cemmap
centre for microdata methods and practice

Point-identifying semiparametric sample selection models with no excluded variable

Dongwoo Kim
Young Jun Lee

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP07/25

UKRI Economic and Social Research Council

POINT-IDENTIFYING SEMIPARAMETRIC SAMPLE SELECTION MODELS WITH NO EXCLUDED VARIABLE

DONGWOO KIM
Department of Economics, Simon Fraser University

YOUNG JUN LEE
Korea Institute for International Economic Policy

Sample selection is pervasive in applied economic studies. This paper develops semiparametric selection models that achieve point identification without relying on exclusion restrictions, an assumption long believed necessary for identification in semiparametric selection models. Our identification conditions require at least one continuously distributed covariate and certain nonlinearity in the selection process. We propose a two-step plug-in estimator that is $\sqrt{n}$-consistent, asymptotically normal, and computationally straightforward (readily available in statistical software), allowing for heteroskedasticity. Our approach provides a middle ground between Lee (2009)'s nonparametric bounds and Honoré and Hu (2020)'s linear selection bounds, while ensuring point identification. Simulation evidence confirms its excellent finite-sample performance. We apply our method to estimate the racial and gender wage disparity using data from the US Current Population Survey. Our estimates tend to lie outside the Honoré and Hu bounds.

Keywords: sample selection, semiparametric identification, sieve estimation, exclusion restriction.

Dongwoo Kim: dongwook@sfu.ca

Young Jun Lee: y.lee@kiep.go.kr

We thank Krishna Pendakur and Myung Hwan Seo for their helpful suggestions. We also greatly benefited from the seminar participants at Seoul National University, Kyung Hee University, and Sogang University. All errors are our own. The authors gratefully acknowledge support from the Social Sciences and Humanities Research Council of Canada under the Insight Development Grant (430-2022-00841) and the Insight Grant (435-2024-0322).

1. INTRODUCTION

Sample selection poses a fundamental challenge in empirical economics, threatening the validity of research findings. When workers self-select into employment or patients choose whether to seek medical care, the resulting datasets systematically exclude critical segments of the population, distorting our understanding of economic relationships and leading to misguided policy prescriptions. Even carefully designed experiments are vulnerable, as systematic attrition introduces selection bias that undermines randomization. While economists have proposed various solutions, they often rely on strong assumptions or yield uninformative bounds. In particular, nonparametric and semiparametric selection models have long been thought to require exclusion restrictions, limiting their practical applicability. This paper challenges that conventional wisdom by introducing a class of semiparametric selection models that achieve point identification without exclusion restrictions. We further develop computationally tractable two-step plug-in estimators that can be readily implemented using standard statistical software.

Heckman (1974, 1979) pioneered correction methods for selection bias using a parametric model that assumes linearity in both selection and outcome equations, along with joint normality of the error terms:

$$Y^* = \alpha + X\beta + V, \quad D = \mathbb{1}[Z\gamma + \varepsilon \geq 0], \quad Y = D \cdot Y^*, \tag{1}$$

where Y^* is the latent outcome, X and Z are row vectors of exogenous covariates, V and ε are mean-zero unobserved heterogeneity terms that are joint normally distributed and independent of (X, Z), with $Var(\varepsilon)$ normalized to 1. Conditional on $X = x$, $Z = z$, and $D = 1$, the mean of the observed outcome is given by:

$$E[Y|x, z, D = 1] = \alpha + x\beta + \sigma_{V\varepsilon}\phi(z\gamma)/\Phi(z\gamma),$$

where $\sigma_{V\varepsilon} = Cov(V, \varepsilon)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density function (p.d.f.) and cumulative distribution function (c.d.f.) respectively.

Although Heckman's model can identify (α, β) when $X = Z$ due to the known functional form of selection bias, it is generally recommended to include at least one variable in

Z that is excluded from X to strengthen identification. Without such an exclusion restriction, the numerical performance of both the two-step and maximum likelihood estimators can be poor, as highlighted in the debates in Duan et al. (1984), Manning et al. (1987), and Hay and Olsen (1984).

By relaxing Heckman's joint normality assumption, econometricians developed semiparametric approaches (Chamberlain, 1986, Ahn and Powell, 1993, Newey, 2009). Later Das et al. (2003) explored fully nonparametric selection models:

$$Y^* = m(X) + V, \quad D = \mathbb{1}[g(Z) + \varepsilon \geq 0], \quad Y = D \cdot Y^*. \tag{2}$$

In both semiparametric and nonparametric models, the exclusion restriction is widely regarded as essential for identification.[1] However, finding an excluded variable is often infeasible in empirical applications. Motivated by this challenge, Lee (2009) proposed a nonparametric bounds approach that does not require exclusion restrictions. By exploiting the selection monotonicity assumption, under which individuals who are observed without treatment would also be observed with treatment, Lee introduced a trimming procedure to adjust for missing data due to selection. His approach (henceforth Lee bounds) is intuitive and easily implemented, making it widely used in empirical studies, particularly in experiments where subjects tend to drop out.

Lee bounds, however, are often too wide to yield meaningful economic insights and its ability to incorporate covariate information is limited.[2] Honoré and Hu (2020) (henceforth HH) later demonstrated that β is partially identified in (1) without distributional assumptions on (V, ε) and in the absence of exclusion restrictions. The HH model, serves as a semiparametric alternative to Lee's, provides tighter bounds than Lee bounds, as it imposes

[1] Lee (2009) stated that "*standard parametric or semiparametric methods for correcting for sample selection require exclusion restrictions that have little justification in this case*" (p. 1072). Similarly, Honoré and Hu (2020) claimed that an exclusion restriction is the *key identifying assumption* in semiparametric selection models.

[2] He proposes that weakly tighter bounds than the unconditional bounds can be obtained by averaging the group-specific effects, weighted by covariate density. However, in practice, discretization is necessary for continuous covariates, and handling a large number of covariates is often infeasible. Semenova (2023) generalizes Lee's approach to a high-dimensional setting.

additional structural assumptions. Despite the growing popularity of partially identifying models, estimating identified sets and conducting inference remain challenging, particularly when the identified set is characterized by a large number of moment inequalities or the distributions of unobserved heterogeneity lack parametric restrictions.

Motivated by these challenges, this paper investigates semiparametric selection models that achieve point identification of β without any excluded variable. Specifically, we relax the linear selection assumption in HH's model, providing a middle ground between Lee's and HH's approaches. Unlike Lee's framework, our method does not impose selection monotonicity, meaning it is not nested within Lee's. We challenge the prevailing belief that an exclusion restriction is necessary for semiparametric selection models. Our approach establishes point identification of β under minimal assumptions without requiring scale normalization or identification at infinity, when there exists at least one continuous variable in X.

Our identification strategy leverages the nonlinearity of the conditional selection probability $p_0(X) := E[D|X]$, which is nonparametrically identified and easily verifiable in practice. Consequently, β can be estimated via a partial linear regression, plugging in estimates of $p_0(\cdot)$ in the nonparametric components approximated by sieves, without needing to estimate both $p_0(X)$ and $E[Y|X]$ nonparametrically. Unlike many existing methods, we do not assume unobserved heterogeneity is independent of regressors, allowing for heteroskedasticity. We demonstrate that our estimators for β are $\sqrt{n}$-consistent, semiparametrically efficient under homoskedasticity, asymptotically normal, and computationally scalable. When unobserved heterogeneity is heteroskedastic, robust standard errors can be computed. As we maintain the linearity of the outcome equation, incorporating a large set of covariates is straightforward. Our proposed two-step semiparametric estimator performs exceptionally well in simulations and an empirical application on gender and racial wage gaps in the United States (US).

We are not the first to consider nonlinearity in the selection process as a means to identify the linear index parameters in the latent outcome equation. For instance, Ahn and Powell (1993) and Newey and Powell (1993) speculated that nonlinearity can yield nonzero semi-

parametric efficiency bounds, as the nonlinear terms in the selection equation may act as excluded variables. In this paper, we formally establish the conditions which suffice to identify the model parameters. More recently, Escanciano et al. (2016) introduced a more general model, assuming $E[Y|X] = F_0(X\beta_0, p_0(X))$, and demonstrated that β_0 can be identified up to scale if X contains at least two continuous variables and $p_0(X)$ is nonlinear in X. They normalize one of the continuous variables' coefficients to 1, as the scale of β_0 is not separately identified from F_0. Unlike our proposed estimator, their approach requires nonparametric estimation of both $p_0(X)$ and $E[Y|X]$ even with linearity in the outcome equation, adding to its complexity.

Pan et al. (2022) propose an integrated nonlinear least squares estimator (Chen, 2010a,b, Chen and Zhou, 2011, 2012) for Escanciano et al. (2016)'s model, eliminating the need for scale normalization. However, they do not provide an identification argument, as they assume β_0 is already identified. Moreover, their estimation procedure relies on multiple layers of kernel regression, each requiring the selection of multiple tuning parameters, along with numerical optimization of an integrated criterion function. This results in a computational burden that scales exponentially with the sample size.

This paper is organized as follows. Section 2 introduces our semiparametric selection model and establishes identification. Section 3 presents our estimators and derives their asymptotic properties. Section 4 evaluates finite-sample performance of our estimator via simulations. Section 5 applies our method to estimating gender and racial wage disparities in the US. Section 6 concludes.

2. THE SEMIPARAMETRIC SELECTION MODEL

We consider a sample selection model where the selection procedure is left unspecified:

$$Y^* = \alpha_0 + X\beta_0 + V, \quad Y = D \cdot Y^*. \tag{3}$$

Let $p_0(x) := P[D = 1|X = x]$ represent the conditional selection probability given $X = x$. We impose the following assumption to identify β_0 without invoking an exclusion restriction.

ASSUMPTION 1: *(i) At least one variable (X_k) in X is continuously distributed; (ii) $p_0(X)$ is continuously differentiable with respect to X_k almost everywhere; (iii) $\partial p_0/\partial x_k \neq 0$ with probability 1; (iv) $E[Y|X, D=1] = X\beta_0 + \lambda_0(p_0(X))$; (v) $\lambda_0(p)$ is continuously differentiable almost everywhere.*

The above assumption requires a continuously distributed variable in X, the smoothness of $p_0(\cdot)$ and $\lambda_0(\cdot)$. It also restricts the selection bias to only depend on the selection probability in a nonparametric form, $\lambda_0(\cdot)$. Additionally, it rules out a flat region in $p_0(\cdot)$. Our model is implied by a special case where

$$D = \mathbb{1}[h_0(X) - \varepsilon \geq 0], \quad X \perp\!\!\!\perp (V, \varepsilon). \tag{4}$$

Suppose ε is continuously distributed with the c.d.f. $F_\varepsilon(\cdot)$. We can normalize the selection process to $D = \mathbb{1}[p_0(X) \geq U]$, where $p_0(X) = F_\varepsilon(h_0(X))$ and $U = F_\varepsilon(\varepsilon) \sim Unif(0,1)$. Unlike this special case, we do not impose the stochastic independence between X and (V, ε). For some β and $\lambda(\cdot)$, define $l(x) := (\beta_0 - \beta)x$ and $b(p) := \lambda_0(p) - \lambda(p)$, both of which are deviations of β and $\lambda(p)$ from the truth. For any observationally equivalent β and λ such that $E[Y|X, D=1] = X\beta + \lambda(p_0(X))$, $l(x) + b(p) = 0$ identically.

Under Assumption 1, β_0 can be point identified. We cannot separately identify the intercept α_0 from the selection bias.[3] Firstly, we consider the simplest case in which X consists of only one continuous variable.

