Federal Reserve Bank of Chicago

# Sample Selection Models Without Exclusion Restrictions: Parameter Heterogeneity and Partial Identification

*Bo E. Honoré and Luojia Hu*

July 2021

WP 2022-33

https://doi.org/10.21033/wp-2022-33

[*] *Working papers are not edited, and all opinions and errors are the responsibility of the author(s). The views expressed do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.*

# Sample Selection Models without Exclusion Restrictions: Parameter Heterogeneity and Partial Identification[*]

Bo E. Honoré[†] Luojia Hu[‡]

July, 2021

## Abstract

This paper studies semiparametric versions of the classical sample selection model (Heckman (1976, 1979)) without exclusion restrictions. We extend the analysis in Honoré and Hu (2020) by allowing for parameter heterogeneity and derive implications of this model. We also consider models that allow for heteroskedasticity and briefly discuss other extensions. The key ideas are illustrated in a simple wage regression for females. We find that the derived implications of a semiparametric version of Heckman's classical sample selection model are consistent with the data for women with no college education, but strongly rejected for women with a college degree or more.

Key Word: Selection, Heterogeneity, Heteroskedasticity, Exclusion Restrictions, Identification
JEL Code: C01, C14, C21, C24

[*] This research was supported by the Gregory C. Chow Econometric Research Program at Princeton University and by the National Science Foundation (Grant Number SES-1530741). 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. We thank the associate editor, and anonymous referee, Sharada Dharmasankar, and Sebastian Roelsgaard for useful comments.

[†] 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 revisits Heckman's classical sample selection model (Heckman (1976, 1979))

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

where $y_i = y_i^*$ is observed if

$$d_i \equiv 1\{w_i'\gamma + \nu_i > 0\} = 1. \tag{2}$$

The variables $w_i$ and $d_i$ are assumed to be observed for everybody, while it is only necessary to observe $x_i$ when $d_i = 1$. The parameter vector, $\beta$, is the object of interest. The intercepts in (1) and (2) are implicitly captured in $\varepsilon_i$ and $\nu_i$, respectively.

In his seminal papers, Heckman (1976, 1979) considered estimation of this model under the assumption that $(\varepsilon_i, \nu_i)$ are distributed according to a bivariate normal distribution independently of $(x_i, w_i)$. Later research, such as Powell (1987), was able to relax the normality assumption provided that there are elements in $w_i$ that are excluded from $x_i$. See Powell (1994) for a survey of this literature. Unfortunately, such exclusion restrictions can sometimes be difficult to find.[1] In Honoré and Hu (2020), we therefore investigated what one can learn about $\beta$ in the model defined by equations (1) and (2) if there are no exclusion restrictions, so $w_i = x_i$, and the only distributional assumption on the pair of errors, $(\varepsilon_i, \nu_i)$, is that it is independent of $x_i$. The parameter vector, $\beta$, is generally[2] not point identified in that case, but it turns out that provided that $\gamma$ is identified up to scale, the identified region for $\beta$ is a line segment in $\mathbb{R}^k$, where $k$ is the dimensionality of $\beta$. The empirical example in Honoré and Hu (2020) suggests that this identified region can be small enough to be empirically useful.

In a series of papers, James Heckman has emphasized the importance of allowing for individual-specific heterogeneity in econometric models (see, for example, Heckman (2001)). In this paper, we consider generalizations of the classical sample selection model that allow for

[1] For example, Krueger and Whitmore (2001) estimated a sample selection model assuming normality "as there is no exclusion restriction."

[2] Chamberlain (1986) shows that one can identify $\beta$ if $x_i$ has unbounded support.

heterogeneity in the main parameter of interest, as well as for conditional heteroskedasticity. Since the standard sample selection model is generally not point identified without exclusion restrictions, the models considered here will also only be partially identified.

Our aim is to provide identified sets that can be empirically useful, although we do not claim that they are sharp. We illustrate the usefulness of the identified sets by constructing identified regions in a simple wage regression with sample selection. In our application, the coefficient on a dummy variable for being white will be the parameter of interest.

Lee (2009) also considered a sample selection model without exclusion restrictions. He focused on the effect of a binary explanatory variable, "treatment", in a sample selection model. Lee's setup is much less parametric than the Heckman sample selection model, and he was able to derive tight bounds for the mean effect of treatment for the subset of individuals who would have been selected into the sample whether or not they are treated. Lee (2009)'s bounds have been used in a number of different contexts, but some papers have pointed out that the Lee bounds can be too wide to be useful in practice. For example, Barrow and Rouse (2018) wrote "Unfortunately, Lee Bounds estimates (Lee, 2009) are quite wide and largely uninformative." Since Lee's bounds are the tightest possible under his assumptions, this suggests that in those cases, either one should give up on estimating sample selection models, or one should maintain more structure. In addition, the parameter that Lee considers is the mean effect of treatment for the subset of individuals who would have been selected into the sample whether or not they are treated. It is not entirely clear why one should be interested in this particular average effect if there is parameter heterogeneity.

The potential for parameter heterogeneity in the outcome equation of a sample selection model is the main motivation for this paper. We also briefly discuss a number of other extensions to the general framework displayed in equations (1) and (2). Specifically, we consider the implications of heteroskedasticity in (1), the potential for identification through nonlinearities in (2), and panel data versions of the basic model. Finally, we briefly consider a potential outcomes version of the sample selection model.

Our approach builds on the insights in Honoré and Hu (2020). We review the basic idea of that paper in Section 2. In Section 3, we consider a model in which the parameter of interest is allowed to be heterogeneous, and Section 4 provides an empirical illustration of the ideas in Section 3. Section 5 allows for heteroskedasticity in (1). The ideas here are illustrated in the empirical Subsection 5.3. Section 6 investigates various generalizations and Section 7 writes the model in terms of potential outcomes. Section 8 concludes.

Throughout the paper, we focus on the parameter on a binary (0/1) explanatory variable (the "treatment"), but we allow for additional continuous "controls." In most of the paper, we also maintain the assumption that the heterogeneous parameter is independent of the random errors in the model.

## 2 Identification Strategy in the Simplest Case

Honoré and Hu (2020) discuss identification without exclusion restrictions in the classical sample selection model

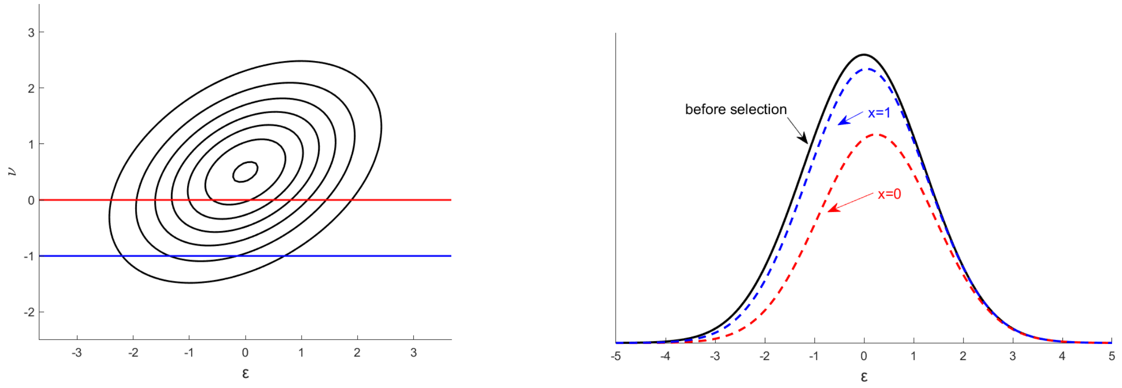
$$y_i^* = x_i'\beta + \varepsilon_i, \tag{3}$$

where $y_i = y_i^*$ is observed if $d_i \equiv 1\{x_i'\gamma + \nu_i > 0\} = 1$, and $(\varepsilon_i, \nu_i)$ is independent of $x_i$. The main insight in that paper is based on the simpler model where there is a single explanatory variable, which is binary taking each of the values 0 and 1 with positive probability. In that case, there is no loss of generality in assuming that $\gamma = 1$.

The left hand panel of Figure 1 displays the joint distribution of $(\varepsilon_i, \nu_i)$ and the two horizontal lines depict the fact that one only observes $y_i$ when $\nu_i > 0$ (when $x_i = 0$) or when $\nu_i > -1$ (when $x_i = 1$). The right hand panel of Figure 1 shows the marginal density of $\varepsilon_i$ before selection, as well as the density times the probability of selection conditional on $\varepsilon_i$ for $x_i = 0$ and $x_i = 1$. The interpretation of the latter two graphs is that the sample selection puts some probability mass at "$\varepsilon_i$ is unobserved"; the remaining mass is then distributed with density given by the two graphs (depending on $x_i = 0$ or $x_i = 1$). Below, we refer to those two graphs as "sub-densities" because they integrate to the probability of selection (as opposed to integrating to 1).

The key assumption in the selection model is that the selection is monotone in $x_i$, meaning that an individual with a particular draw of $(\varepsilon_i, \nu_i)$ who is selected into the sample when $x_i = 0$, would also be selected with $x_i = 1$. This implies that the sub-density of $\varepsilon_i$ when $x_i$ is

Figure 1: Distribution of $\varepsilon$ Before and After Selection



(a) Joint Distribution of $(\varepsilon, \nu)$

(b) Distribution of $\varepsilon$

1 in Figure 1 is above the sub-density when $x_i$ is 0. This in turn implies that the sub-density of $y_i - \beta$ for $x_i = 1$ is above the sub-density of $y_i$ for $x_i = 0$. Honoré and Hu (2020) show that this characterizes the sharp identified region for $\beta$. The paper then uses the same insight to construct a sharp identified set for $\beta$ in the case $x_i$ is multidimensional and not necessarily binary. Finally, Honoré and Hu (2020) propose estimation of a non-sharp identified region for $\beta$ by considering interval probabilities rather than densities.

The sample selection equation (2) is essential for the approach in Honoré and Hu (2020). This equation implies that the sample selection is monotone in $w_i'\gamma$ for a given draw of the errors $(\varepsilon_i, \nu_i)$, and it is this monotonicity that leads to comparisons like the one in the right hand side of Figure 1.

# 3 Parameter Heterogeneity

Heckman (2001) and others have emphasized the importance of heterogeneity. This is also implicit in the analysis in Lee (2009), who derived bounds for the average parameter value in a certain subset of the population.

One way to introduce heterogeneity in the sample selection model is by allowing a subset of the parameters to vary across individuals. For example, if $x_{i1}$ is the variable of interest, then one might specify the model

$$y_i^* = x_{1i}\beta_{1i} + x_{2i}'\beta_2 + \varepsilon_i,$$

where $y_i = y_i^*$ is observed if $d_i = 1$, where $d_i = 1\{x_i'\gamma + \nu_i > 0\}$, and where $x_i = (x_{1i}, x_{2i})$. Without a scale normalization of $\nu_i$, $\gamma$ is at best identified up to scale, and the sign of each element of $\gamma$ is identified. We assume the first element[3] of $\gamma$ is not 0, and we normalize $\gamma$ so that $|\gamma_1| = 1$.