PROPOSITION 1: *Let X be a scalar, nondegenerate, continuously distributed random variable. Let Assumption 1 (ii)–(v) hold. If there exist two distinct values x' and x'' in the support of X such that $p_0(x') = p_0(x'') = p'$, β_0 is identified and λ_0 is identified up to an additive constant.*

[3]For point identification of the intercept α_0, see Heckman (1990) and Andrews and Schafgans (1998). Unlike our paper, both papers relies on the"identification at infinity" argument and an exclusion restriction.

PROOF: For any observational equivalent β and $\lambda(\cdot)$ and for both x' and x'', $l(x') + b(p_0(x')) = l(x'') + b(p_0(x'')) = 0$. And therefore,

$$l(x') + b(p') - l(x'') - b(p') = (\beta_0 - \beta)(x' - x'') = 0.$$

As we assume $x' \neq x''$, β must be identical to β_0 so that β_0 is identified. Furthermore, by continuous differentiability of $p_0(\cdot)$ and $\lambda_0(\cdot)$ (Assumption 1 (ii)–(iii)),

$$\beta_0 - \beta + \frac{\partial b(p)}{\partial p}\frac{\partial p}{\partial x_k} = \frac{\partial b(p)}{\partial p}\frac{\partial p}{\partial x_k} = 0. \tag{5}$$

By Assumption 1(iv), $\partial p_0/\partial x_k \neq 0$ and hence $\partial b(p)/\partial p = 0$ implying that $b(p)$ is constant *i.e.*, $b(p) = C$ for an unknown constant C. This means $\lambda_0(p) = \lambda(p) + C$ so that $\lambda_0(\cdot)$ is identified up to a constant. *Q.E.D.*

This proposition shows that more than nonlinearity is required in the selection process, as it rules out monotonicity of $p_0(\cdot)$. Consider $h_0(\cdot)$ in the special case described above. If $h_0(\cdot) = X + 0.5X^2$, β_0 is not point identified. When X is binary, $p_0(X) = \gamma_0 + \gamma_1 X$ is fully nonparametric and therefore the parameters in the latent outcome equation are not point identified as shown in HH.

Now we consider more general cases where X is multidimensional. Suppose two elements of X, X_k and X_j, are continuously distributed. The following proposition shows that the model is point identifying.

PROPOSITION 2: *Let Assumption 1 hold. Assume there exists another continuously distributed element X_j of X, such that $p_0(\cdot)$ is continuously differentiable w.r.t. X_j and $\frac{\partial p_0}{\partial X_j} \neq 0$ with probability 1. If $\frac{\partial p_0}{\partial X_k} \not\propto \frac{\partial p_0}{\partial X_j}$, β_0 is identified and λ_0 is identified up to a constant.*

PROOF: Partially differentiating $l(x) + b(p) = 0$ w.r.t. x_k and x_j yields

$$\beta_{0k} - \beta_k + \frac{\partial b(p)}{\partial p}\frac{\partial p}{\partial x_k} = 0, \quad \beta_{0j} - \beta_j + \frac{\partial b(p)}{\partial p}\frac{\partial p}{\partial x_j} = 0. \tag{6}$$

This implies the following identity:

$$\frac{\partial b(p)}{\partial p} = \frac{\beta_k - \beta_{0k}}{\partial p/\partial x_k} = \frac{\beta_j - \beta_{0j}}{\partial p/\partial x_j} \Rightarrow \frac{\beta_k - \beta_{0k}}{\beta_j - \beta_{0j}} = \frac{\partial p/\partial x_k}{\partial p/\partial x_j}. \tag{7}$$

which only holds either $\partial p/\partial x_k \propto \partial p/\partial x_j$ or $\beta_k - \beta_{0k} = \beta_j - \beta_{0j} = 0$. Therefore, by ruling out $\partial p/\partial x_k \propto \partial p/\partial x_j$, we can conclude $\beta_k - \beta_{0k} = \beta_j - \beta_{0j} = 0$ so that $\partial b(p)/\partial p = 0$. As $b(p)$ is constant, the whole parameter vector β_0 is identified and $\lambda_0(\cdot)$ is identified up to a constant. *Q.E.D.*

In general, the marginal effect of x_k on p is not proportional to that of x_j. Identification fails in the special case (4) when $p_0(X) = F_\varepsilon(X\gamma)$ because $\frac{\partial p_0/\partial x_k}{\partial p_0/\partial x_j} = \frac{\gamma_k f_\varepsilon(X\gamma)}{\gamma_j f_\varepsilon(X\gamma)} = \gamma_k/\gamma_j$ where $F_\varepsilon(\cdot)$ and $f_\varepsilon(\cdot)$ are the c.d.f. and p.d.f. of ε. Therefore, the nonlinearity of $p_0(\cdot)$ enables us to identify the model parameters.

Lastly, we consider the case where X_k is the only continuous variable in X. Without loss of generality, let X_j be a binary variable for all $j \neq k$, as any discrete variable can be equivalently expressed as a set of dummy variables.

PROPOSITION 3: *Let Assumption 1 hold. For some X_j, let x' be a vector where $x_j = 1$, and let x'' denote an otherwise identical vector where $x_j = 0$. Further assume there exists x''' an otherwise identical vector to x' except x_k such that $p_0(x''') = p_0(x'')$. Unless $(x'_k - x'''_k)$ is constant for all values of x'_k, β_0 is identified and λ_0 is identified up to a constant.*

PROOF: Let p' and p'' denote $p_0(x')$ and $p_0(x'')$ respectively. Then,

$$l(x') + b(p') - l(x'') - b(p'') = \beta_{0j} - \beta_j + b(p') - b(p'') = 0. \tag{8}$$

Let p''' denote $p_0(x''')$. As $p''' = p''$,

$$l(x') + b(p') - l(x''') - b(p''') = (\beta_{0k} - \beta_k)(x'_k - x'''_k) + b(p') - b(p'') = 0. \tag{9}$$

Combining (8) and (9) yields $\beta_{0j} - \beta_j = (\beta_{0k} - \beta_k)(x'_k - x'''_k)$. By ruling out the possibility that $x'_k - x'''_k$ is constant across all values of x'_k, we can conclude that

$$\beta_{0k} = \beta_k, \quad \beta_{0j} = \beta_j, \quad \frac{\partial b(p)}{\partial p} = 0.$$

Hence, β_0 is identified and λ_0 is identified up to a constant. *Q.E.D.*

Unless the required change in x_k to achieve p'' with $x_j = 0$ from p' with $x_j = 1$ is constant for all values of x'_k, the model parameters are identified. This proposition rules out the case in which $p_0(X) = F_\varepsilon(X\gamma)$ because $p'' = F_\varepsilon(x''\gamma) = F_\varepsilon(x'\gamma - \gamma_j) = F_\varepsilon\left(x'\gamma + (x'''_k - x'_k)\gamma_k\right)$ and hence $x'_k - x'''_k = \gamma_j/\gamma_k$ for all x'_k.

REMARK 1: (Parameter heterogeneity) We do not explicitly allow the parameters vary across individuals. In applied studies, heterogeneous treatment effects are often concerned. As briefly discussed in Honoré and Hu (2024), it is straightforward to extend our results to allow for treatment heterogeneity. Modifying the model (3) as:

$$Y_i^* = \alpha_i + X_i\beta_i + V_i, \quad Y_i = D_i \cdot Y_i^*, \quad p_i = E[D_i|X_i], \tag{10}$$

We allow the parameter vector β_i to be individual-specific. If we assume $(\alpha_i, \beta_i) \perp\!\!\!\perp (X_i, D_i, V_i)$ following Honoré and Hu (2024), we yield

$$E[Y_i|D_i = 1, X_i] = E[\alpha_i] + X_i E[\beta_i] + \lambda^*(p_i).$$

Therefore, under parameter heterogeneity, our model (3) identifies $E[\beta_i]$ and $\lambda(\cdot) := \lambda^*(\cdot) + E[\alpha_i]$.

The identification results in this section show that our semiparametric selection model can point identify the model parameters without an excluded variable as long as there is at least one continuously distributed covariate. In applied economic studies, continuous variables such as age, income, and price are not uncommon. Hence, our semiparametric model can be quite generally applicable to many modern data sets. Furthermore, it is natural that the true selection process exhibit some degree of nonlinearity. As the conditional selection

probability is nonparametrically identified, it is simple to check the nonlinearity in the selection probability. When there exists strong empirical evidence of selection nonlinearity, our model becomes a strong alternative to the HH's model, as our model is more robust and point-identifying.

3. THE ESTIMATORS

Given the identification results, we now propose a class of tractable two-step plug-in semiparametric estimators. We first begin by demonstrating that these estimators are consistent, semiparametrically efficient, and asymptotically normal when the individual selection probability, $p_i = p_0(X_i) = \mathbb{E}[D_i|X_i]$, is observed. Subsequently, we prove that replacing p_i with $\hat{p}_i$, a consistent estimator of p_i, does not affect the asymptotic behavior of our estimators when $\hat{p}_i$ converges to p_i at a sufficiently fast rate as $n \to \infty$. Then we describe how the estimator can be practically implemented.

3.1. *When p_i is known*

Suppose we have an i.i.d. sample, $\{W_i\}_{i=1}^n$, where $W_i := (Y_i, D_i, X_i, p_i)$. As p_i is observed, the model parameters $\theta_0 = (\beta_0, \lambda_0) \in \Theta := \mathcal{B} \times \Lambda$ can be estimated by the least squares (LS) procedure:

$$\tilde{\theta}_n = \left(\tilde{\beta}_n, \tilde{\lambda}_n\right) = \underset{(\beta,\lambda)\in\mathcal{B}\times\Lambda}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} D_i \left(Y_i - X_i'\beta - \lambda(p_i)\right)^2.$$

As λ is infinite-dimensional, we approximate the unknown function $\lambda \in \Lambda$ by sieves, $\lambda_n \in \Lambda_n$, where Λ_n is an approximating function space (such as polynomials, trigonometric polynomials, splines, and orthogonal wavelets) that becomes dense in Λ as $n \to \infty$.

Let $\Lambda_n = \left\{\lambda_n(\cdot) = R_{K(n)}(\cdot)'\gamma : \gamma \in \mathbb{R}^{K(n)}\right\}$, where $R_{K(n)} = \left[r_1(\cdot), \ldots, r_{K(n)}(\cdot)\right]'$ denote a vector of basis functions. Let $\Theta_n = \mathcal{B} \times \Lambda_n$ be the sieve space for $\theta = (\beta, \lambda(\cdot))$ and $\bar{K}(n) = \dim(\beta) + K(n)$. For notational simplicity, we omit the subscript $K(n)$ and write $R_{K(n)}(\cdot) = R(\cdot)$. Define

$$D = diag(D_1, \ldots, D_n), \quad X = [X_1, \ldots, X_n]', \quad y = (Y_1, \ldots, Y_n)', \quad p_0 = (p_1, \cdots, p_n)',$$

For an arbitrary vector $\tilde{p} := (\tilde{p}_1, \cdots, \tilde{p}_n)$ where $\tilde{p}_i \in [0,1]$, define $R(\tilde{p}) = D[R(\tilde{p}_1), \ldots, R(\tilde{p}_n)]'$, and $Q(\tilde{p}) = R(\tilde{p})\left(R(\tilde{p})' R(\tilde{p})\right)^{-1} R(\tilde{p})'$. Then the LS estimator of β is written as:

$$\tilde{\beta}_n = \left((DX)'(I - Q(p_0))(DX)/n\right)^{-1} (DX)'(I - Q(p_0))(Dy)/n.$$

We establish asymptotic properties of our sieve LS (SLS) estimator using the results in Chen (2007). Let Θ be equipped with a norm $\|\theta\|_s = |\beta|_e + \|\lambda\|_\infty$, where $|\cdot|_e$ denotes the Euclidean norm and $\|\lambda\|_\infty = \sup_{p\in[0,1]} |\lambda(p)|$ is the supremum norm. We introduce the Hölder class of functions. Let $[m]$ be the largest nonnegative integer such that $[m] < m$. A real-valued function λ on $[0,1]$ is said to be in the Hölder space $\Lambda^m([0,1])$ if it is $[m]$ times continuously differentiable on $[0,1]$ and

$$\max_{\ell \le [m]} \sup_p \left| \frac{\partial^\ell \lambda(p)}{\partial p^\ell} \right| + \sup_{p,p'} \left| \frac{\partial^{[m]} \lambda(p)}{\partial p^{[m]}} - \frac{\partial^{[m]} \lambda(p')}{\partial p^{[m]}} \right| / \left|p - p'\right|^{m-[m]}$$

is finite. We impose the following conditions to derive asymptotic properties of the SLS estimator.

ASSUMPTION 2: *(i)* $\{W_i\}_{i=1}^n$ *are i.i.d.; (ii) the support of* X_i, $\mathcal{X}$, *is compact; (iii) the density of* p_i *is bounded and bounded away from zero on the compact subset of* $[0,1]$.

ASSUMPTION 3: *(i)* $\lambda \in \Lambda^m([0,1])$ *with* $m > 1/2$*; (ii)* $\forall \lambda \in \Lambda^m([0,1]), \exists \lambda_n(p;\gamma) \in \Lambda_n$ *such that* $\|\lambda_n - \lambda\|_\infty = O\left(K(n)^{-m}\right)$ *with* $K(n) = O\left(n^{1/(2m+1)}\right)$.

ASSUMPTION 4: $\sigma_0^2(X_i, D_i) := \mathbb{E}\left[D_i (Y_i - X_i'\beta_0 - \lambda_0(p_i))^2 | X_i = x, D_i = 1\right]$ *are positive and bounded uniformly over* $x \in \mathcal{X}$.

ASSUMPTION 5: $\Theta := \mathcal{B} \times \Lambda$ *is compact under* $\|\cdot\|_s$.

Under these regularity conditions, the consistency of the estimator is obtained by Proposition 3.3 of Chen (2007) as $\|\tilde{\theta}_n - \theta_0\| = O_p\left(n^{-m/(2m+1)}\right)$.

We now show that the parametric components of the SLS estimator, $\tilde{\beta}_n$, is asymptotically normal. Let $\tilde{X}_i = D_i X_i - \mathbb{E}[D_i X_i | p_i, D_i = 1]$.

ASSUMPTION 6: *(i)* $\mathbb{E}\left[\tilde{X}_i'\tilde{X}_i\right]$ *is positive definite; (ii) each element of* $\mathbb{E}\left[D_i X_i | p_i, D_i = 1\right]$ *belongs to the Hölder space* $\Lambda^{m_j}\left([0,1]\right)$ *with* $m_j > 1/2$ *for* $j = 1, \ldots, dim(\beta)$.

ASSUMPTION 7: $\beta_0 \in \text{int}\left(\mathcal{B}\right)$.

Assumption 6(i) is satisfied when β_0 and λ_0 are identified. Applying Proposition 4.5 of Chen (2007), we obtain the following asymptotic normality of $\tilde{\beta}_n$.

PROPOSITION 4: *Let Assumptions 2–7 hold. Then,*

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) \to_d N\left(0, \mathbb{E}\left[\tilde{X}_i'\tilde{X}_i\right]^{-1} \mathbb{E}\left[\sigma_0^2\left(X_i, D_i\right)\tilde{X}_i'\tilde{X}_i\right] \mathbb{E}\left[\tilde{X}_i'\tilde{X}_i\right]^{-1}\right).$$

If the error term is homoskedastic i.e., $\sigma_0(X_i, D_i)$ is constant, the SLS estimator $\tilde{\beta}_n$ is semiparametrically efficient. When the error term exhibits heteroskedasticity, efficient estimation can be achieved through the sieve generalized least squares (SGLS) estimator. However, in applied economic studies, the standard practice is to report heteroskedasticity-robust standard errors rather than employing the GLS approach. Consequently, in our simulations and empirical applications, we proceed with robust standard errors calculated using the asymptotic variance formula reported in Proposition 4.

3.2. *When* p_i *is replaced by* $\hat{p}_i$

As p_i is never observed in practice, $\tilde{\beta}_n$ is an infeasible estimator. Suppose p_i is consistently estimated by an estimator $\hat{p}_i := \hat{p}_n\left(X_i\right) = p_0\left(X_i\right) + O_p(n^{-1/3})$. Define $\hat{p} = (\hat{p}_1, \cdots, \hat{p}_n)'$. Replacing p_0 with $\hat{p}$ in $\tilde{\beta}_n$, we yield the following feasible estimator:

$$\hat{\beta}_n = \left((DX)'\left(I - Q\left(\hat{p}\right)\right)(DX)/n\right)^{-1}(DX)'\left(I - Q\left(\hat{p}\right)\right)(Dy)/n.$$

We will show that $\hat{\beta}_n$ and $\tilde{\beta}_n$ have the same asymptotic distribution by leveraging the results in Song (2012, 2014). Define

$$\hat{b}_n = \left[(DX)'(Dy)/n; (DX)'(DX)/n\right],$$
$$\hat{a}_n\left(\tilde{p}\right) = \left[(DX)'Q\left(\tilde{p}\right)(Dy)/n; (DX)'Q\left(\tilde{p}\right)(DX)/n\right].$$

Let d_X denote $\dim(X)$ and define $H(a,b) = (b_2 - a_2)^{-1}(b_1 - a_1)$ where a_2 and b_2 are the right $d_X \times d_X$ subblocks, and a_1 and b_1 are the left $d_X \times 1$ subblocks of a and b. Then, we can write $\|\hat{\beta}_n - \tilde{\beta}_n\|$ as $\|H\left(\hat{a}_n(\hat{p}), \hat{b}_n\right) - H\left(\hat{a}_n(p_0), \hat{b}_n\right)\|$. From the continuously differentiability of H, we have

$$\|H\left(\hat{a}_n(\hat{p}), \hat{b}_n\right) - H\left(\hat{a}_n(p_0), \hat{b}_n\right)\| \leq C\|\hat{a}_n(\hat{p}) - \hat{a}_n(p_0)\| + o_p\left(\|\hat{a}_n(\hat{p}) - \hat{a}_n(p_0)\|\right).$$

For the RHS of the above inequality, observe that

$$\hat{a}_n(\hat{p}) - \hat{a}_n(p_0) = \underbrace{\hat{a}_n(\hat{p}) - a_0(\hat{p})}_{:=A_n(\hat{p})} - \underbrace{\{\hat{a}_n(p_0) - a_0(p_0)\}}_{:=A_n(p_0)} + \underbrace{a_0(\hat{p}) - a_0(p_0)}_{:=B_n}, \tag{11}$$

where $a_0(p) := \mathbb{E}[D_i X_i \mathbb{E}[Z_i|p_i, D_i = 1]]$ with $Z_i = [Y_i; X_i]$.

As $\|\hat{p} - p_0\| = o_p(1)$, $A_n(\hat{p})$ and $A_n(p_0)$ in (11) can be shown to be $O_p\left(n^{-1/2}\right)$. First, we observe that

$$\begin{aligned} A_n(\tilde{p}) =& \frac{1}{n}\sum_{i=1}^{n} D_i X_i \left[D_i R(\tilde{p}_i)\left(R(\tilde{p})' R(\tilde{p})\right)^{-1} R(\tilde{p})' DZ - D_i \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1]\right] \\ &+ \frac{1}{n}\sum_{i=1}^{n} \{D_i X_i \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1] - \mathbb{E}[D_i X_i \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1]]\}. \end{aligned} \tag{12}$$

We obtain the asymptotic linear representation of $\sqrt{n}A_n(\tilde{p})$ as follows:

$$\sqrt{n}A_n(\tilde{p}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_i(\tilde{p}) + o_p(1), \tag{13}$$

for some i.i.d. $\psi_i(\cdot)$ such that $\mathbb{E}[\psi_i(\tilde{p})] = 0$, where $o_p(1)$ is uniform local around p_0. Then, applying the maximal inequality yields the stochastic equicontinuity of $A_n(\cdot)$ as shown in Andrews (1994). By Lemma B3 in Song (2014), we rewrite the first element of the RHS of (12) to:

$$\frac{1}{n}\sum_{i=1}^{n} D_i \mathbb{E}[X_i|\tilde{p}_i, D_i = 1](Z_i - \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1]) + o_p\left(n^{-1/2}\right),$$

uniformly over $\tilde{p} \in B(p_0; c_n) := \{\tilde{p} : \|\tilde{p} - p_0\| < c_n\}$, which implies that $\sqrt{n}A_n(\tilde{p})$ is, uniformly over $p \in B(p_0; c_n)$, equal to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \mathbb{E}[X_i|\tilde{p}_i, D_i = 1](Z_i - \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1])$$
$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{D_i X_i \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1] - \mathbb{E}[D_i X_i \mathbb{E}[Z_i|\tilde{p}_i, D_i = 1]]\} + o_p(1).$$

From this uniform linear representation, we have $\sup_{\tilde{p} \in B(p_0; c_n)} |\sqrt{n}A_n(\tilde{p})| = O_p(1)$, which means that both $A_n(\hat{p})$ and $A_n(p_0)$ are $O_p(n^{-1/2})$. Furthermore, we have

$$\sqrt{n}(A_n(\hat{p}) - A_n(p_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\psi_i(\tilde{p}) - \psi_i(p_0)] + o_p(1), \tag{14}$$

whose asymptotic variance, $\mathbb{E}\left[(\psi_i(\tilde{p}) - \psi_i(p_0))^2\right]$, goes to 0 as $\tilde{p} \to p_0$ under minor regularity conditions for $\psi_i(\tilde{p})$. This result also works when $\tilde{p} = \hat{p}$ as shown in Lemma 1 provided in the appendix. Hence we conclude $A_n(\hat{p}) - A_n(p_0) = o_p\left(n^{-1/2}\right)$ as $\hat{p} \to p_0$.

Lastly, we show that the term B_n in (11) is also $o_p(n^{-1/2})$. We extend Song (2014)'s results, which are derived under the linear selection procedure, to the cases with nonlinear, possibly nonmonotone selection. Under regularity conditions, the function $a_0(\tilde{p})$ is sufficiently smooth in $\tilde{p}$ around p_0 so that there exist constants $C > 0$ and $\varepsilon \in (0, 1/2]$ such that for each $\eta \in (0, \varepsilon]$,

$$\sup_{\tilde{p} \in B(p_0; \eta)} \|a_0(\tilde{p}) - a_0(p_0)\| \leq C\eta^2. \tag{15}$$

The formal proof of (15) is provided in Lemma 2 in the appendix. Hence, $\|a_0(\hat{p}) - a_0(p_0)\| = O_p\left(\eta_n^2\right)$ if $\|\hat{p} - p_0\| \leq \eta_n$. As we consider $\hat{p}$ converging to p_0 at a cube-root rate, we obtain $\|a_0(\hat{p}) - a_0(p_0)\| = o_p\left(n^{-1/2}\right)$ by taking $\eta_n = n^{-1/3} \log n$. This concludes that $\|\hat{\beta}_n - \tilde{\beta}_n\| = o_p\left(n^{-1/2}\right)$, which implies that $\hat{\beta}_n$ and $\tilde{\beta}_n$ have the same asymptotic distribution.