Except where we explicitly state otherwise, we will assume that $\beta_{1i}$, $(\varepsilon_i, \nu_i)$, and $x_i$ are independent. For identification of the distribution of $\beta_{1i}$, this can be relaxed somewhat by conditioning on $x_{2i}$. We do not pursue this because it is unlikely to be useful in practice when $x_{2i}$ is multidimensional and contains continuously distributed variables. The assumption that $\beta_{1i}$ is independent of $(\varepsilon_i, \nu_i)$ is strong; however, it is clear that some assumption of this type is necessary in order to make statements about, say, the population mean of $\beta_{1i}$.[4] The assumptions that $\nu_i$ is independent of $x_i$ and that $\gamma$ is constant again imply that the sample selection is monotone in $x_i'\gamma$.

## 3.1 Binary Regressor

We first consider the case with only one explanatory variable, $x_i$, which is binary. We assume that $\gamma_1 = 1$, so that the sample selection is more severe when $x_i$ is 0 than when it is 1. We observe

$$y_i = x_i\beta_i + \varepsilon_i \qquad \text{if} \qquad x_i + \nu_i > 0.$$

The key observation again is that the selection is monotone in $x_i$. Individuals who are selected with $x_i = 0$ would also be selected with $x_i = 1$ and with the same $(\varepsilon_i, \nu_i)$. For any

[3] If all elements of $\gamma$ are 0, then this is known from the population distribution of the data, and in that case there is no sample selection bias.

[4] For example, while Lee (2009) does not make such an assumption, the bounds derived in that paper are for the average treatment effect for the individuals who would have been selected into the sample whether or not they were treated. Conditional expectations like that can only be turned into population-wide expectations by making additional assumptions.

set $A$, we therefore have

$$\begin{aligned} P\left(\varepsilon_i \in A, \nu_i > -1\right) &= P\left(\varepsilon_i \in A, \nu_i > 0\right) + P\left(\varepsilon_i \in A, 0 \geq \nu_i > -1\right) \\ &\geq P\left(\varepsilon_i \in A, \nu_i > 0\right). \end{aligned}$$

It therefore follows that

$$P\left(\varepsilon_i + b \in A, \nu_i > -1\right) \geq P\left(\varepsilon_i + b \in A, \nu_i > 0\right) \text{ for any } b. \tag{4}$$

In this special case, $\beta_i$ only matters when $x_i = 1$. As a result, we do not need to assume that $\beta_i$ and $x_i$ are independent. Instead we assume that $(\varepsilon_i, \nu_i)$ is independent of $x_i$ and that $\beta_i$ is independent of $(\varepsilon_i, \nu_i)$ conditional on $x_i = 1$. The distribution of $\beta_i$ conditional on $x_i = 1$, $F_\beta$, belongs to some class of distributions $\mathcal{F}_\beta$. The class of distributions could be a parametric family of distributions, or $F_\beta$ could be left nonparametric. Typically, the class of distributions for $\beta_i$ will include degenerate distributions, in which case the model with parameter homogeneity becomes a special case of the model considered here. With this, (4) implies

$$\int P\left(\varepsilon_i + b \in A, \nu_i > -1\right) dF_\beta\left(b\right) \geq \int P\left(\varepsilon_i + b \in A, \nu_i > 0\right) dF_\beta\left(b\right). \tag{5}$$

Recalling that $y_i = \beta_i + \varepsilon_i$ when $x_i = 1$ and $y_i = \varepsilon_i$ when $x_i = 0$, equation (5) becomes

$$P\left(y_i \in A, d_i = 1 \middle| x_i = 1\right) \geq \int P\left(y_i + b \in A, d_i = 1 \middle| x_i = 0\right) dF_\beta\left(b\right),$$

so one identified set for $F_\beta$ is

$$\left\{F \in \mathcal{F}_\beta : P\left(y_i \in A, d_i = 1 \middle| x_i = 1\right) \geq \int P\left(y_i + b \in A, d_i = 1 \middle| x_i = 0\right) dF\left(b\right)\right\}. \tag{6}$$

For example, if we restrict $\beta$ to be discrete, taking on $K$ values with probabilities $\pi_k$, then the distribution of $\beta$ must belong to

$$\Bigg\{ \left(K, \{\beta_k, \pi_k\}_{k=1}^K\right) : P\left(y_i \in A, d_i = 1 | \, x_i = 1\right) \\ \geq \sum_k \pi_k P\left(y_i + \beta_k \in A, d_i = 1 | \, x_i = 0\right) \text{ for all } A, \sum_k \pi_k = 1 \Bigg\}.$$

This places restrictions on $K$ and $\{\beta_k, \pi_k\}_{k=1}^K$. In the applications below, we specify $K$, but it could in principle be considered a parameter to be estimated or bounded.

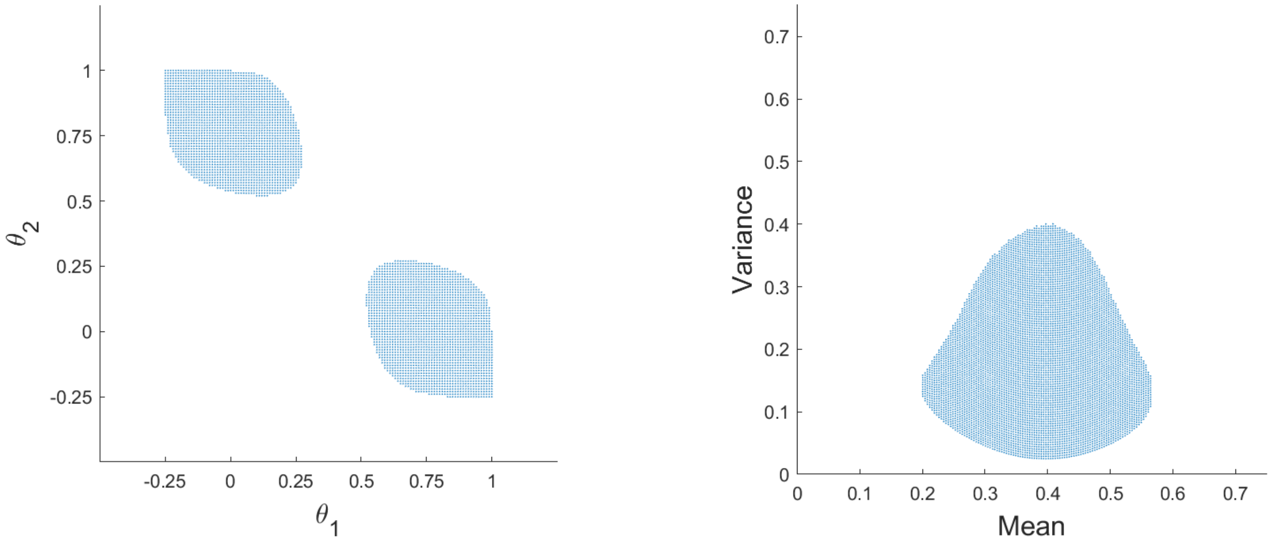
Since $\beta_i$ is allowed to depend on $x_i$, the restriction that $\beta_i$ takes a finite number of values, $K$, is similar in spirit to the group-specific heterogeneity studied in Bonhomme and Manresa (2015).

**Example 1** *Let* $(\varepsilon_i, \nu_i)$ *be distributed according to a bivariate normal distribution with* $E[\varepsilon_i] = 0$, $E[\nu_i] = 1$, $V[\varepsilon_i] = 1$, $V[\nu_i] = 1$ *and* $\operatorname{cov}(\varepsilon_i, \nu_i) = 0.5$, *and let* $P(\beta_{1i} = 0) = P(\beta_{1i} = 1) = \frac{1}{2}$. *If the researcher knows that* $\beta_{1i}$ *has a discrete distribution with two equally likely points of support,* $\theta_1$ *and* $\theta_2$, *but does not know the distribution of* $(\varepsilon_i, \nu_i)$, *then the identified region for* $(\theta_1, \theta_2)$ *is as depicted in the left hand panel of Figure 2. The identified set is symmetric around the 45-degree line since the two points of support enter the model symmetrically.*

It is difficult to graphically present the identified set of a parameter of dimension higher than two. In those cases, it can be useful to present identified regions for summary statistics of the parameter. For example, when the object of interest is a distribution, one might present the joint identified set for the mean and the variance.

**Example 2 (Continuation of Example 1)** *Consider the same data generating progress as in Example 1. If the researcher knows that* $\beta_{1i}$ *has a discrete distribution with two points of support, but does not know the associated probabilities, then the identified region for the mean and variance of* $\beta_{1i}$ *is as depicted in the right hand panel of Figure 2. Combined with 1, this example illustrates that the sample selection model with parameter heterogeneity is far from point-identified even when the distribution of the heterogeneity is tightly parameterized, but that one can still construct informative bounds on objects of interest such as the mean and the variance of the heterogeneous parameter.*

Figure 2: Examples 1 and 2



(a) Identified Region for points of support

(b) Identified region for $(E[\beta_{1i}], SD[\beta_{1i}])$

The same line of argument can be used to find the identified set for the distribution of $\beta_{1i}$ when $\gamma = -1$, that is, the case where the selection is more severe when $x_i$ is 1 than when it is 0. It is

$$\Bigg\{ (\beta_k, \pi_k) : P\left(y_i \in A, d_i = 1 | \, x_i = 1\right) \\ \le \sum_k \pi_k P\left(y_i + \beta_k \in A, d_i = 1 | \, x_i = 0\right) \text{ for all } A, \sum_k \pi_k = 1 \Bigg\}.$$

## 3.2 Generalization to Multiple $x$

We now turn to the more general case where $x_{1i}$ is still a binary (0/1) "treatment", but there are additional explanatory variables. For simplicity, we assume that these have homogeneous parameters and that $\gamma_1$ is positive (and normalized to 1). Specifically,

$$y_i^* = x_{1i}\beta_{1i} + x_{2i}'\beta_2 + \varepsilon_i,$$

where $y_i = y_i^*$ is observed when $d_i \equiv 1\left\{x_i'\gamma + \nu_i > 0\right\}$ equals 1. We maintain the assumption that $\beta_{1i}$, $(\varepsilon_i, \nu_i, )$ and $x_i$ are independent. We also assume that a set of sufficient conditions for $\gamma$ to be identified up to scale are satisfied[5] (see for example Klein and Spady (1993)) and that conditional on $d_i = 1$, $(\varepsilon_i, x'_{2i}, x'_i\gamma)$ satisfy the conditions on $(U_i, X_i, Z_i)$ in Robinson (1988). The distribution of the heterogeneous parameter, $\beta_{1i}$, can in principle be continuous or discrete, although we restrict it to be discrete in the application in the next section. We assume that $E[\beta_{1i}]$ is finite. The selection equation, $d_i \equiv 1\{x'_i\gamma + \nu_i > 0\}$, is a monotonicity assumption which states that if $y_i$ is observed for an individual with $x'_i\gamma = \xi_1$, then $y_i$ would also be observed if $x'_i\gamma = \xi_2 > \xi_1$ and $(\varepsilon_i, \nu_i)$ is left unchanged.