3.3. *Practical implementations*

Given the asymptotic results, one can consider a wide range of estimators in the first stage estimation of $p_i = P[D_i = 1|X_i]$. Any consistent nonparametric estimator $\hat{p}_n(\cdot)$ that converges to p_0 at a cube-root rate or faster can be employed. The convergence rate of the first stage estimation depends on the smoothness of p_0 and the number of continuous elements in X_i, denoted as d_c. With a relatively low d_c, it is quite feasible that standard kernel or sieve estimators converge faster than the $n^{-1/3}$ rate. In a high-dimensional setting, a high level of smoothness for p_0 is necessary to ensure a sufficiently fast rate. Recall that m denotes the Hölder smoothness of p_0. The rate condition is satisfied when $m > d_c$. Suppose that the true selection process is $D_i = \mathbb{1}[g_k(X_i) \geq U_i]$ where $g_k(\cdot)$ is a k-th order polynomial of X_i and U_i is continuously distributed with the c.d.f. $F_U(\cdot)$ which belongs to a smooth parametric class. Consequently, $p_0(X_i) = F_U(g_k(X_i))$ and $p_0(\cdot)$ is infinitely continuously differentiable, implying $m = \infty$. Therefore, if one would like to impose the selection procedure outlined above, standard kernel or sieve methods can be employed with a high d_c. Alternatively, additional structural restrictions can be imposed, such as additivity, i.e. $p_0(X_{1i}, X_{2i}) = p_{10}(X_{1i}) + p_{20}(X_{2i})$.

In practical implementation, we propose the following simple two-step procedure.

- Step 1: using the full sample, estimate p_i using an appropriate nonparametric estimator and obtain the fitted values $\hat{p}_i$.
- Step 2: conditional on $D_i = 1$, estimate β_0 and λ_0 using the SLS estimator.

In the simulations and empirical studies conducted in this paper, we employ the sieve maximum likelihood estimator in Step 1, using piecewise polynomial basis functions. In Step 2, we regress Y_i on X_i and piecewise polynomial transformations of $\hat{p}_i$ using ordinary least squares (OLS). The OLS standard errors of $\hat{\beta}$ provided in standard statistical programs such as Stata, R, and Matlab are asymptotically valid standard errors under homoskedasticity. In practice, researchers often desire heteroskedasticity-robust or cluster-robust standard errors. The same 'sandwich' formula can be used to compute robust standard errors. Therefore, the two-step procedure outlined here can be readily utilized in any statistical software

without necessitating a new implementation package, which makes our proposal particularly attractive to applied researchers.

REMARK 2: (Nonlinearity test) Nonlinearity in the first stage can be empirically tested. When sieve MLE is employed in the first stage $Y_i = \mathbb{1}\big[\, g(X_i) + U_i > 0\big]$ where $g(X_i)$ is approximated by $\sum_{k=1}^{K} \gamma_k \, \phi_k(X_i)$, significant coefficients of high-order sieve terms imply nonlinearity. Alternatively, one can also consider linear index specifications such as probit and logit. Let l_{lin} and l_{sieve} denote maximized log likelihoods of linear index and sieve-based models respectively. Under the null hypothesis $H_0 : g(X) = X'\beta$, the distribution of $l_{LR} = 2(l_{sieve} - l_{lin})$ is asymptotically $\chi^2_{K-d_X}$ (assuming K is fixed). Reject H_0 (and thus reject linearity) if $l_{LR} > \chi^2_{K-d_X,\alpha}$ where $\chi^2_{K-d_X,\alpha}$ is the critical value of the chi-square distribution with $K - d_X$ degrees of freedom at the significance level α.

4. SIMULATIONS

In this section, we evaluate the finite sample performances of our semiparametric estimator using known data-generating processes (DGPs). For each DGP, we repeat 1,000 iterations in each of which we draw a Monte Carlo sample of size $n = 5,000$. We first investigate the single-covariate case using the following DGP:

$$Y = D \cdot (\beta_0 + X_1\beta_1 + 2 \cdot V), \quad D = \mathbb{1}[\alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3 + U \geq 0], \tag{16}$$

$$X_1 \sim N(0,1), \quad \begin{bmatrix} V \\ U \end{bmatrix} | X_1 \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix} \right). \tag{17}$$

In this instance, β_0 is not separately identified from $E[V|X, D = 1] = \lambda_0(p_0)$. The identification of β_1 is contingent upon the parameter values $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)$. This is because the conditional selection probability $p_0(X) = E[D|X]$ must not exhibit strict monotonicity. We employ the proposed two-step sieve-based approach to estimate the model. For alternative estimators, we consider the ordinary least squares (OLS) estimator conditional on selection ($D = 1$) assuming random selection, which is commonly referred to as the two-part model (TPM), and the maximum likelihood estimator under the Heckman selection model (HSM), both of which are misspecified. We also compare our estimator to the oracle

estimator, which incorporates the true functional form of $p_0(X)$ given the selection bias is expressed using the inverse Mills ratio.[4]

We consider two selection designs: (a) $\alpha = (0.6, 1.50, -0.5, -0.05)$ and (b) $\alpha = (0.4, 1.50, 0.2, 0.05)$. In both designs, the parameter values are chosen to ensure that the selection probability, $P(D = 1)$, is approximately 60%. Let $h(X) := \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3$ be the selection index. The shape of h is displayed in Figure 1 and the performances of estimators are reported in Table I and Figure 2. Under design (a), $h(\cdot)$ is not monotone, so our two-step sieve estimator for β_1 is well centered around the true value. Its root-mean-squared error (RMSE) is close to that of the oracle estimator. Conversely, with design (b), β_1 remains unidentified so it suffers from a large RMSE as expected. In both designs, the OLS exhibits substantial misspecification bias. The Heckman's MLE performs poorly in the non-monotone design due to misspecification but performs very well in the monotone design. This is because the selection index is close to linear in the effective support of X and the error distribution is correctly specified in the monotone design.

TABLE I

FINITE SAMPLE PERFORMANCES OF ESTIMATORS (SINGLE COVARIATE CASE)

	Non-monotone selection				Monotone selection			
	TPM	HSM	Kim&Lee	Oracle	TPM	HSM	Kim&Lee	Oracle
RMSE	0.524	0.109	0.083	0.071	0.693	0.059	0.354	0.096
Bias	-0.522	0.092	-0.001	-0.004	-0.692	-0.015	-0.001	-0.003

We next generate Monte Carlo samples from the following DGP (referred to as DGP1 henceforth), where X consists of two continuously distributed variables and the unobservables are joint normally distributed as (17):

$$Y = D \cdot (\beta_0 + X_1\beta_1 + X_2\beta_2 + 2 \cdot V), \tag{18}$$

$$D = \mathbb{1}[\alpha_0 + \alpha_1 X_1 + \alpha_2 X_1^2 + \alpha_3 X_1^3 + \alpha_4 X_1 X_2 + \alpha_5 X_2 + \alpha_6 X_2^2 + U \geq 0]. \tag{19}$$

[4]In the oracle estimation, we first estimate α using probit regression of D on $Z = (1, X_1, X_1^2, X_1^3)$. Subsequently, we employ $\lambda_0(\hat{p}_0(Z)) = \phi(Z\hat{\alpha})/\Phi(Z\hat{\alpha})$ to correct the selection bias.

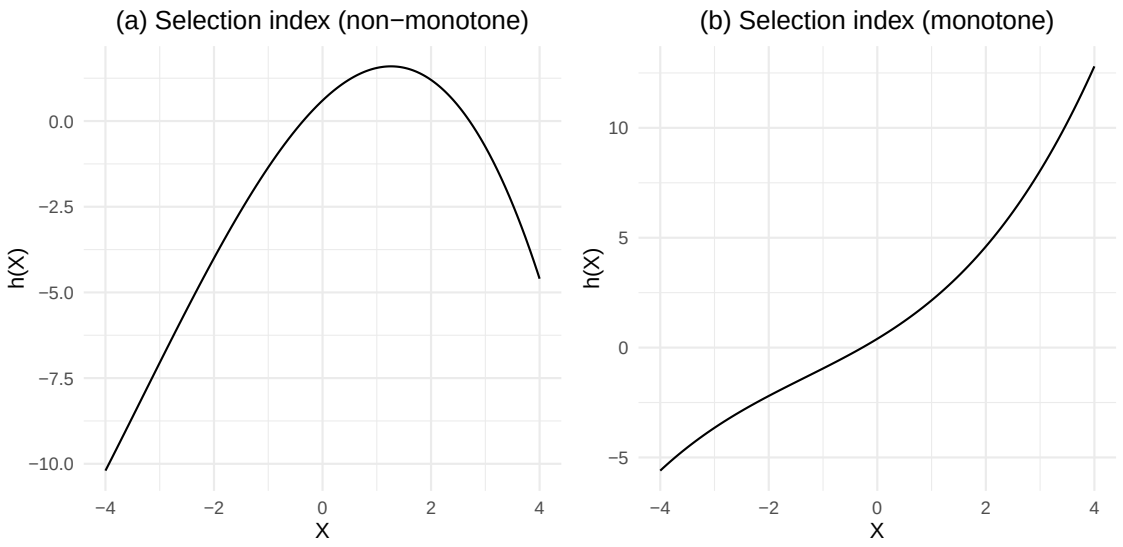


FIGURE 1.—Selection index designs

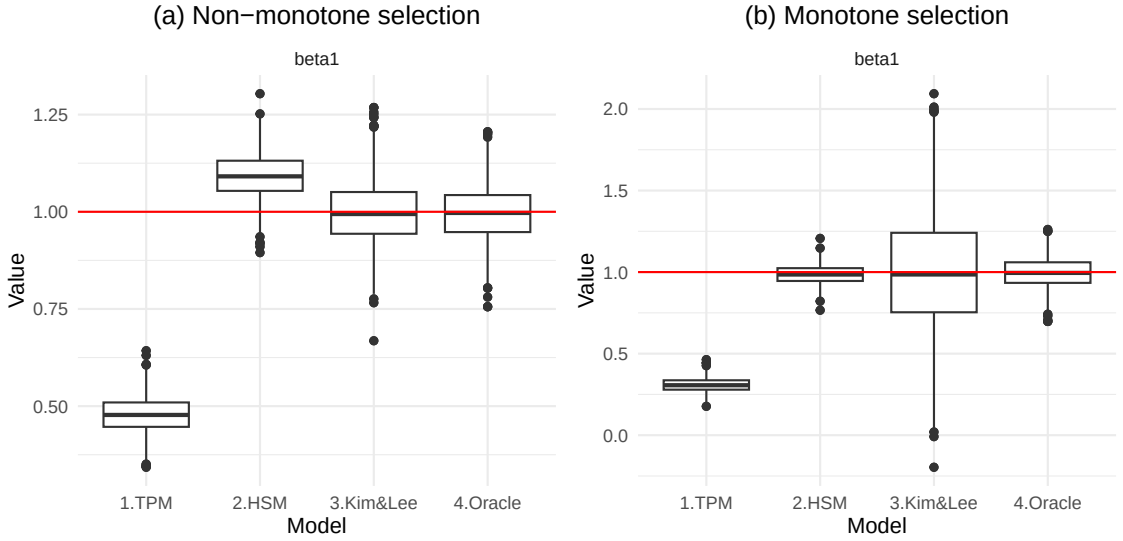


FIGURE 2.—Finite sample performances of Estimators: Single covariate case

Note: This figure shows a box plot for each parameter estimator. The main rectangular box shows the interquartile range (IQR). The thick line inside the box represents the median. Whiskers reach to the furthest data points within $1.5\times$ IQR. Dots beyond the whiskers are potential outliers. The red horizontal line indicates the true parameter value.

X_1 and X_2 are drawn from the standard normal distribution and independent of each other. The parameter values are set as:

$$\alpha = (1.5, 0.5, -0.5, 0.2, 0.5, 1.0, -0.5), \quad \beta = (0.5, 0.5, 0.25).$$

The average selection probability across Monte Carlo samples is 52%.

In the latest design (DGP2 henceforth), we consider the scenario where X consists of a continuously distributed variable, $X_1 \sim N(0,1)$, and a binary variable, $X_2 \sim Bernoulli(0.5)$. The remaining elements of DGP2 are otherwise identical to DGP1 except the selection process:

$$D = \mathbb{1}[\alpha_0 + \alpha_1 X_1 + \alpha_2 X_1^2 + \alpha_3 X_1^3 + \alpha_4 X_1 X_2 + \alpha_5 X_2 + \alpha_6 X_1^2 X_2 + \alpha_7 X_1^3 X_2 + U \geq 0].$$

The parameter values are set as:

$$\alpha = (0.2, -0.2, -0.5, 0.3, 0.1, 0.5, -0.3, 0.2), \quad \beta = (0.5, 0.5, 0.25).$$

The average selection probability is 66% under this DGP.

In both DGPs, we have at least one continuous covariate and the selection probability function $p_0(\cdot)$ exhibits sufficient nonlinearity, so our model point identifies β_1 and β_2. As showcased in Figures 3-4, the TPM and the Heckman selection model are misspecified and hence the OLS and MLE suffer from large bias for both DGPs. Table II displays the RMSE and mean bias of each estimator. Heckman's MLE works particularly badly in DGP2. In contrast, our semiparametric estimator performs exceptionally well in DGP1 for both parameters as the oracle estimator outperforms our estimator by a very slight margin in terms of root-mean-squared errors (RMSE) and mean bias. In DGP2, it performs similarly to the oracle estimator for β_1, but shows a larger RMSE (0.130) than the oracle estimator (0.085) for β_2, possibly due to limited variations in X_2.

Finally, we evaluate the performance of Lee (2009)'s and Honoré and Hu (2020)'s bounds approaches using DGP2. We do not consider DGP1 because there is no binary treatment variable for which Lee's bounds are applicable. We use a sample size of 100,000 instead of 5,000, which we use for point estimators. This is because the HH bounds are

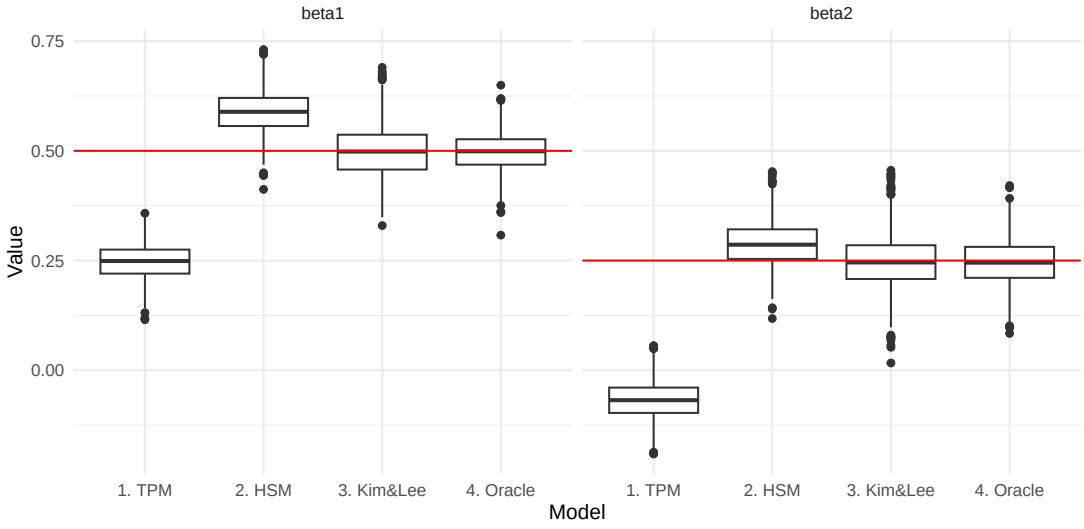


FIGURE 3.—Comparison of Estimators: DGP1 (selection probability = 0.52)

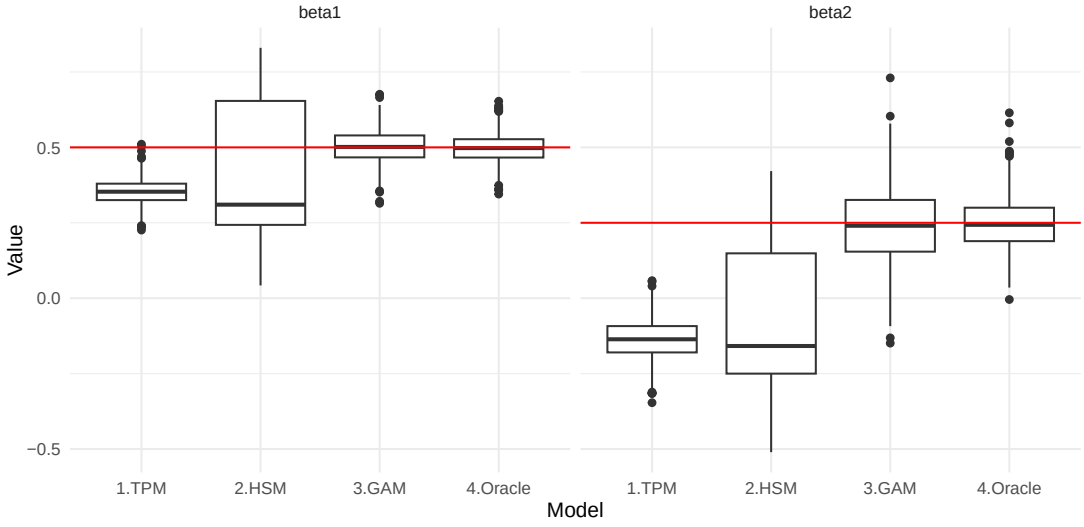


FIGURE 4.—Comparison of Estimators: DGP2 (selection probability = 0.66)

Note: This figure shows a box plot for each parameter estimator. The main rectangular box shows the interquartile range (IQR). The thick line inside the box represents the median. Whiskers reach to the furthest data points within $1.5\times$ IQR. Dots beyond the whiskers are potential outliers. The red horizontal line indicates the true parameter value.

not reliably estimated with a moderate sample size, with which the bounds are often empty (in 93 iterations out of 1,000) or uninformative (including zero within the bounds in 615 iterations out of 1,000). With the 100,000 sample size, both bounds are reliably estimated.

TABLE II

FINITE SAMPLE PERFORMANCES OF ESTIMATORS

		DGP1				DGP2			
		TPM	HSM	Kim&Lee	Oracle	TPM	HSM	Kim&Lee	Oracle
RMSE	β_1	0.255	0.100	0.060	0.045	0.153	0.224	0.056	0.047
	β_2	0.321	0.065	0.063	0.051	0.393	0.383	0.130	0.085
Bias	β_1	-0.252	0.088	-0.001	-0.003	-0.148	-0.078	0.002	-0.003
	β_2	-0.318	0.038	-0.002	-0.003	-0.388	-0.318	-0.008	-0.004

Figure 5 displays the box plots of HH's and Lee's bounds. It is not surprising to observe that the Lee bounds consistently contain the true parameter value for β_2 because the bounds are very wide in this setup. The Lee bounds are never informative about the sign of the treatment effect as they include zero in every simulation under DGP2. In contrast, the HH bounds are significantly tighter than the Lee bounds. However, in most iterations, the HH bounds are not informative and never contain the true value because the model misspecifies the selection process.

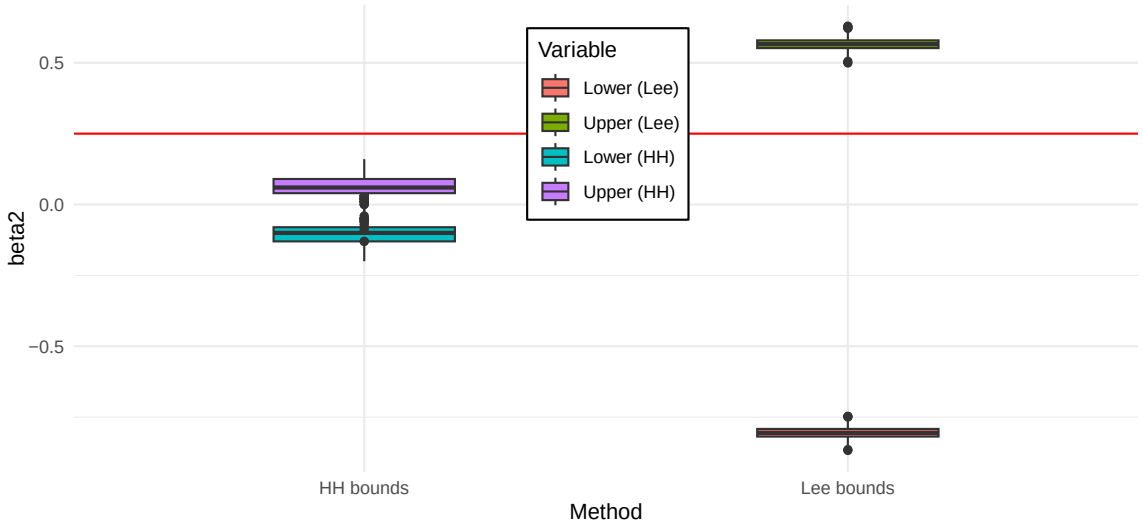


FIGURE 5.—Honore and Hu (HH) bounds and Lee bounds on DGP2 $(n = 100,000)$

These simulation exercises clearly demonstrate the practical usefulness of our semiparametric estimator. When at least one continuous covariate is present and the selection process exhibits nonlinearity, our estimator performs exceptionally well even with a modest sample size. The first-stage selection procedure is nonparametrically identified so assessing the nonlinearity in the selection equation is practically easy. In contrast to HH's partially identifying linear selection model, our semiparametric model offers greater flexibility by not imposing linearity in the first stage while still point-identifying the parameters of interest. Consequently, our estimator can serve as a valuable alternative when the Lee bounds are excessively wide to provide meaningful insights. However, if there is no continuous variable or the selection process is genuinely linear, the HH bounds would be an excellent alternative to the Lee bounds.