To construct bounds for the distribution of $\beta_{1i}$, one can pick an arbitrary $b_2$ and apply (6) with $y$ replaced by $y - x'_2 b_2$. This gives a (possibly empty) identified set for the distribution of $\beta_{1i}$ for each $b_2$. One identified set for the distribution of $\beta_{1i}$ is then the union (over $b_2$) of these. Unfortunately, this approach is difficult to implement, unless the dimensionality of $x_{2i}$ is small. We therefore pursue an alternative approach.

Conditional on selection, and conditional on $\beta_{1i}$, we have

$$y_i = x_{1i}\beta_{1i} + x'_{2i}\beta_2 + g(x'_i\gamma) + u_i,$$

where $g(x'_i\gamma) = E[\varepsilon_i | x_i, x'_i\gamma + \nu_i > 0]$ and $E[u_i | x_i, \beta_{1i}] = 0$. This implies that

$$\begin{aligned} E[y_i | x'_i\gamma] &= E[x_{1i} | x'_i\gamma] E[\beta_{1i} | x'_i\gamma] + E[x_{2i} | x'_i\gamma] \beta_2 + g(x'_i\gamma) \\ &= E[x_{1i} | x'_i\gamma] \beta_{1i} - E[x_{1i} | x'_i\gamma] (\beta_{1i} - E[\beta_{1i} | x'_i\gamma]) + E[x_{2i} | x'_i\gamma] \beta_2 + g(x'_i\gamma) \end{aligned}$$

and therefore

$$y_i - E[y_i | x'_i\gamma] = (x_{1i} - E[x_{1i} | x'_i\gamma]) \beta_{1i} + (x_{2i} - E[x_{2i} | x'_i\gamma])' \beta_2 - E[x_{1i} | x'_i\gamma] (\beta_{1i} - E[\beta_{1i} | x'_i\gamma]) + u_i.$$

As in Honoré and Hu (2020), $(x_{1i} - E[x_{1i} | x'_i\gamma]) + (x_{2i} - E[x_{2i} | x'_i\gamma])' \gamma_2 = 0$. We therefore

[5] When $\gamma$ is not point-identified up to scale, the approach below can be applied to each point in the identified set for $\gamma$.

have

$$\begin{aligned}
y_i - E\left[y_i | x_i'\gamma\right] &= -\left(x_{2i} - E\left[x_{2i} | x_i'\gamma\right]\right)' \gamma_2 \beta_{1i} + \left(x_{2i} - E\left[x_{2i} | x_i'\gamma\right]\right)' \beta_2 \\
&\quad -E\left[x_{1i} | x_i'\gamma\right] \left(\beta_{1i} - E\left[\beta_{1i} | x_i'\gamma\right]\right) + u_i \\
&= \left(x_{2i} - E\left[x_{2i} | x_i'\gamma\right]\right)' \left(\beta_2 - \gamma_2 \beta_{1i}\right) - E\left[x_{1i} | x_i'\gamma\right] \left(\beta_{1i} - E\left[\beta_{1i} | x_i'\gamma\right]\right) + u_i \\
&= \left(x_{2i} - E\left[x_{2i} | x_i'\gamma\right]\right)' \left(\beta_2 - \gamma_2 E\left[\beta_{1i} | x_i'\gamma\right]\right) - E\left[x_{1i} | x_i'\gamma\right] \left(\beta_{1i} - E\left[\beta_{1i} | x_i'\gamma\right]\right) + u_i \\
&\quad - \left(x_{2i} - E\left[x_{2i} | x_i'\gamma\right]\right)' \gamma_2 \left(\beta_{1i} - E\left[\beta_{1i} | x_i'\gamma\right]\right)
\end{aligned}$$

Since the last three terms have mean 0 conditional on $x_i$, and $\beta_{1i}$ is assumed to be independent of $x_i$ (so $E\left[\beta_{1i} | x_i'\gamma\right] = E\left[\beta_{1i}\right]$), this implies that

$$\alpha_2 \equiv \left(\beta_2 - \gamma_2 E\left[\beta_{1i}\right]\right)$$

is identified provided that $\left(x_{2i} - E\left[x_{2i} | x_i'\gamma\right]\right)$ has full rank.

Having identified $\alpha_2$, we write

$$\begin{aligned}
y_i^* - x_{2i}'\alpha_2 &= y_i^* - x_{2i}'\left(\beta_2 - \gamma_2 E\left[\beta_{1i}\right]\right) = x_{1i}\beta_{1i} + x_{2i}'\beta_2 + \varepsilon_i - x_{2i}'\left(\beta_2 - \gamma_2 E\left[\beta_{1i}\right]\right) \\
&= x_{1i}\beta_{1i} + \left(x_{2i}'\gamma_2\right) E\left[\beta_{1i}\right] + \varepsilon_i,
\end{aligned}$$

or

$$y_i^* - x_{2i}'\alpha_2 - \left(x_{2i}'\gamma_2\right) E\left[\beta_{1i}\right] = x_{1i}\beta_{1i} + \varepsilon_i.$$

In other words

$$y_i^* - x_{2i}'\alpha_2 - \left(x_{2i}'\gamma_2\right) E\left[\beta_{1i}\right] = \varepsilon_i \qquad \text{when} \qquad x_{1i} = 0 \tag{7}$$

and

$$y_i^* - x_{2i}'\alpha_2 - \left(x_{2i}'\gamma_2\right) E\left[\beta_{1i}\right] = \beta_{1i} + \varepsilon_i \qquad \text{when} \qquad x_{1i} = 1. \tag{8}$$

Thinking of the left hand side as a dependent variable, equations (7) and (8) have the same structure as the problem in Section 3.1. The main difference is that to arrive at (7) and (8) we assumed independence between $\beta_{1i}$ and $x_i$. This was not necessary when there is a single, binary, explanatory variable. Moreover, the selection probability is now monotone in the index $x_i'\gamma$.

Combining (7) and (8), we have

$$y_i^* - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E[\beta_{1i}] + 1\{x_{1i} = 0\}\, \beta_{1i} = \beta_{1i} + \varepsilon_i.$$

Hence, for any interval $A$ and for $\xi_1 < \xi_2$,

$$\begin{aligned} &P\left(\left(y_i^* - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E[\beta_{1i}] + 1\{x_{1i} = 0\}\, \beta_{1i}\right) \in A, d_i = 1 |\, x_i'\gamma = \xi_2\right) \geq \\ &P\left(\left(y_i^* - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E[\beta_{1i}] + 1\{x_{1i} = 0\}\, \beta_{1i}\right) \in A, d_i = 1 |\, x_i'\gamma = \xi_1\right). \end{aligned} \tag{9}$$

Therefore we can construct an identified set for the distribution of $\beta_{1i}$, $F$, as

$$\Big\{F \in \mathcal{F}_\beta : \int P\Big(\left(y_i - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E_F[\beta_{1i}] + b\right) \in A, d_i = 1 |\, x_i'\gamma = \xi_2\Big) dF(b)$$
$$\geq \int P(\ \left(y_i - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E_F[\beta_{1i}] + b\right) \in A, d_i = 1 |\, x_i'\gamma = \xi_1) dF(b)\Big\}$$

for all $\xi_1 < \xi_2$. We use the notation $E_F[\beta_{1i}]$ as a reminder that the expectation of $\beta_{1i}$ in (9) will depend on $F$.

For each $F$ in the identified set for the distribution of $\beta_{1i}$, the average treatment effect is $E_F[\beta_{1i}]$. The remaining parameter vector, $\beta_2$, is given by $\alpha_2 + \gamma_2 E_F[\beta_{1i}]$ where $\gamma_2$ is identified from the semiparametric discrete choice model $d_i \equiv 1\{x_i'\gamma + \nu_i > 0\}$ with $\gamma = (\gamma_1, \gamma_2')'$ and the normalization that $\gamma_1 = 1$.

# 4 Empirical Illustration

To illustrate the approach outlined above, we consider a simple sample selection model for wages for females. The question is how to make statements about the coefficient on being white, $\beta_1$, without exclusion restrictions.

We first estimate the model under joint normality of the errors using the maximum likelihood estimator and Heckman's two-step estimator. The parameters of the model are not point-identified without a distributional assumption on the errors. We therefore next apply the method in Honoré and Hu (2020) to construct a confidence region for $\beta_1$ under the assumption that this parameter is homogenous. After that, we use the approach discussed in Section 3.2 to estimate a (two point) discrete distribution for $\beta_1$.

Using the Current Population Survey from 1982 to 2018, we construct a data set of 1,060,351 females aged 25 to 65. Of them, 552,446 are working and have a recorded (real) wage. The explanatory variable of interest is a dummy for being white, and the additional explanatory variables are the unemployment rate, a time trend, age, age-squared, and two education indicators (one for some college, and one for college and beyond).

Table 1 reports the maximum likelihood estimates that assume joint normality of $(\varepsilon_i, \nu_i)$ for the full sample and for the three subsamples defined by educational group. Table 2 reports the corresponding 2-step estimates. Comparing the estimates in Table 1 and 2 makes it very clear that the normality assumption is violated. For example, under the null that the normality assumption is satisfied, the standard error of the difference in the estimates of the coefficient on being white for the full sample would be 0.0036. The difference in the point estimates is 0.0225, which leads to a t-statistic of more than 6. The values of the corresponding t-statistics for the three subsample are all above 2.3 in absolute value.