5. EMPIRICAL APPLICATION: GENDER AND RACIAL WAGE GAPS IN THE US

We now demonstrate the empirical usefulness of our semiparametric model and its estimator using real-world data. We estimate the gender and racial wage disparities in the US. The reservation wage varies between different genders and ethnicities. Upon selection into employment, the distribution of unobserved factors can differ from that of the unemployed. Therefore, the effect of sample selection on observed wages should be taken into account to accurately calculate wage gaps. Following Mora (2008) and Honoré and Hu (2020) where they focus on racial wage gaps, we analyze Current Population Survey (CPS) data on wages from Arizona, California, New Mexico, and Texas. The data set covers the years 2003–2016 and includes 129,907 women. Among them, 26,698 are third-generation Mexican-Americans, while 103,209 are non-Hispanic whites. The remaining 118,418 men comprise 21,402 third-generation Mexican-Americans and 97,016 non-Hispanic whites. All individuals in the sample are aged between 25 and 62. In terms of employment, the percentage of women working is 64% for third-generation Mexican-Americans and 61% for non-Hispanic whites. The employment rates for men are 71% for Mexican-Americans and 67% for non-Hispanic whites, respectively.

The gender wage gap is estimated for Mexican-Americans and non-Hispanic whites separately to nonparametrically control for ethnicity. We use the log inflation-adjusted hourly

wage as the outcome variable. In the latent outcome equation, we estimate the coefficient on the female dummy with age, age squared, experience, experience squared, education dummies (less than high school, some college, college, and advanced degree such as master's and doctorate, with high school as the base category), dummies for being a veteran and being married, state dummies (New Mexico as a base state), and year dummies as control variables. Age and experience serve as continuously distributed covariates in the selection equation for our semiparametric model. For the racial wage gap, we estimate the coefficient on the Mexican-American dummy with the same set of control variables separately for men and women.

We first estimate the Lee and HH bounds. As the Lee bounds are fully nonparametric, we compute the bounds conditional on the education level (high school and college) with no other covariates. For the HH bounds, we use the full set of control variables. We closely follow Honoré and Hu (2020)'s implementation except that we employ probit regression in the first stage estimation of selection parameters in lieu of logit. The results are still very similar to the original results of HH with the logit first stage. Table III presents the estimated bounds. For the racial wage disparities, the Lee bounds are not informative for college graduates, as they contain zero. For high school graduates, the bounds range from -25% to -7.5% for men and from -21% to -4.1% for women. In contrast, the HH bounds are highly informative and significantly narrower than the Lee bounds. The HH bounds range between -11.4% and -10.3% for men and between -8.9% and -6.6% for women. Regarding the gender wage gap, the Lee bounds suggest substantially lower wages for females, *ceteris paribus*. The bounds are wider for high school graduates (-32.5% to -10.5% for Mexican-Americans and -37.2% to -14.2% for whites). For college graduates, the bounds lie between -24.0% and -15.2% for Mexican-Americans, and between -28.4% and -11.9% for whites, indicating a potentially smaller gender wage gap among college graduates. The HH bounds for the gender wage gaps (-21.9% to -14.4% for Mexican-Americans and -22% to 17% for whites) are narrower than the Lee bounds but not as tight as for the racial gaps.

For point estimators, like in the simulation experiments, we consider the two-part model ("TPM") using the OLS conditional on employment assuming random selection, the Heck-

TABLE III

ESTIMATED LEE (2009) AND HH (2020) BOUNDS FOR RACIAL AND GENDER WAGE GAPS

Source	Category	Racial Gap	Category	Gender Gap
Lee	Men, Highschool	[-0.249, -0.074]	Mexican, Highschool	[-0.325, -0.105]
	Women, Highschool	[-0.210, -0.041]	White, Highschool	[-0.372, -0.142]
	Men, College	[-0.205, 0.015]	Mexican, College	[-0.240, -0.152]
	Women, College	[-0.235, 0.035]	White, College	[-0.284, -0.119]
HH	Men	[-0.114, -0.103]	Mexican	[-0.219, -0.144]
	Women	[-0.089, -0.066]	White	[-0.220, -0.170]

man selection two-step estimator ("HS 2step") and MLE ("HS MLE"), and our proposed semiparametric two -step estimator ("KL"). In the first stage estimation for the selection probability, we employ the sieve maximum likelihood estimator and predict $\hat{p}_0(\cdot)$, by including piecewise-polynomial (cubic b-spline) basis functions of age and experience with 5 knots, and their interactions with dummy variables. Most coefficients on sieve terms in the first stage estimation are highly significant across all the subsamples, indicating strong nonlinearity in the selection process. Given the prediction for the selection probability, $\hat{p}_i$, from the first stage, we estimate a partial linear model where the bias correction term $\lambda(\cdot)$ is approximated by cubic b-spline basis functions with 7 knots.

The estimation results are shown in Table IV for the racial wage gaps. It is surprising that the OLS assuming random selection and the Heckman selection approach using the MLE produce the same estimate of the racial wage gap for both men (-11.3%) and women (-7.8%). The Heckman two-step estimator, on the other hand, gives quite different results from the OLS and Heckit MLE with inflated standard errors. As it does not exploit the full information in the model, the two-step estimator tend to be less reliable. The OLS and Heckman MLE generally produce almost identical coefficient estimates for all covariates. In Heckman's approach, the null hypothesis of no correlation between the error terms cannot be rejected. Both OLS and Heckman MLE estimates are contained in the HH bounds, meaning that the linear selection models fail to capture any selection bias. As we can see in the first stage estimation, linearity of the selection process is strongly rejected, so the linear

TABLE IV

WAGE REGRESSION AND RACIAL WAGE DISPARITY

	Men				Women			
	TPM	HS 2step	HS MLE	KL	TPM	HS 2step	HS MLE	KL
mexican	-0.113	-0.084	-0.113	-0.087	-0.078	-0.013	-0.078	-0.065
	(0.005)	(0.012)	(0.005)	(0.009)	(0.005)	(0.017)	(0.005)	(0.007)
age	0.078	0.112	0.079	0.108	0.111	0.213	0.113	0.133
	(0.006)	(0.014)	(0.006)	(0.010)	(0.007)	(0.026)	(0.007)	(0.009)
age^2	0.000	-0.001	0.000	-0.001	0.000	-0.001	0.000	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
exp	-0.025	-0.045	-0.025	-0.043	-0.069	-0.127	-0.070	-0.082
	(0.005)	(0.009)	(0.005)	(0.007)	(0.006)	(0.016)	(0.006)	(0.007)
exp^2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
less than hs	-0.169	-0.222	-0.170	-0.217	-0.174	-0.372	-0.177	-0.227
	(0.012)	(0.023)	(0.012)	(0.016)	(0.014)	(0.050)	(0.015)	(0.017)
some college	0.052	0.043	0.051	0.045	0.033	0.017	0.033	0.027
	(0.009)	(0.011)	(0.009)	(0.010)	(0.011)	(0.014)	(0.011)	(0.011)
college	0.235	0.205	0.235	0.210	0.156	0.084	0.155	0.134
	(0.022)	(0.026)	(0.023)	(0.023)	(0.025)	(0.036)	(0.025)	(0.026)
adv degrees	0.258	0.194	0.257	0.205	0.200	0.113	0.199	0.173
	(0.031)	(0.041)	(0.031)	(0.034)	(0.034)	(0.047)	(0.034)	(0.035)
veteran	-0.001	0.015	-0.001	0.013	0.030	0.037	0.030	0.032
	(0.006)	(0.008)	(0.006)	(0.007)	(0.016)	(0.020)	(0.016)	(0.016)
married	0.135	0.185	0.136	0.178	0.034	-0.079	0.033	0.010
	(0.004)	(0.019)	(0.005)	(0.012)	(0.004)	(0.028)	(0.005)	(0.007)
california	0.151	0.140	0.151	0.141	0.204	0.178	0.204	0.199
	(0.007)	(0.009)	(0.007)	(0.008)	(0.007)	(0.011)	(0.007)	(0.008)
arizona	0.042	0.052	0.042	0.050	0.098	0.103	0.098	0.099
	(0.009)	(0.010)	(0.009)	(0.009)	(0.009)	(0.012)	(0.009)	(0.009)
texas	0.015	0.045	0.015	0.041	0.030	0.064	0.031	0.038
	(0.007)	(0.014)	(0.008)	(0.010)	(0.008)	(0.013)	(0.008)	(0.008)

Note: The values in parentheses are standard errors.

TABLE V

WAGE REGRESSION AND GENDER WAGE DISPARITY

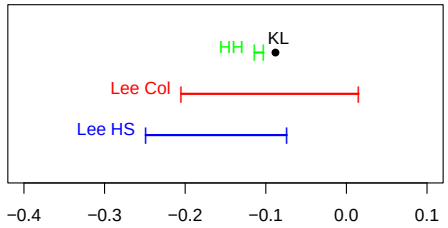
	Mexican				White			
	TPM	HS 2step	HS MLE	KL	TPM	HS 2step	HS MLE	KL
female	-0.193	-0.215	-0.195	-0.179	-0.209	-0.186	-0.159	-0.211
	(0.006)	(0.028)	(0.007)	(0.012)	(0.003)	(0.016)	(0.004)	(0.004)
age	0.056	0.075	0.058	0.044	0.103	0.077	0.045	0.107
	(0.009)	(0.025)	(0.010)	(0.013)	(0.006)	(0.019)	(0.006)	(0.006)
age^2	0.000	0.000	0.000	0.000	-0.001	0.000	0.000	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
exp	-0.019	-0.029	-0.020	-0.013	-0.053	-0.037	-0.019	-0.056
	(0.007)	(0.014)	(0.007)	(0.008)	(0.005)	(0.012)	(0.005)	(0.005)
exp^2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
less than hs	-0.186	-0.216	-0.188	-0.168	-0.159	-0.108	-0.042	-0.166
	(0.014)	(0.040)	(0.015)	(0.019)	(0.012)	(0.037)	(0.013)	(0.013)
some college	0.078	0.079	0.078	0.075	0.033	0.039	0.047	0.030
	(0.012)	(0.012)	(0.012)	(0.013)	(0.009)	(0.010)	(0.010)	(0.009)
college	0.323	0.315	0.323	0.316	0.171	0.192	0.217	0.165
	(0.029)	(0.031)	(0.029)	(0.030)	(0.021)	(0.026)	(0.023)	(0.021)
adv degrees	0.408	0.388	0.406	0.404	0.201	0.234	0.268	0.194
	(0.040)	(0.047)	(0.040)	(0.041)	(0.029)	(0.037)	(0.032)	(0.029)
vetetran	0.069	0.059	0.068	0.076	0.000	-0.010	-0.024	0.002
	(0.013)	(0.018)	(0.013)	(0.014)	(0.006)	(0.010)	(0.007)	(0.006)
married	0.095	0.100	0.095	0.091	0.081	0.089	0.097	0.081
	(0.006)	(0.009)	(0.006)	(0.007)	(0.003)	(0.007)	(0.004)	(0.003)
california	0.147	0.149	0.147	0.146	0.185	0.196	0.209	0.185
	(0.012)	(0.012)	(0.012)	(0.012)	(0.006)	(0.010)	(0.007)	(0.006)
arizona	0.052	0.060	0.053	0.047	0.074	0.072	0.066	0.074
	(0.015)	(0.018)	(0.015)	(0.016)	(0.007)	(0.008)	(0.008)	(0.007)
texas	-0.029	-0.016	-0.028	-0.039	0.035	0.021	0.001	0.037
	(0.012)	(0.021)	(0.012)	(0.014)	(0.006)	(0.012)	(0.007)	(0.006)

Note: The values in parentheses are standard errors.

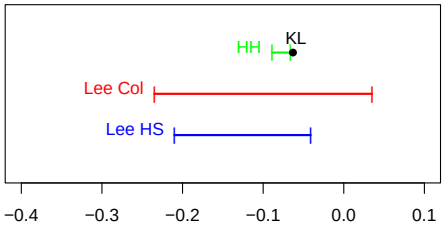
selection models are misspecified regardless of the assumption on the error terms. On the contrary, our semiparametric estimator shows a smaller magnitude of the racial wage disparity which is outside the HH bounds for both (-8.7%) men and women (-6.5%). Figure 6(a)-(b) compares our point estimates with the bounds estimates.

Our estimator also corrects selection bias in the coefficient estimates for the other covariates. The semiparametric estimator yields smaller wage premiums for higher education degrees (particularly for advanced degrees) for both men and women. Veteran status provides a higher wage premium for women than men, whose veteran premium is virtually negligible. Married men earn significantly higher wages than unmarried men, while married and unmarried women exhibit no significant difference in their hourly wage rates. The standard errors of our semiparametric estimates remain comparable to those obtained using Heckman MLE. These results effectively demonstrate the strong efficiency of our semiparametric estimator. There is a minimal difference in the estimated state fixed effects between the estimators. Both men and women are the highest-paid in California, followed by Arizona. The wage premium associated with residing in California and Arizona is higher for women compared to men by approximately 5% points relative to their counterparts in New Mexico.

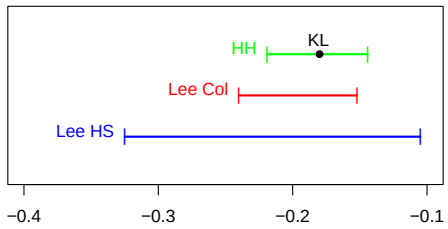
The results on the gender wage gap also show interesting patterns as shown in Table V. For Mexican Americans, the OLS and Heckit MLE again produce the same estimates (around -19.5%). In contrast, our estimator indicates a smaller magnitude of the gender wage disparity (-18%). As the HH bounds in this case are quite wide, all the point estimates are contained in the bounds. For non-Hispanic whites, the patterns are quite the opposite. Heckit MLE seems to over-correct the selection bias, delivering a much smaller magnitude of the gender wage gap (-15.9%) than OLS (-20.9%). It also indicates much larger premiums on higher degrees (college and advanced degrees) compared to high school diploma than the OLS. These patterns are totally flipped in the semiparametric estimation. Our estimator produces a very similar estimate of the gender wage gap (-21.1%) to the OLS, while it produces lower wage premiums of higher degrees. Interestingly, the Heckman MLE esti-



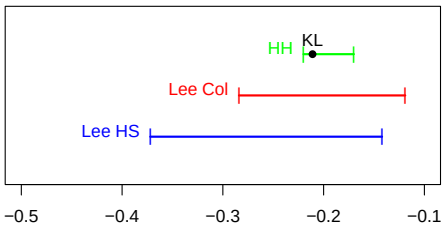
(a) Racial wage gap (men)



(b) Racial wage gap (women)



(c) Gender wage gap (Mexican-Americans)



(d) Gender wage gap (Non-Hispanic whites)

FIGURE 6.—Bounds and point estimates of wage gaps across gender and racial groups

mate does not lie in the HH bounds, whereas our semiparametric estimate is still contained in the bounds as shown in Figure 6(c)-(d).

Finally, we incorporate heteroskedasticity of the error term in our semiparametric model and compute the heteroskedasticity-robust standard errors of the coefficients. Table VI presents the results. The robust standard errors are generally almost identical to the standard errors computed under the homoskedasticity assumption. The robust standard errors tend to be slightly larger than the non-robust errors, but occasionally slightly smaller.

This empirical application demonstrates that the widely used bounds approach proposed by Lee (2009) can yield uninformative bounds in analyzing crucial labor market outcomes, such as wages. The HH bounds offer a potential alternative, as they tend to provide tighter bounds. However, even the feasible non-sharp version of the HH bounds (as the sharp char-

acterization relies on an uncountable infinity of moment inequalities) are computationally intensive. Moreover, the inference for these bounds hinges on resampling, which can be computationally demanding. In contrast, our semiparametric estimator is straightforward to implement in standard statistical packages like Stata and R, as it is a simple two-step plugin estimator. Any nonparametric estimator that satisfies the rate condition outlined in Section 3 can be used for the first stage of estimation, which calculates the selection probability. The second stage can then be executed using standard partial linear regression. The asymptotically valid standard errors are computationally straightforward and incorporating heteroskedasticity is also very tractable. The estimator is efficient, as demonstrated in this application and simulations. Therefore, our estimator presents a valuable alternative that

TABLE VI

SEMIPARAMETRIC WAGE REGRESSION WITH HETEROSKEDASTICITY ROBUST STANDARD ERRORS

	Racial Wage Gap				Gender Wage Gap			
	Men		Women		Mexican		White	
Variable	Coef	s.e.	Coef	s.e.	Coef	s.e.	Coef	s.e.
wage gap	-0.087	(0.009)	-0.065	(0.006)	-0.179	(0.013)	-0.211	(0.005)
age	0.108	(0.010)	0.134	(0.008)	0.044	(0.013)	0.107	(0.007)
age2	-0.001	(0.000)	-0.001	(0.000)	0.000	(0.000)	-0.001	(0.000)
exp	-0.043	(0.007)	-0.082	(0.006)	-0.013	(0.008)	-0.056	(0.005)
exp2	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)
less than hs	-0.217	(0.016)	-0.227	(0.017)	-0.168	(0.019)	-0.166	(0.013)
some college	0.045	(0.010)	0.027	(0.010)	0.075	(0.012)	0.030	(0.009)
college	0.210	(0.024)	0.133	(0.025)	0.316	(0.028)	0.165	(0.022)
adv degree	0.205	(0.034)	0.171	(0.034)	0.404	(0.039)	0.194	(0.030)
vet	0.013	(0.007)	0.032	(0.014)	0.076	(0.014)	0.002	(0.006)
marital	0.178	(0.012)	0.010	(0.007)	0.091	(0.007)	0.081	(0.003)
calif	0.141	(0.008)	0.199	(0.007)	0.146	(0.013)	0.185	(0.006)
arizo	0.050	(0.009)	0.099	(0.009)	0.047	(0.016)	0.074	(0.007)
texas	0.041	(0.010)	0.038	(0.008)	-0.039	(0.014)	0.037	(0.006)

Note: 'Wage gap' is the coefficient estimate on the 'Mexican American' dummy for the racial wage gap and the coefficient estimate on the 'female' dummy for the gender wage gap. 's.e.' is the heteroskedasticity robust standard error.

can be easily applied in cases where the bounds approaches fail to provide informative results, while it remains more robust than linear selection models. For researchers interested in correcting sample selection bias without resorting to unjustifiable parametric or distributional assumptions, we recommend reporting point estimates from our semiparametric selection model.

6. CONCLUDING REMARKS

In this paper, we investigate point identification and efficient estimation of semiparametric selection models without imposing an exclusion restriction. We do not restrict the selection process to be linear, demonstrating identification of the model parameters when there is at least one continuous covariate and the linearity of the selection process is violated. The primary objective of our paper is to challenge the long-held belief that an exclusion restriction is necessary for semiparametric selection models. Bounds approaches for selection models are often motivated by this misconception. We present convenient and practical semiparametric estimators that accommodate non-monotone selection, heteroskedastic error, multiple control variables, and simple asymptotically valid inference. Our recommendation for applied researchers is to report point estimates using our semiparametric method when their preferred bounds are not sufficiently informative. The identifying conditions are readily verifiable in practice, as researchers simply need to ensure the presence of a continuous variable in the data and reject the linearity of the selection process.

In our simulations and empirical applications, we demonstrate that our method provides more robust estimates of parameters of interest compared to linear selection models such as the Heckman selection model and Honoré and Hu (2020)'s model. While our semiparametric approach is not necessarily nested within Lee's fully nonparametric model, it imposes more restrictive assumptions than Lee's. Our model can permit parameter heterogeneity but it does not allow treatment effects to vary across different subpopulations. Extending our results to the case where the treated group and the untreated group have different treatment effects beyond the assumption made in Honoré and Hu (2024) would be an intriguing avenue for future research. Another promising research direction would be identification of semiparametric sample selection models with endogenous regressors.

REFERENCES

AHN, HYUNGTAIK AND JAMES L POWELL (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58 (1-2), 3–29. [3, 4]

ANDREWS, DONALD WK (1994): "Empirical process methods in econometrics," *Handbook of econometrics*, 4, 2247–2294. [13]

ANDREWS, DONALD WK AND MARCIA MA SCHAFGANS (1998): "Semiparametric estimation of the intercept of a sample selection model," *The Review of Economic Studies*, 65 (3), 497–517. [6]

CHAMBERLAIN, GARY (1986): "Asymptotic efficiency in semi-parametric models with censoring," *journal of Econometrics*, 32 (2), 189–218. [3]

CHEN, SONGNIAN (2010a): "An integrated maximum score estimator for a generalized censored quantile regression model," *Journal of Econometrics*, 155 (1), 90–98. [5]

——— (2010b): "Non-parametric identification and estimation of truncated regression models," *The Review of Economic Studies*, 77 (1), 127–153. [5]

CHEN, SONGNIAN AND XIANBO ZHOU (2011): "Semiparametric estimation of a bivariate Tobit model," *Journal of econometrics*, 165 (2), 266–274. [5]

——— (2012): "Semiparametric estimation of a truncated regression model," *Journal of Econometrics*, 167 (2), 297–304. [5]

CHEN, XIAOHONG (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," Elsevier, vol. 6 of *Handbook of Econometrics*, 5549–5632. [11, 12]

DAS, MITALI, WHITNEY K NEWEY, AND FRANCIS VELLA (2003): "Nonparametric estimation of sample selection models," *The Review of Economic Studies*, 70 (1), 33–58. [3]

DUAN, NAIHUA, WILLARD G MANNING, CARL N MORRIS, AND JOSEPH P NEWHOUSE (1984): "Choosing between the sample-selection model and the multi-part model," *Journal of Business & Economic Statistics*, 2 (3), 283–289. [3]

ESCANCIANO, JUAN CARLOS, DAVID JACHO-CHÁVEZ, AND ARTHUR LEWBEL (2016): "Identification and estimation of semiparametric two-step models," *Quantitative Economics*, 7 (2), 561–589. [5]

HAY, JOEL W AND RANDALL J OLSEN (1984): "Let them eat cake: a note on comparing alternative models of the demand for medical care," *Journal of Business & Economic Statistics*, 2 (3), 279–282. [3]

HECKMAN, JAMES (1974): "Shadow prices, market wages, and labor supply," *Econometrica: journal of the econometric society*, 679–694. [2]

——— (1990): "Varieties of selection bias," *The American Economic Review*, 80 (2), 313–318. [6]

HECKMAN, JAMES J (1979): "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, 153–161. [2]

HONORÉ, BO E AND LUOJIA HU (2020): "Selection without exclusion," *Econometrica*, 88 (3), 1007–1029. [3, 19, 22, 23, 30]

——— (2024): "Sample selection models without exclusion restrictions: Parameter heterogeneity and partial identification," *Journal of Econometrics*, 243 (1-2), 105360. [9, 30]

LEE, DAVID S. (2009): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 76 (3), 1071–1102. [3, 19, 24, 28]

MANNING, WILLARD G, NAIHUA DUAN, AND WILLIAM H ROGERS (1987): "Monte Carlo evidence on the choice between sample selection and two-part models," *Journal of econometrics*, 35 (1), 59–82. [3]

MORA, RICARDO (2008): "A nonparametric decomposition of the Mexican American average wage gap," *Journal of Applied Econometrics*, 23 (4), 463–485. [22]