To estimate the semiparametric version of the sample selection model that acknowledges that the coefficients are only partially identified without exclusion restrictions, we turn the constraints in (9) into a finite number of moment inequalities by first dividing the range of $x_i'\widehat{\gamma}$ into five regions, $C_\ell$, defined by quintiles. For a given candidate distribution of $\beta_{1i}$, we then divide the range of $y - x_2'\widehat{\alpha}_2 - (x_{2i}'\widehat{\gamma}_2)\, E_F[\beta_{1i}]$ into ten regions, $A_j$, defined by the deciles of $y - x_2'\widehat{\alpha}_2 - (x_{2i}'\widehat{\gamma}_2)\, E_F[\beta_{1i}]$. This gives moment conditions of the type

$$\begin{aligned} E\left[1\left\{(y_i^* - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E_F[\beta_{1i}] + 1\{x_{1i}=0\}\,\beta_{1i}) \in A_j, d_i = 1\right\} \middle|\, x_i'\gamma \in C_\ell\right] \geq \\ E\left[1\left\{(y_i^* - x_{2i}'\alpha_2 - (x_{2i}'\gamma_2)\, E_F[\beta_{1i}] + 1\{x_{1i}=0\}\,\beta_{1i}) \in A_j, d_i = 1\right\} \middle|\, x_i'\gamma \in C_{\ell-1}\right]. \end{aligned} \tag{10}$$

Since the distribution of $\nu$ is left unspecified, we should in principle estimate $\gamma$ semiparametrically, for example by employing the maximum rank estimator of Han (1987) or the estimator proposed by Klein and Spady (1993). These can be difficult and computationally expensive to calculate. Below, we calculate confidence sets by subsampling, and we therefore

Table 1: Parametric Estimation under Normality (MLE)

| | All | No College | Some College | College Plus |
|---|---|---|---|---|
| White | 0.062 | 0.117 | 0.060 | 0.020 |
| | (0.002) | (0.003) | (0.003) | (0.003) |
| Unemployment Rate | 0.012 | 0.011 | 0.006 | 0.006 |
| | (0.000) | (0.001) | (0.001) | (0.001) |
| Year | 0.004 | 0.002 | 0.000 | 0.007 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| Age | 0.241 | 0.272 | 0.528 | 0.502 |
| | (0.006) | (0.015) | (0.013) | (0.011) |
| Age-Squared | -0.021 | -0.028 | -0.057 | -0.053 |
| | (0.001) | (0.002) | (0.002) | (0.001) |
| Some College | 0.195 | | | |
| | (0.002) | | | |
| College Plus | 0.528 | | | |
| | (0.002) | | | |
| Constant | 1.112 | 0.923 | 0.527 | 1.083 |
| | (0.016) | (0.041) | (0.032) | (0.026) |
| Observations | 1,060,351 | 519,750 | 264,233 | 276,368 |

Standard errors in parentheses

estimate $\gamma$ by a logit maximum likelihood. Following Robinson (1988), we estimate $\alpha_2$ by regressing $y_i - \widehat{E}\left[y_i | x_i'\widehat{\gamma}\right]$ on $\left(x_{2i} - \widehat{E}\left[x_{2i} | x_i'\widehat{\gamma}\right]\right)$, where the $\widehat{E}$'s are constructed by kernel estimation. The estimator of the distribution of $\beta_{1i}$ is then defined by minimizing the sum of the squares of the negative deviations between the sample analogs of the left and right hand sides of (10). Specifically, we define

$$
\begin{aligned}
&R_{j\ell}(F) \\
&\quad = E_F\left[\widehat{E}\left[1\left\{\left(y_i^* - x_{2i}'\widehat{\alpha}_2 - \left(x_{2i}'\widehat{\gamma}_2\right) E_F\left[\beta_{1i}\right] + 1\left\{x_{1i} = 0\right\}\beta_{1i}\right) \in A_j, d_i = 1\right\} \middle| x_i'\widehat{\gamma} \in C_\ell\right]\right],
\end{aligned}
$$

where $\widehat{E}$ refers to sample averages as well as averaging $\beta_{1i}$ over the distribution $F$, and

Table 2: Parametric Estimation under Normality (2-Step)

| | All | No College | Some College | College Plus |
|---|---|---|---|---|
| White | 0.040 | 0.081 | 0.073 | 0.095 |
| | (0.004) | (0.012) | (0.006) | (0.027) |
| Unemployment Rate | 0.016 | 0.017 | -0.001 | -0.007 |
| | (0.001) | (0.002) | (0.002) | (0.006) |
| Year | 0.005 | 0.003 | -0.002 | 0.001 |
| | (0.000) | (0.000) | (0.001) | (0.002) |
| Age | 0.037 | 0.012 | 0.976 | 1.684 |
| | (0.033) | (0.085) | (0.127) | (0.322) |
| Age-Squared | 0.007 | 0.005 | -0.117 | -0.233 |
| | (0.004) | (0.011) | (0.017) | (0.049) |
| Some College | 0.144 | | | |
| | (0.008) | | | |
| College Plus | 0.449 | | | |
| | (0.012) | | | |
| Constant | 1.715 | 1.641 | -0.672 | -2.344 |
| | (0.095) | (0.234) | (0.339) | (0.927) |
| Observations | 1,060,351 | 519,750 | 264,233 | 276,368 |

Standard errors in parentheses

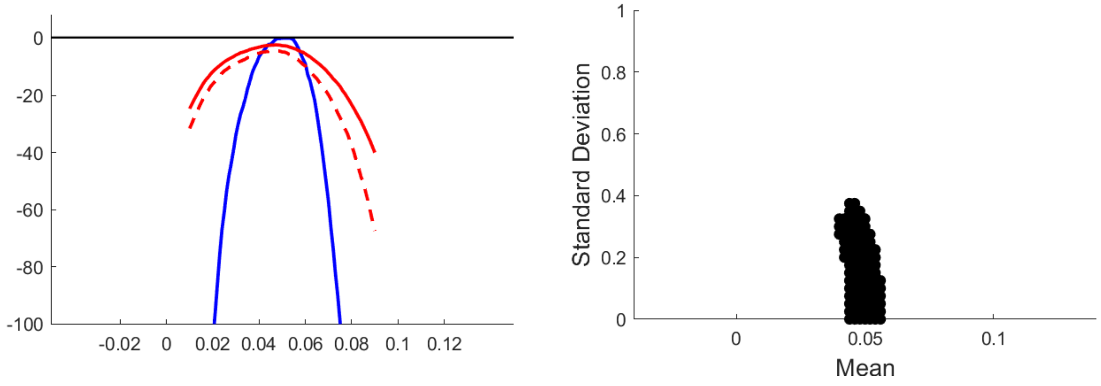
$E_F\left[\beta_{1i}\right]$ is the expectation of $\beta_{1i}$ calculated using $F$. We then calculate the objective function

$$Q_n\left(F\right) = -\sum_{\ell,j} \max\left\{R_{j,\ell-1}\left(F\right) - R_{j,\ell}\left(F\right), 0\right\}^2 \tag{11}$$

for the distributions, $F$, under consideration.

We first calculate the identified region for the coefficient on the dummy for being white ($\beta_1$) in a model with homogeneous parameters. In other words, $F$ is degenerate. In this case, the approach here is the same as that in Honoré and Hu (2020). The left hand panel of Figure 3 displays the objective function, (11), as a function of the parameter, as well as the 20% (the solid red line) and 5% (the dashed red line) critical value functions calculated using sub-sampling (see Canay and Shaikh (2017)) for the full sample. We generate 1,000 subsamples, each having sample size equal to 50,000. The left hand panel of Figure 3 shows the 95% confidence interval for the coefficient on being white to be $(0.042, 0.060)$. This overlaps

Figure 3: Estimated Effect. Full Sample.



(a) Homogenous Effect

(b) Heterogeneous Effect

with the confidence intervals suggested by the maximum likelihood estimator and by the 2-step estimator which assume normality. It is quite time-consuming to calculate the critical values by subsampling. For the remaining results, we therefore generate 250 subsamples and report 20% critical regions.

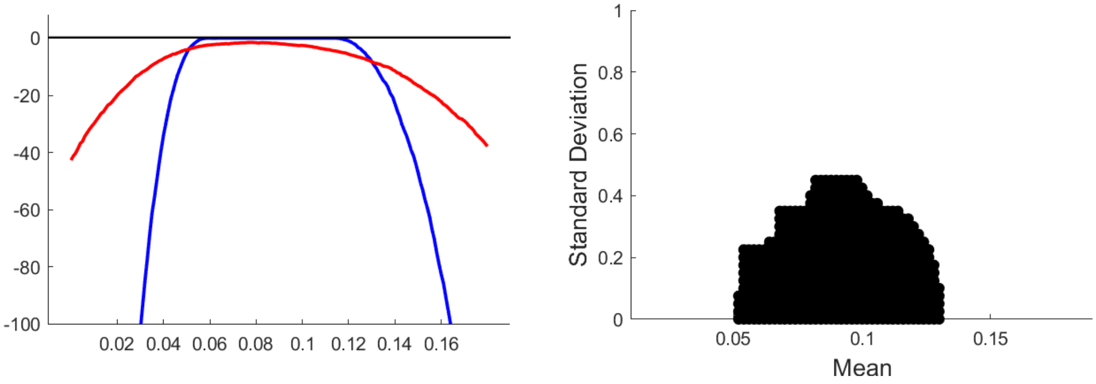
In the right panel of Figure 3, we report the 80% confidence region for the identified set in a model in which the coefficient on the dummy for being white is allowed to take on two values, $\theta_1$ and $\theta_2$, with probabilities $p$ and $1-p$. We restrict $\theta_1$ and $\theta_2$ to be between $-1$ and 1, and the grid for $p$ is $0, 0.05, 0.10, ..., 0.50$. The confidence set is calculated by sub-sampling as above.

The identified region in the right panel of Figure 3 contains points for which the standard deviation of the parameter of interest is 0. This is consistent with the fact that the left hand panel of Figure 3 gives a non-empty confidence region under the assumption of parameter homogeneity. On the other hand, the identified regions also contain points for which the standard deviation of $\beta_1$ is quite high relative to its mean. The identified set for the average effect, $E\left[\beta_{1i}\right]$, is fairly small, although it does include points that are lower than the 80% confidence region that would be obtained under the assumption that $\beta_1$ is homogenous.[6]

Figures 4, 5, and 6 show the estimated effects for the three subsamples defined by educa-

[6] As pointed out by a referee, it might not be interesting to estimate the effect of being White *conditional* on education. We performed the calculations leading to Figure 3 excluding education as an explanatory variable. This model is strongly rejected by the data, and we therefore do not pursue this further in this paper.

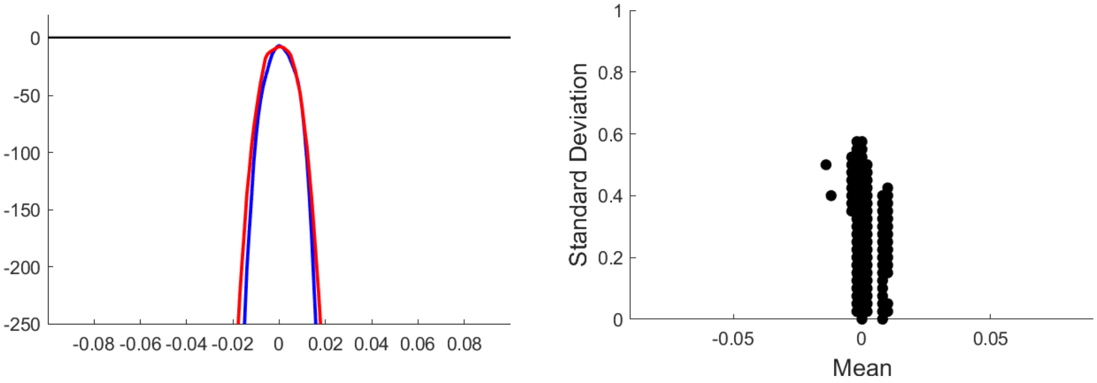
Figure 4: Estimated Effect. No College.



(a) Homogenous Effect

(b) Heterogeneous Effect

Figure 5: Estimated Effect. Some College.



(a) Homogenous Effect

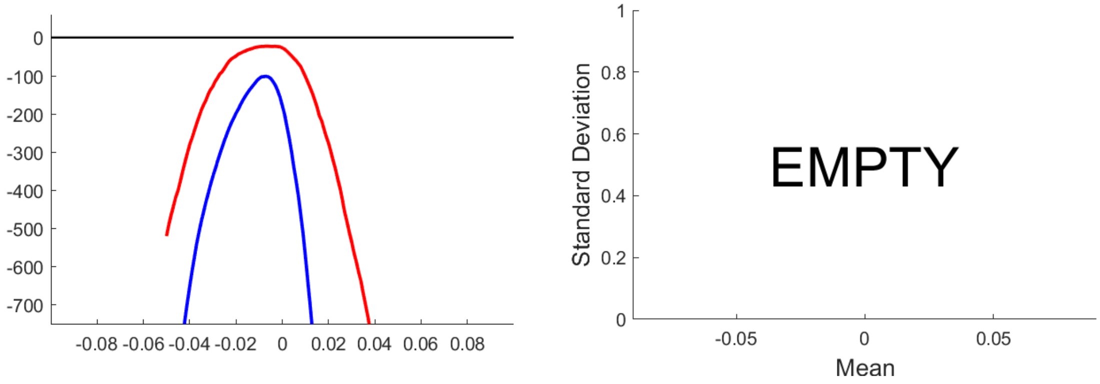
(b) Heterogeneous Effect

tion group. Since the sample sizes for these subsets are smaller than for the full data set (see Table 1), we use sub-sample sizes of 30,000, 20,000, and 20,000 for the three subsamples.

The most striking finding in Figures 4, 5, and 6 is that the confidence set is empty for the subset of observations with at least a college degree. This suggests that the simple sample selection model is inconsistent with the data. We also find it interesting that the location of the maximum of the objective function for this group is slightly negative. This is consistent with the pseudo-maximum-likelihood estimate of the coefficient of being white in Table 1 being low (0.0195, with a robust standard error of 0.0033) for this sample.

The left hand panel of Figure 5 is difficult to read because we have kept the scale of the x-axis the same across Figures 3 to 6. Figure 7 shows the same objective function using

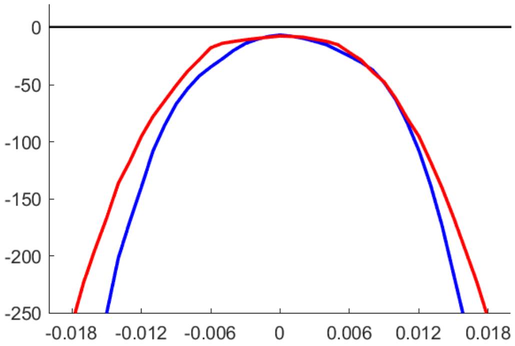
Figure 6: Estimated Effect. College Plus.



(a) Homogenous Effect

(b) Heterogeneous Effect

Figure 7: Estimated Effect. Some College.

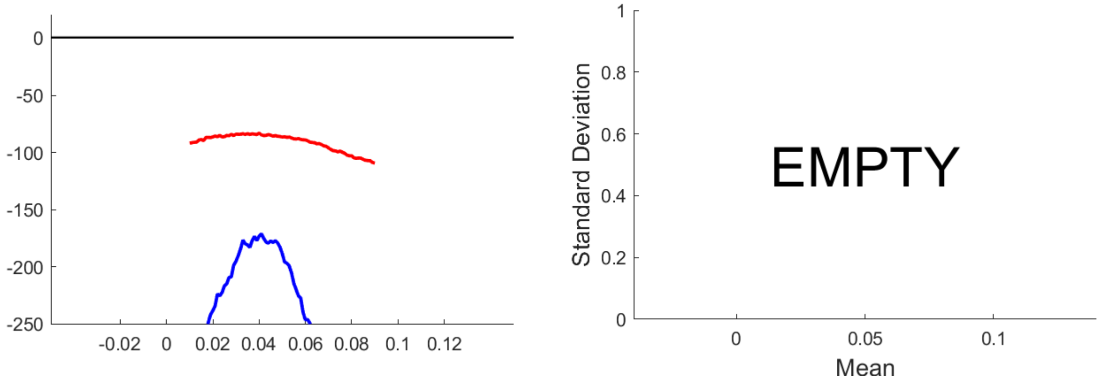


a different scale. Formally speaking, it suggests a very small 80% confidence interval for a homogeneous $\beta_1$. However, it also suggests that this apparent precision is due to the fact that the model is only marginally not rejected by the data.

We also note that the scale of the objective functions in the left hand panels of Figures 3 and 4 are quite different from those in Figures 5 and 6. Informally, this hints at the samples of women with some college and at least a college degree being more at odds with the simple sample selection model than the other samples.

Figure 4 suggests that the derived implications of the classical sample selection model are consistent with the data. As noted above, the parametric maximum likelihood and two-step estimation results obtained by estimating the model under the assumption of normality lead to parameter estimates that are statistically significantly different from each other,

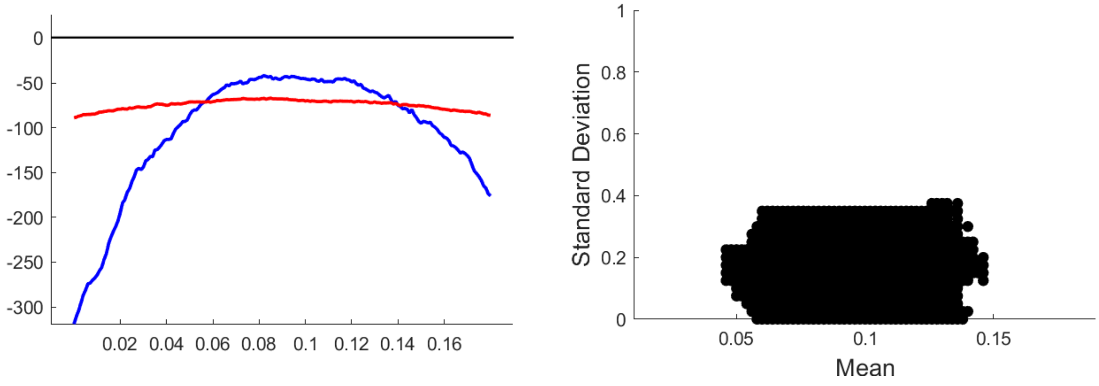
Figure 8: Estimated Effect Using More Bins. Full Sample.



(a) Homogenous Effect

(b) Heterogeneous Effect

Figure 9: Estimated Effect Using More Bins. No College.



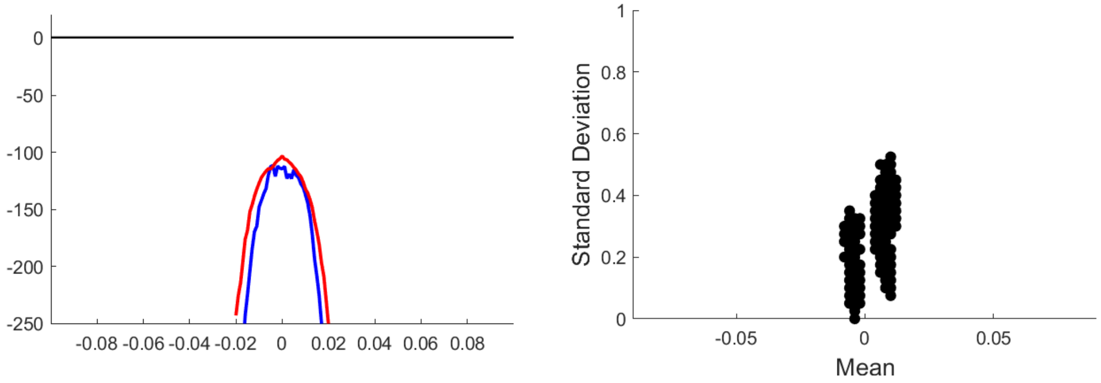
(a) Homogenous Effect

(b) Heterogeneous Effect

suggesting that the normality assumption is violated. For example, the absolute value of the t-tests for testing equality of each coefficient ranges between 2.1 and 3.1.

The equation (10) aggregates the constraints in equation (9) into 50 moment inequalities. One might worry that will lead to important loss of information. In Figures 8, 9, 10, and 11, we present the confidence sets that are obtained by dividing both the distribution of $x_i'\widehat{\gamma}$ and of $y - x_2'\widehat{\alpha}_2 - (x_{2i}'\widehat{\gamma}_2)\, E_F[\beta_{1i}]$ into 15 intervals, each based on their percentiles. As anticipated, this leads to smaller confidence regions, and the model is now rejected on the full sample.

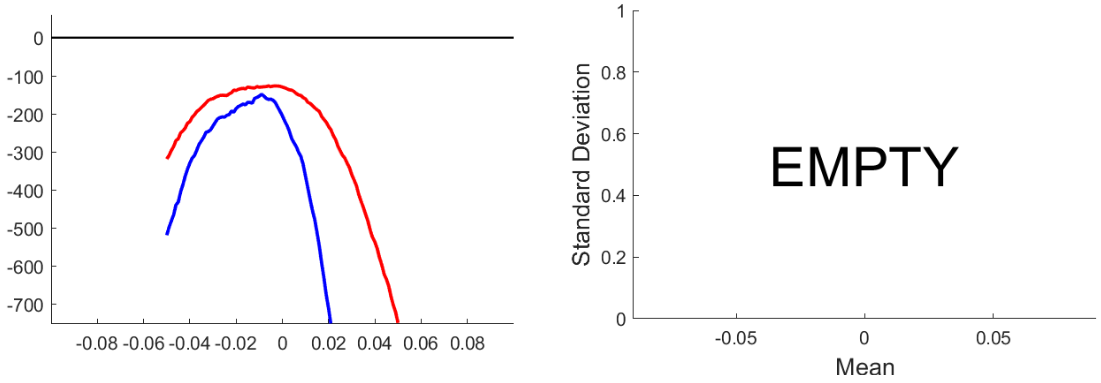
Figure 10: Estimated Effect Using More Bins. Some College.



(a) Homogenous Effect

(b) Heterogeneous Effect

Figure 11: Estimated Effect Using More Bins. College Plus.