NEWEY, WHITNEY K (2009): "Two-step series estimation of sample selection models," *The Econometrics Journal*, 12 (suppl_1), S217–S229. [3]

NEWEY, WHITNEY K AND JAMES L POWELL (1993): "Efficiency bounds for some semiparametric selection models," *Journal of Econometrics*, 58 (1-2), 169–184. [4]

PAN, ZHEWEN, XIANBO ZHOU, AND YAHONG ZHOU (2022): "Semiparametric Estimation of a Censored Regression Model Subject to Nonparametric Sample Selection," *Journal of Business & Economic Statistics*, 40 (1), 141–151. [5]

SEMENOVA, VIRA (2023): "Generalized lee bounds," *arXiv preprint arXiv:2008.12720*. [3]

SONG, KYUNGCHUL (2012): "On the smoothness of conditional expectation functionals," *Statistics & Probability Letters*, 82 (5), 1028–1034. [12, 34, 35]

——— (2014): "Semiparametric models with single-index nuisance parameters," *Journal of Econometrics*, 178, 471–483. [12, 13, 14]

APPENDIX A: LEMMAS FOR ASYMPTOTIC PROPERTIES OF THE ESTIMATORS

We provide lemmas that are used in proving the asymptotic results in Section 3. Let $X = [X_1', X_2']' \in \mathbb{R}^{d_1+d_2}$ be a random vector on a probability space, and let $\mathcal{P}$ be a collection of Borel measurable real maps on $\mathbb{R}^{d_1+d_2}$ such that $p(X)$ is a continuous random variable for each $p \in \mathcal{P}$. Suppose X_1 is a vector of continuous variables and X_2 is a vector of discrete varibles that take values from $\{x_1, \ldots, x_M\}$. Let S_m and $S_{1,m}$ be the supports of $X \cdot \mathbb{1}[X_2 = x_m]$ and $X_1 \cdot \mathbb{1}[X_2 = x_m]$.

We make the following assumptions under which Lemma 1 is derived.

ASSUMPTION 8: *(i)* $\mathbb{E}[Y|p = \cdot]$ *and* $\mathbb{E}[X|p = \cdot]$ *are twice continuously differentiable with derivatives bounded uniformly over* $p \in B(p_0; \varepsilon)$ *with some* $\varepsilon > 0$*; (ii) for some* $\varepsilon > 0$*,* $P[D = 1|p_0 = p] > \varepsilon$ *for all* $p \in [0, 1]$.

ASSUMPTION 9: *There exists $\varepsilon > 0$ such that for each $p \in B(p_0; \varepsilon)$, (i) p is continuous and its conditional density function given $D = 1$ is bounded uniformly over $p \in B(p_0; \varepsilon)$ and bounded away from zero on the interior of its support uniformly over $p \in B(p_0; \varepsilon)$; and (ii) the set $\{p(x); p \in B(p_0; \varepsilon), x \in S_m\}$ is an interval of finite length for all $1 \leq m \leq M$.*

The following lemma proves that the sum of the terms in the RHS of (14), in which $\tilde{p}$ is replaced with $\hat{p}$, is $o_p(1)$.

LEMMA 1: *Let Assumptions 8–9 hold. Then,*

$$\mathbb{E}\left[\left(\psi_{1i}(\hat{p}_n) - \psi_{1i}(p_0)\right)^2\right], \ \mathbb{E}\left[\left(\psi_{2i}(\hat{p}_n) - \psi_{2i}(p_0)\right)^2\right], \ \mathbb{E}\left[\left(\psi_{3i}(\hat{p}_n) - \psi_{3i}(p_0)\right)^2\right] \to 0,$$

as $\hat{p}_n \to_p p_0$, where

$$\psi_{1i}(p) := D_i \mathbb{E}[X_i | p_i, D_i = 1] Z_i, \ \psi_{2i}(p) := D_i \mathbb{E}[X_i | p_i, D_i = 1] \mathbb{E}[Z_i | p_i, D_i = 1]$$

$$\psi_{3i}(p) := D_i X_i \mathbb{E}[Z_i | p_i, D_i = 1] - \mathbb{E}[D_i X_i \mathbb{E}[Z_i | p_i, D_i = 1]]$$

PROOF: Since

$$\begin{aligned}
\psi_{1i}(\hat{p}_n) - \psi_{1i}(p_0) =& D_i \left(\mathbb{E}[X_i | \hat{p}_{ni}, D_i = 1] - \mathbb{E}[X_i | p_{0i}, D_i = 1]\right) Z_i, \\
\psi_{2i}(\hat{p}_n) - \psi_{2i}(p_0) =& D_i \left(\mathbb{E}[X_i | \hat{p}_{ni}, D_i = 1] - \mathbb{E}[X_i | p_{0i}, D_i = 1]\right) \mathbb{E}[Z_i | \hat{p}_{ni}, D_i = 1] \\
&+ D_i \mathbb{E}[X_i | p_{0i}, D_i = 1] \left(\mathbb{E}[Z_i | \hat{p}_{ni}, D_i = 1] - \mathbb{E}[Z_i | p_{0i}, D_i = 1]\right), \\
\psi_{3i}(\hat{p}_n) - \psi_{3i}(p_0) =& D_i X_i \left(\mathbb{E}[Z_i | \hat{p}_{ni}, D_i = 1] - \mathbb{E}[Z_i | p_{0i}, D_i = 1]\right) \\
&+ \mathbb{E}[D_i X_i \mathbb{E}[Z_i | \hat{p}_{ni}, D_i = 1]] - \mathbb{E}[D_i X_i \mathbb{E}[Z_i | p_{0i}, D_i = 1]],
\end{aligned}$$

it follows from Assumption 8 that $\psi_{ki}(\hat{p}_n) \to_p \psi_{ki}(p_0)$ for $k = 1, 2, 3$. Furthermore, ψ_{ki}^2 is uniformly integrable from Assumption 9, which completes the statement. *Q.E.D.*

Now we further make the following assumptions to derive (15), which is used to prove that B_n in (11) is $o_p\left(n^{-1/2}\right)$, in Lemma 2.

ASSUMPTION 10: *For $r \geq 4$, $\sup_{x \in \mathcal{X}} \mathbb{E}\left[|Y|^r | X = x\right] + \sup_{x \in \mathcal{X}} \|X\|^r < \infty$.*

ASSUMPTION 11: *For each $m = 1, \ldots, M$, (i) $\mathbb{E}[Y|X_1 = x, (X_2, D) = (x_m, 1)]$ and the inverse image of $p \in \mathcal{P}$ are Lipschitz continuous; and (ii) there are a finite number of partitions of S_m such that $p(X)$ is monotone in each partition.*

LEMMA 2: *Let Assumptions 9–11 hold. Then, there exist $C > 0$ and $\varepsilon > 0$ such that for each $\eta \in (0, \varepsilon]$,*

$$\sup_{p:\|p-p_0\|\leq\eta} \|a(p) - a(p_0)\| \leq C\eta^2.$$

Proof. By Theorem 1 of Song (2012), the statement in the lemma holds when Assumption 2.(ii) of Song (2012) is satisfied. Let A be a subset of $[0, 1]$. The inverse image $p^{-1}(A)$ is the set of all points in the support of X that map into A:

$$p^{-1}(A) = \{x \in \mathbb{R}^{d_1+d_2} : p(x) \in A\}.$$

Choose $a_1 \leq a_2 \leq a_3 \leq a_4$, and let $B_1 := [a_2, a_3]$ and $B_2 := [a_1, a_4]$, such that $p^{-1}(B_1) \cap S_m \neq \emptyset$. Let $A_{1,p,m} := p^{-1}(B_1) \cap S_m$ and $A_{2,p,m} := p^{-1}(B_2) \cap S_m$, so that, with $b_{L,p,m} := \inf\{p(x) : x \in S_m\}$ and $b_{U,p,m} := \sup\{p(x) : x \in S_m\}$,

$$A_{1,p,m} = \{x \in S_m : c_{2,p,m} \leq p(x) \leq c_{3,p,m}\}, \; A_{2,p,m} = \{x \in S_m : c_{1,p,m} \leq p(x) \leq c_{4,p,m}\},$$

where

$$c_{1,p,m} := \max\{a_1, b_{L,p,m}\}, \quad c_{2,p,m} := \max\{a_2, b_{L,p,m}\},$$
$$c_{3,p,m} := \min\{a_3, b_{U,p,m}\}, \quad c_{4,p,m} := \min\{a_4, b_{U,p,m}\}.$$

Note that the Hausdorff metric between $A_{1,p,m}$ and $A_{2,p,m}$ is defined by

$$d(A_{1,p,m}, A_{2,p,m}) := \max\left\{\sup_{a\in A_{1,p,m}} \inf_{b\in A_{2,p,m}} \|a - b\|, \sup_{b\in A_{2,p,m}} \inf_{a\in A_{1,p,m}} \|a - b\|\right\}.$$

Since $\sup_{a\in A_{1,p,m}} \inf_{b\in A_{2,p,m}} \|a-b\| = 0$,

$$\begin{aligned} d\left(A_{1,p,m}, A_{2,p,m}\right) &= \sup_{b\in A_{2,p,m}} \inf_{a\in A_{1,p,m}} \|a-b\| \\ &\leq \sup_{b\in A_{21,p,m}} \inf_{a\in A_{1,p,m}} \|a-b\| + \sup_{b\in A_{22,p,m}} \inf_{a\in A_{1,p,m}} \|a-b\| \\ &\leq \sup_{b\in A_{21,p,m}} \inf_{a\in \bar{A}_{11,p,m}} \|a-b\| + \sup_{b\in A_{22,p,m}} \inf_{a\in \bar{A}_{12,p,m}} \|a-b\|, \end{aligned}$$

where

$$A_{21,p,m} = \left\{x \in S_m : c_{1,p,m} \leq p\left(x\right) \leq c_{2,p,m}\right\},\ \bar{A}_{11,p,m} = \left\{x \in S_m : p\left(x\right) = c_{2,p,m}\right\},$$

$$A_{22,p,m} = \left\{x \in S_m : c_{3,p,m} \leq p\left(x\right) \leq c_{3,p,m}\right\},\ \bar{A}_{12,p,m} = \left\{x \in S_m : p\left(x\right) = c_{3,p,m}\right\}.$$

Let $S_m^1, \ldots, S_m^{n_m}$ be partitions of S_m such that $p\left(X\right)$ is monotone in each partition. Then

$$\sup_{b\in A_{21,p,m}} \inf_{a\in \bar{A}_{11,p,m}} \|a-b\| \leq \sum_{k=1}^{n_m} \sup_{b\in A_{21,p,m}\cap S_m^k} \inf_{a\in \bar{A}_{11,p,m}} \|a-b\|,$$

$$\sup_{b\in A_{22,p,m}} \inf_{a\in \bar{A}_{12,p,m}} \|a-b\| \leq \sum_{k=1}^{n_m} \sup_{b\in A_{22,p,m}\cap S_m^k} \inf_{a\in \bar{A}_{12,p,m}} \|a-b\|$$

For each $k = 1, \ldots, n_m$, there exists c_k such that

$$\sup_{b\in A_{21,p,m}\cap S_m^k} \inf_{a\in \bar{A}_{11,p,m}} \|a-b\| \leq c_k\left(a_2 - a_1\right),$$

$$\sup_{b\in A_{22,p,m}\cap S_m^k} \inf_{a\in \bar{A}_{12,p,m}} \|a-b\| \leq c_k\left(a_4 - a_3\right),$$

which confirms that Assumption 2.(ii) of Song (2012) is satisfied.