(a) Homogenous Effect

(b) Heterogeneous Effect

# 5 Heteroskedasticity

Most of the classical literature on the estimation of semiparametric sample selection models assumes that the error in the outcome equation is independent of the explanatory variables. Donald (1995) is a notable early exception. That paper assumes joint normality of the errors, but allows their variance matrix to depend arbitrarily on the explanatory variables. In other words, Donald (1995) introduces heteroskedasticity by multiplying the errors by an unknown function of $x$. In the spirit of this, Chen and Khan (2003) allow for multiplicative heteroskedasticity in both the selection equation and the outcome equation. That paper maintains an exclusion restriction.[7]

In this subsection, we first investigate identification when the error in the outcome equation is allowed to have multiplicative heteroskedasticity and then proceed to the more general case. We maintain the assumption that the error in the selection equation is independent of the explanatory variable. The reason is that heteroskedasticity in the selection equation can actually aid in the identification of $\beta$. See, for example, Ahn and Powell (1993), Chen and Khan (2003), Escanciano, Jacho-Chávez, and Lewbel (2016) and the discussion in Section 6.2.

## 5.1 Univariate $x$

We start by introducing multiplicative heteroskedasticity in the outcome equation of the simple model (3), where the only explanatory variable is binary. In this subsection, we assume that the parameter is homogeneous, so the only complication is the heteroskedasticity. In order to simplify the exposition, we focus on the case where the sample selection is more severe when $x_i = 0$ than when $x_i = 1$ (i.e. $\gamma$ in the sample selection equation is positive and normalized to 1).

Multiplicative heteroskedasticity only makes sense after one has controlled for the level (for example, the mean or the median) of the errors. We therefore write

[7] Klein and Vella (2009) consider a related model with a dummy endogenous variable. That paper allows for heteroskedasticity in both equations, and the heteroskedasticity in the selection equation is assumed to be multiplicative; see also Klein and Vella (2010).

$$y_i = \beta_0 + x_i\beta + \sigma(x_i)\,\varepsilon_i \qquad \text{is observed if} \qquad x_i + \nu_i > 0. \tag{12}$$

When $\sigma(x)$ is constant, $\beta_0$ becomes part of $\varepsilon_i$. Otherwise, $\beta_0$ is the level around which the multiplicative heteroskedasticity operates. Below, we use $\sigma$ to denote $\sigma(1)$. Since the distribution of $\varepsilon_i$ is unspecified, there is no loss of generality in assuming that $\sigma(0) = 1$.

As before, the monotonicity of the selection equation implies that

$$P(\varepsilon_i \in A, \nu_i > 0) \leq P(\varepsilon_i \in A, \nu_i > -1)$$

for any $A$. Since $\varepsilon_i$ is independent of $x_i$, and with $d_i = 1\{x_i + \nu_i > 0\}$, this can be written as

$$P(\varepsilon_i \in A, d_i = 1|\, x_i = 0) \leq P(\varepsilon_i \in A, d_i = 1|\, x_i = 1).$$

As $y_i = \varepsilon_i + \beta_0$ when $x_i = 0$, and $y_i = \beta_0 + \beta + \sigma\varepsilon_i$ when $x_i = 1$, the true $(\beta_0 + \beta, \sigma)$ must therefore satisfy

$$P(y_i - \beta_0 \in A, d_i = 1|\, x_i = 0) \leq P(\,(y_i - \beta - \beta_0)/\,\sigma \in A, d_i = 1|\, x_i = 1).$$

This gives the following identified set for $(\beta_0, \beta, \sigma)$:

$$\left\{(b_0, b, s) : P(y_i - b_0 \in A, d_i = 1|\, x_i = 0) \leq P(\,(y_i - b_0 - b)/\,s \in A, d_i = 1|\, x_i = 1) \text{ for all } A\right\}.$$

This can be extended to a model with non-multiplicative heteroskedasticity. Suppose that when $x_i = 0$ we observe

$$y_i = \varepsilon_i^0 \qquad \text{if} \qquad \nu_i > 0,$$

and when $x_i = 1$ we observe

$$y_i = \beta + \varepsilon_i^1 \qquad \text{if} \qquad \nu_i > -1.$$

As before, $x_i$ is independent of the errors $(\varepsilon_i^0, \varepsilon_i^1, \nu_i)$, and $\beta_0$ is now implicit in $\varepsilon_i^0$ and $\varepsilon_i^1$. Of course, without restrictions on the distributions of $\varepsilon_i^0$ and $\varepsilon_i^1$, $\beta$ will be unidentified since one can incorporate it in $\varepsilon_i^1$. We therefore have in mind that their distributions, $F_0$ and $F_1$, belong to some class of distributions, $\mathcal{F}$. For example, $F_0$ and $F_1$ could be restricted to having mean or median equal to 0.

For simplicity, assume that $\mathcal{F}$ restricts both $\varepsilon_i^0$ and $\varepsilon_i^1$ to have continuous, strictly increasing cumulative distribution functions, $F_0$ and $F_1$, respectively. Then $\varepsilon_i^0$ is distributed like $F_0^{-1}(F_1(\varepsilon_i^1))$.

As before, we have the inequality

$$P\left(\varepsilon_i^0 \in A, d_i = 1 \middle| x_i = 0\right) \leq P\left(\varepsilon_i^0 \in A, d_i = 1 \middle| x_i = 1\right),$$

or

$$P\left(\varepsilon_i^0 \in A, d_i = 1 \middle| x_i = 0\right) \leq P\left(F_0^{-1}\left(F_1\left(\varepsilon_i^1\right)\right) \in A, d_i = 1 \middle| x_i = 1\right).$$

Let $g(\cdot) = F_0^{-1}(F_1(\cdot))$. Using the fact that $\varepsilon_i^0 = y_i$ when $x_i = 0$ and that $\varepsilon_i^1 = y_i - \beta$ when $x_i = 1$, we then have

$$P\left(y_i \in A, d_i = 1 | x_i = 0\right) \leq P\left(g\left(y_i - \beta\right) \in A, d_i = 1 | x_i = 1\right).$$

So one identified set for $\beta$ is

$$\begin{aligned}\big\{b : &\text{There exists an increasing function, } g(\cdot) = F_0^{-1}(F_1(\cdot)) \text{ with } F_0, F_1 \in \mathcal{F}\text{, such that} \\ &\qquad P\left(y_i \in A, d_i = 1 | x_i = 0\right) \leq P\left(g\left(y_i - b\right) \in A, d_i = 1 | x_i = 1\right) \text{ for all } A\big\}.\end{aligned}$$

Restrictions on the form of heteroskedasticity will appear as restrictions on the function $g$ in the expression above. For example, with the multiplicative heteroskedasticity above, $\varepsilon_i^0 = \beta_0 + \varepsilon_i$ and $\varepsilon_i^1 = \beta_0 + \beta_1 + \sigma\varepsilon_i$. Therefore $F_1(a) = F_0(\beta_0 + (a - \beta_0 - \beta_1)/\sigma)$ and $g(y_i) = F_0^{-1}(F_1(y_i)) = \beta_0 + (a - \beta_0 - \beta_1)/\sigma$.

## 5.2 Multiple $x$ and Heteroskedasticity

Allowing for heteroskedasticity is more complicated when the model includes additional explanatory variables.

Consider the model

$$y_i^* = \beta_0 + x_i'\beta + \sigma\left(x_i\right)\varepsilon_i,$$

where $\sigma$ belongs to a class of heteroskedasticity functions. When $\sigma\left(x_i\right)$ is constant, $\beta_0$ can be incorporated into $\varepsilon_i$. When $\sigma\left(x_i\right)$ is not constant, $\beta_0$ is implicitly defined as the central tendency parameter around which the multiplicative heteroskedasticity operates.

For the true heteroskedasticity parameter, $\sigma$,

$$y_i^* / \sigma\left(x_i\right) = \beta_0 / \sigma\left(x_i\right) + x_i / \sigma\left(x_i\right)' \beta + \varepsilon_i. \tag{13}$$

Suppose that the function $\sigma$ is known and is not a constant. The explanatory variables in (13) are then not the same as in the sample selection equation, and the key assumption for identification is that conditional on $x_i'\gamma$, $\left[1/\sigma\left(x_i\right), x_i/\sigma\left(x_i\right)\right]$ has "full rank" (i.e., is not contained in a linear subspace of $\mathbb{R}^{k+1}$ (with probability 1)). This will typically be satisfied unless $\sigma\left(x_i\right)$ is constant.

One possible approach for bounding $\beta$ (in the population) would then be to apply the following procedure to each candidate function, $\sigma\left(x_i\right)$. If $\left[1/\sigma\left(x_i\right), x_i/\sigma\left(x_i\right)\right]$ has full rank conditional on $x_i'\gamma$, then this identifies $\beta\left(\sigma\right)$. It must then be the case that

$$\begin{aligned} &P\left(\left.\left(y_i - x_i'\beta\left(\sigma\right)\right)/\sigma\left(x_i\right) \in A, d_i = 1\right| x_i'\gamma = \xi_1\right) \\ &\qquad \leq P\left(\left.\left(y_i - x_i'\beta\left(\sigma\right)\right)/\sigma\left(x_i\right) \in A, d_i = 1\right| x_i'\gamma = \xi_2\right) \end{aligned} \tag{14}$$

for all $A$ and $\xi_1 < \xi_2$. If that is not the case, then that $\sigma\left(x_i\right)$ can be eliminated from the identified set. If $\left[1/\sigma\left(x_i\right), x_i/\sigma\left(x_i\right)\right]$ does not have full rank conditional on $x_i'\gamma$, then Honoré and Hu (2020) delivers the identified set for $\beta$ for that $\sigma$. If that identified set is empty, then $\sigma$ can be eliminated from the identified set.

Suppose, for example,

$$y_i^* = \beta_0 + x_{1i}\beta_1 + x_{2i}'\beta_2 + \sigma\left(x_i\right)\varepsilon_i,$$

where $x_{1i}$ is binary and one specifies the heteroskedasticity function to be a function of $x_{1i}$

alone:

$$\sigma\left(x_i\right)=\begin{cases}1 & \text{if} \quad x_{1i}=0\\ \sigma & \text{if} \quad x_{1i}=1.\end{cases}$$

In this case $x_i/\,\sigma\left(x_i\right)$ will not have full rank conditional on $x_i'\gamma$ if $x$ is composed of all interactions between $x_{1i}$ and a vector of variables $w_i$ (in other words, $x_{2i}=\left(\left(1-x_i\right)\cdot w_i, x_i\cdot w_i\right)$).

### 5.3 Empirical Illustration

The discussion in Section 4 suggests that our simple specification of the classical sample selection model is strongly rejected for the sample of women with a college degree or more. In this section, we explore whether the data are consistent with the derived implications of a version of the sample selection model in which the errors are heteroskedastic as a function of being white. Except for allowing for heteroskedasticity, the specification is the same as in Section 4.

For a set of values of $\sigma$ (bounded away from 1), we consider the model

$$\begin{aligned} y_i^*/\sigma &= \beta_0/\sigma+\beta_1/\sigma+\left(x_{2i}/\sigma\right)'\beta_2+\varepsilon_i \quad \text{when} \quad x_{1i}=1 \quad \text{and} \\ y_i^* &= \beta_0+x_{2i}'\beta_2+\varepsilon_i \quad \text{when} \quad x_{1i}=0, \end{aligned}$$

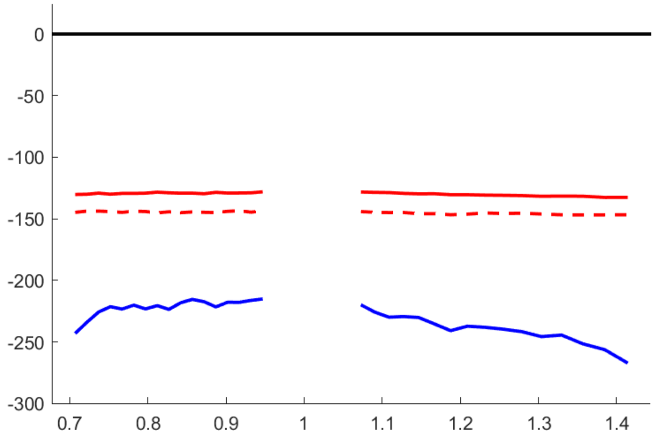
and $y_i=y_i^*$ is observed whenever $x_{1i}+x_{2i}'\gamma_2+\nu_i>0$. Here, $x_{1i}$ is an indicator for being white. For each value of $\sigma$, we estimate this model using the estimator of the semiparametric sample selection model proposed by Powell (1987).[8] We bound $\sigma$ away from 1, because when $\sigma$ is 1, the key identifying exclusion restriction for Powell's estimator is not satisfied, and we expect the inference to be arbitrarily poor when $\sigma$ is arbitrarily close to 1.

Consider a collection of sets, $A_j$ and $C_\ell$, where the $C_\ell$'s are increasing in the sense that the every element in $C_{\ell-1}$ is below each element in $C_\ell$. We define

$$R_{j,\ell}\left(\sigma\right)=\widehat{E}\left[1\left\{\left(y_i^*/\,\sigma\left(x_i\right)-\widehat{\beta}_0\Big/\,\sigma\left(x_i\right)-x_i'\widehat{\beta}\Big/\,\sigma\left(x_i\right)\right)\in A_j, d_i=1\right\}\Big|\, x_i'\widehat{\gamma}\in C_\ell\right].$$

[8] We estimate the selection equation using a logit model, and we use a normal kernel and a bandwidth of 0.05 for the outcome equation.

Figure 12: Allowing for Heteroskedasticity



The functions $R_{j,\ell}(\sigma)$ estimate the terms on each side of the inequality in (14) aggregated over a set of values of $\xi$. In the population, and at the true parameter values, $R_{j,\ell}(\sigma) \leq R_{j,k}(\sigma)$ for $\ell < k$. We therefore define the objective function

$$Q_n(\sigma) = -\sum_{\ell,j} \max\{R_{j,\ell-1}(F) - R_{j,\ell}(F), 0\}^2 .$$

If the model is correct and the sample is large, $Q_n(\sigma)$ should be close to 0 at the true $\sigma$.

As in Section 4, we calculate the 20% (the solid red line) and 5% (the dashed red line) critical value functions by subsampling using 1,000 subsamples with 20,000 observations in each. The resulting graphs are shown in Figure 12.

It is clear from Figure 12 that for the sample of women with a college degree or more, our simple specification of the classical sample selection model is still strongly rejected after allowing for heteroskedasticity.

# 6 Generalizations

In this section, we discuss identification in several extensions of the sample selection model above. Throughout, we assume that $\gamma_1$ is positive and normalized to 1. In other words, we assume that everything else equal, the sample selection is less severe when $x_{1i}$ is one than when it is zero.

## 6.1 $\boldsymbol{\beta}$ depends on $\boldsymbol{\varepsilon}$

In the discussion in Section 3, we assumed that $\beta_i$ is independent of $(\varepsilon_i, \nu_i)$. This is a strong assumption. Here, we illustrate one way to proceed under the alternative (strong) assumption that $\beta_i$ is a deterministic function of $\varepsilon_i$. To simplify the exposition, we focus on the case where there is only one binary explanatory variable:

$$y_i = \beta\left(\varepsilon_i\right) x_i + \varepsilon_i \qquad \text{if} \qquad x_i + \nu_i > 0.$$

For any interval, $A$, we have $P\left(\varepsilon_i \in A, d_i = 1 \middle| x_i = 0\right) \leq P\left(\varepsilon_i \in A, d_i = 1 \middle| x_i = 1\right)$ and hence

$$P\left(\beta\left(\varepsilon_i\right) + \varepsilon_i \in A, d_i = 1 \middle| x_i = 0\right) \leq P\left(\beta\left(\varepsilon_i\right) + \varepsilon_i \in A, d_i = 1 \middle| x_i = 1\right),$$

or

$$P\left(\beta\left(y_i\right) + y_i \in A, d_i = 1 \middle| x_i = 0\right) \leq P\left(y_i \in A, d_i = 1 \middle| x_i = 1\right).$$

This provides a set of restrictions which can be used to bound the function $\beta\left(\cdot\right)$.

## 6.2 Identification Through Possible Nonlinearity

The heteroskedasticity in Section 5.2 transformed the explanatory variable in the outcome equation from $x_i$ to $x_i/\sigma\left(x_i\right)$, making the explanatory variables in the outcome equation a nonlinear function of the explanatory variable in the selection equation. This gives an exclusion restriction which can be used to achieve identification of $\beta$ for a known function, $\sigma$. In the spirit of Escanciano, Jacho-Chávez, and Lewbel (2016), we can also consider identification through nonlinearities in the *selection* equation. The basic idea in that paper is that the nonlinearity in the selection equation can act as an excluded variable in the outcome equation (see also the discussion in Ahn and Powell (1993)).

To explore this avenue for identification in a model with parameter heterogeneity, we again start with the equation

$$y_i^* = x_{1i}\beta_{1i} + x_{2i}'\beta_2 + \varepsilon_i, \tag{15}$$

where $\beta_{1i}$ is assumed to be independent of $(x_i, \nu_i, \varepsilon_i)$.

Without (much) loss of generality, assume that $d_i = 1\{p(x_i) > \nu_i\}$ where $\nu_i$ is uniform (so $P(d_i = 1|\, x_i) = p(x_i)$). Then

$$y_i = x_{1i}\beta_{1i} + x_{2i}'\beta_2 + h(p(x_i)) + u_i,$$

where $h(p(x_i)) = E[\varepsilon_i|\, p(x_i) > \nu_i]$, $u_i = \varepsilon_i - E[\varepsilon_i|\, p(x_i) > \nu_i]$, and $E[u_i|\, x_i, d_i = 1] = 0$.

Hence,

$$\begin{aligned} y_i - E[y_i|\, p(x_i), d_i = 1, \beta_{1i}] = {} & (x_{1i} - E[x_{1i}|\, p(x_i), d_i = 1, \beta_{1i}])\, \beta_{1i} \\ & + (x_{2i}' - E[x_{2i}'|\, p(x_i), d_i = 1, \beta_{1i}])\, \beta_2 + \widetilde{u}_i, \end{aligned}$$

where $\widetilde{u}_i = u_i - E[u_i|\, p(x_i), d_i = 1, \beta_{1i}]$ has conditional mean 0. Since $\beta_{1i}$ is assumed to be independent of $(x_i, \nu_i, \varepsilon_i)$, this becomes

$$\begin{aligned} y_i - E[y_i|\, p(x_i), d_i = 1] = {} & (x_{1i} - E[x_{1i}|\, p(x_i), d_i = 1])\, E[\beta_{1i}] \\ & + (x_{2i}' - E[x_{2i}'|\, p(x_i), d_i = 1])\, \beta_2 + \widetilde{\widetilde{u}}_i, \end{aligned}$$

where $\widetilde{\widetilde{u}}_i = \widetilde{u}_i + (x_{1i} - E[x_{1i}|\, p(x_i), d_i = 1, \beta_{1i}])\, (\beta_{1i} - E[\beta_{1i}])$ has conditional mean 0. This identifies $(E[\beta_{1i}], \beta_2)$ subject to a rank condition on $((x_{1i} - E[x_{1i}|\, p(x_i), d_i = 1]), (x_{2i}' - E[x_{2i}'|\, p(x_i), d_i = 1]))$.

## 6.3 Panel Data

We finally note that the general approach outlined in this paper also applies to panel data versions of the sample selection model like the one studied in Kyriazidou (1997):

$$y_{it}^* = x_{it}'\beta + \alpha_i + \varepsilon_{it}$$

where $y_{it}^*$ is observed whenever $x_{it}'\gamma + \delta_i + \nu_{it} > 0$. Here $\alpha_i$ and $\delta_i$ play the roles of fixed effects in the outcome and selection equations, respectively. It is well known that $\gamma$ is identified up to scale subject to regularity conditions (see Manski (1987)), so from an identification point of view, we can consider it known.

If $x_{i2}'\gamma > x_{i1}'\gamma$, then $P(\varepsilon_{i2} \in A, d_{i2} = 1) \geq P(\varepsilon_{i1} \in A, d_{i1} = 1)$ as above. This implies that $P(\varepsilon_{i2} + \alpha_i \in A, d_{i2} = 1) \geq P(\varepsilon_{i1} + \alpha_i \in A, d_{i1} = 1)$. Writing this in terms of the observed $y_{it}$, we therefore have $P(y_{i2} - x_{i2}'\beta \in A, d_{i2} = 1) \geq P(y_{i1} - x_{i1}'\beta \in A, d_{i1} = 1)$. This suggests an identified set for $\beta$ of the type

$$\begin{aligned}\{\beta : P\left(y_{i2} - x_{i2}'\beta \in A, d_{i2} = 1 \middle| x_{i2}'\gamma > x_{i1}'\gamma\right) & \\ \geq P\left(y_{i1} - x_{i1}'\beta \in A, d_{i1} = 1 \middle| x_{i2}'\gamma > x_{i1}'\gamma\right) & \text{ for all } A\}.\end{aligned}$$

# 7 Potential Outcomes

The key to the relative simplicity of the identified region discussed so far is that the heterogeneous parameter has been multiplied by a binary $x_{1i}$. This implies that when $x_{1i} = 0$, the distribution of $y_i$ differs from the distribution of $\varepsilon_i$ only because of the selection and the additional controls, $x_{2i}$. The heterogeneity of $\beta_{1i}$ plays no role when $x_{1i} = 0$. In some cases, this might seem somewhat artificial. For example, in the empirical illustration in Section 4, the model would be different if we redefine $x_{1i}$ to be 0 for whites and 1 for non-whites. One way to overcome this is to use the potential outcomes setup frequently used in the program evaluation literature. Within the structure of the selection model discussed here, one would specify the potential outcomes as[9]

$$y_i^* = \begin{cases} \beta_{0i} + x_{2i}'\beta_2 + \varepsilon_i & \text{when} \quad x_{1i} = 0 \\ \beta_{1i} + x_{2i}'\beta_2 + \varepsilon_i & \text{when} \quad x_{1i} = 1, \end{cases}$$

where $y_i$ is observed if $x_i'\gamma + \nu_i > 0$ . We assume that $(\beta_{0i}, \beta_{1i})$ is independent of $(\varepsilon_i, \nu_i, x_i)$, but $\beta_{0i}$ and $\beta_{1i}$ need not be independent of each other.

[9] Here $\beta_2$ is the same whether $x_{1i}$ is 0 or 1. This is easily relaxed by interacting $x_{2i}$ with $x_{1i}$.

To fix ideas, we first consider the case where there are no additional controls ($x_{2i}$) and where the selection is more severe when $x_{1i} = 0$ than when $x_{1i} = 1$ (i.e., $\gamma_1 = 1$). In this case,

$$P\left(\varepsilon_i \in A,\, d_i = 1 \middle| x_{1i} = 0\right) \leq P\left(\varepsilon_i \in A,\, d_i = 1 \middle| x_{1i} = 1\right)$$

for any set $A$. This implies that

$$P\left(\left(\varepsilon_i + \widetilde{\beta}_{0i} + \widetilde{\beta}_{1i}\right) \in A,\; d_i = 1 \middle| x_{1i} = 0\right) \leq P\left(\left(\varepsilon_i + \widetilde{\beta}_{0i} + \widetilde{\beta}_{1i}\right) \in A,\; d_i = 1 \middle| x_{1i} = 1\right), \tag{16}$$

where $\widetilde{\beta}_{0i}$ and $\widetilde{\beta}_{1i}$ are independent of each other, with $\widetilde{\beta}_{0i}$ drawn from the marginal distribution of $\beta_{0i}$ and $\widetilde{\beta}_{1i}$ drawn from the marginal distribution of $\beta_{1i}$.

In terms of the observable data, $y_i$ is distributed like $\varepsilon_i + \widetilde{\beta}_{0i}$ when $x_{1i} = 0$ and like $\varepsilon_i + \widetilde{\beta}_{1i}$ when $x_{1i} = 1$. Equation (16) can therefore be written as

$$P\left(\left(y_i + \widetilde{\beta}_{1i}\right) \in A, d_i = 1 \middle| x_{1i} = 0\right) \leq P\left(\left(y_i + \widetilde{\beta}_{0i}\right) \in A,\; d_i = 1 \middle| x_{1i} = 1\right),$$

where $\widetilde{\beta}_{0i}$ and $\widetilde{\beta}_{1i}$ are independent of the data and distributed as the marginal distributions of $\beta_{0i}$ and $\beta_{1i}$, respectively. This yields constraints on the marginal distributions of $\beta_{0i}$ and $\beta_{1i}$.

Additional controls, $x_{2i}$, can be dealt with as in Section 3. Conditional on selection, and conditional on $(\beta_{0i}, \beta_{1i})$, we have

$$y_i = \left(1 - x_{1i}\right)\beta_{0i} + x_{1i}\beta_{1i} + x_{2i}'\beta_2 + g\left(x_i'\gamma\right) + u_i,$$

where $g\left(x_i'\gamma\right) = E\left[\varepsilon_i \middle| x_i, x_i'\gamma + \nu_i > 0\right]$ and $E\left[u_i \middle| x_i, \beta_{1i}\right] = 0$.

Using that $\left(x_{1i}-E\left[\left.x_{1i}\right|x_i'\gamma\right]\right)=-\left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)'\gamma_2$, we therefore have

$$\begin{aligned}
y_i-E\left[\left.y_i\right|x_i'\gamma\right] &= \left(1-x_{1i}\right)\beta_{0i}+x_{1i}\beta_{1i}+\left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)'\beta_2+u_i \\
&\quad -\left(1-E\left[\left.x_{1i}\right|x_i'\gamma\right]\right)E\left[\beta_{0i}\right]-E\left[\left.x_{1i}\right|x_i'\gamma\right]E\left[\beta_{1i}\right] \\
&= \left(1-x_{1i}\right)\left(E\left[\beta_{0i}\right]+\beta_{0i}-E\left[\beta_{0i}\right]\right)+x_{1i}\left(E\left[\beta_{1i}\right]+\beta_{1i}-E\left[\beta_{1i}\right]\right) \\
&\quad +\left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)'\beta_2+u_i-\left(1-E\left[\left.x_{1i}\right|x_i'\gamma\right]\right)E\left[\beta_{0i}\right]-E\left[\left.x_{1i}\right|x_i'\gamma\right]E\left[\beta_{1i}\right] \\
&= \left(1-x_{1i}\right)E\left[\beta_{0i}\right]+\left(1-x_{1i}\right)\left(\beta_{0i}-E\left[\beta_{0i}\right]\right)+x_{1i}E\left[\beta_{1i}\right]+x_{1i}\left(\beta_{1i}-E\left[\beta_{1i}\right]\right) \\
&\quad +\left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)'\beta_2+u_i-\left(1-E\left[\left.x_{1i}\right|x_i'\gamma\right]\right)E\left[\beta_{0i}\right]-E\left[\left.x_{1i}\right|x_i'\gamma\right]E\left[\beta_{1i}\right] \\
&= \left(x_{1i}-E\left[\left.x_{1i}\right|x_i'\gamma\right]\right)E\left[\beta_{1i}-\beta_{0i}\right]+ \\
&\quad \left(1-x_{1i}\right)\left(\beta_{0i}-E\left[\beta_{0i}\right]\right)+x_{1i}\left(\beta_{1i}-E\left[\beta_{1i}\right]\right)+\left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)'\beta_2+u_i \\
&= \left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)'\left(\beta_2-\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]\right) \\
&\quad +\left(1-x_{1i}\right)\left(\beta_{0i}-E\left[\beta_{0i}\right]\right)+x_{1i}\left(\beta_{1i}-E\left[\beta_{1i}\right]\right)+u_i
\end{aligned}$$

The term $\left(1-x_{1i}\right)\left(\beta_{0i}-E\left[\beta_{0i}\right]\right)+x_{1i}\left(\beta_{1i}-E\left[\beta_{1i}\right]\right)+u_i$ has mean 0 conditional on $x_i$, and we can therefore identify $\alpha_2\equiv\left(\beta_2-\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]\right)$ by regressing $\left(y_i-E\left[\left.y_i\right|x_i'\gamma\right]\right)$ on $\left(x_{2i}-E\left[\left.x_{2i}\right|x_i'\gamma\right]\right)$.

With this, we have

$$\begin{aligned}
y_i^*-x_{2i}'\alpha_2 &= y_i^*-x_{2i}'\left(\beta_2-\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]\right) \\
&= \left(1-x_{1i}\right)\beta_{0i}+x_{1i}\beta_{1i}+x_{2i}'\beta_2+\varepsilon_i-x_{2i}'\left(\beta_2-\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]\right) \\
&= \left(1-x_{1i}\right)\beta_{0i}+x_{1i}\beta_{1i}+x_{2i}'\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]+\varepsilon_i
\end{aligned}$$

or

$$y_i^*-x_{2i}'\alpha_2+\left(1-x_{1i}\right)\beta_{1i}+x_{1i}\beta_{0i}-x_{2i}'\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]=\beta_{0i}+\beta_{1i}+\varepsilon_i.$$

The marginal distributions of $\beta_{0i}$ and $\beta_{1i}$ must therefore satisfy

$$\begin{aligned}
&P\left(\left(y_i^*-x_{2i}'\alpha_2+\left(1-x_{1i}\right)\beta_{1i}+x_{1i}\beta_{0i}-x_{2i}'\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]\right)\in A,d_i=1\right|x_i'\gamma=\xi_2\left.\right) \\
&\quad \geq P\left(\left(y_i^*-x_{2i}'\alpha_2+\left(1-x_{1i}\right)\beta_{1i}+x_{1i}\beta_{0i}-x_{2i}'\gamma_2E\left[\beta_{1i}-\beta_{0i}\right]\right)\in A,d_i=1\right|x_i'\gamma=\xi_1\left.\right)
\end{aligned}$$

for all intervals $A$ and $\xi_2 > \xi_1$.

## 8 Conclusion

Semiparametric sample selection models are generally not point-identified without exclusion restrictions. In earlier work, Honoré and Hu (2020) derived the sharp identified region of the parameters in such a model. In this paper, we extend that analysis to allow for parameter heterogeneity and heteroskedasticity while maintaining the basic linearity, independence and monotonicity assumptions of the classical sample selection model. We also discuss a potential outcomes version of the sample selection model, identification through nonlinearities, and panel data versions of the model.

We illustrate the key insights in a simple wage regression for females, where the parameter of interest is the coefficient on a dummy variable for being white. We find that for the full sample, neither the introduction of parameter heterogeneity nor heteroskedasticity is sufficient for the data to be consistent with the model. The classical sample selection model is especially at odds with the data for the subsample of women with a college degree or more.

## 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.

Barrow, L., and C. E. Rouse (2018): "Financial Incentives and Educational Investment: The Impact of Performance-based Scholarships on Student Time Use," Education Finance and Policy, 13(4), 419–448.

Bonhomme, S., and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data," Econometrica, 83(3), 1147–1184.

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.

Chen, S., and S. Khan (2003): "Semiparametric Estimation of a Heteroskedastic Sample Selection Model," Econometric Theory, 19(6), 1040–1064.

Donald, S. G. (1995): "Two-step estimation of heteroskedastic sample selection models," Journal of Econometrics, 65(2), 347 – 380.

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

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.

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," Journal of Political Economy, 109(4), 673–748.

Honoré, B. E., and L. Hu (2020): "Selection Without Exclusion," Econometrica, 88(3), 1007–1029.

Klein, R., and F. Vella (2009): "A semiparametric model for binary response and continuous outcomes under index heteroscedasticity," Journal of Applied Econometrics, 24(5), 735–762.

——— (2010): "Estimating a class of triangular simultaneous equations models without exclusion restrictions," Journal of Econometrics, 154(2), 154–164.

Klein, R. W., and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," Econometrica, 61(2), 387–421.

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.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," Econometrica, 65, 1335–1364.

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. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," Econometrica, 55, 357–62.

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.