A Semiparametric Machine Learning Estimator for Sample Selection Models

Sebastian T. Roelsgaard[1,3] and Luke Taylor[2]

Abstract

We propose a semiparametric machine learning estimator for sample selection models that is asymptotically normal and $\sqrt{n}$-consistent, and show its improvement over existing methods in an empirical Monte Carlo. The simulated data relates to prison sentencing and is closely aligned to real court case data from North Carolina. The improvement is particularly pronounced in the absence of an exclusion restriction and when the proportion of censored observations increases; in some cases, mean squared error is reduced by more than 99%. Finite-sample evidence of the accuracy of estimated standard errors is also provided.

JEL Classification Codes: C14; C34

Keywords: Sample Selection; Machine Learning; Semiparametric Estimation; Empirical Monte Carlo

We are grateful to Bo Honoré, José Luis Montiel Olea for insightful and inspiring comments. We thank Michal Kolesár, Carol Shou, and Daniel Wilhelm for helpful feedback. We are thankful to Malene Vindfeldt Skals for proofreading the paper. All remaining errors are our own.

[1]Department of Economics, Princeton University, Julis Romo Rabinowitz Building, Princeton, NJ 08540, United States. Email: roelsgaard@princeton.edu.
[2]Department of Economics, Aarhus University, Fuglesangs Alle 4, Aarhus V 8210, Denmark. Email: lntaylor@econ.au.dk.
[3]Dale T. Mortensen Center, Department of Economics, Aarhus University.

1 Introduction

This paper proposes a semiparametric machine learning (SPML) estimator for sample selection models and provides instructions on its implementation. The estimator exploits any potential nonlinearity in the selection function to rectify the bias from sample selection – even in the absence of an exclusion restriction – and is shown to be asymptotically normal and $\sqrt{n}$-consistent. We showcase the estimator's superiority, in terms of mean squared error, compared to conventional methods in an empirical Monte Carlo, where the data is constructed to closely mimic a real dataset from the North Carolina court system. Finally, we show that the estimator continues to perform admirably even when the data generating process is modified to ideally suit the well-known [Heckman, 1979] estimator.

To fix ideas, consider the sample selection model

$$Y^* = X'\beta + U \tag{1}$$

$$S = 1\{m(W) \geq V\}, \tag{2}$$

where $Y^* \in \mathbb{R}$, $U \in \mathbb{R}$, $V \in \mathbb{R}$, $X \in \mathbb{R}^p$, and $W \in \mathbb{R}^q$, where $q \geq p$.

We assume $\mathbb{E}[U|W] = 0$, $Y = Y^*$ is observed only if $S = 1$, and X is a subvector[1] of W. If W contains elements that are not in X, we say that these elements satisfy an exclusion restriction, as they affect selection but not the outcome. We assume throughout that (U, V) are continuously distributed with connected support and that $W \perp (U, V)$. We use capital letters to denote random vectors and variables, and lowercase letters, with a subscript, to indicate realizations of these.

This model implies

$$E[Y|W, S = 1] = X'\beta + f(m(W)), \tag{3}$$

where $f(m(W)) \equiv E[U|m(W), S = 1] = E[U|W, S = 1]$, which, in general, is non-zero if U

[1]By a subvector we mean that all elements in X are elements in W as well, but W may contain elements not in X. Formally, if $W = (W_1, W_2, \ldots, W_q)$ there exists indices $\{j_1, j_2, ..., j_p\} \subseteq \{1, 2, \ldots, q\}$ such that $x_i = (W_{j_1}, W_{j_2}, \ldots, W_{j_p})$.

and V are not independently distributed, causing bias in the ordinary least squares (OLS) estimate of β.

In general, methods to nullify the selection bias involve directly controlling for $f(m(W))$. However, this term may exhibit perfect multicollinearity with X, resulting in failure to identify β. The identifying assumption in this paper[2] is therefore that $f(\cdot)$ and $m(W)$ are such that there does not exist a parameter vector γ such that $f(m(W)) = X'\gamma$.

The shape of the function $f(\cdot)$ depends on the joint distribution of (U, V). Therefore, if, say, $m(W) = X'\delta$ for some δ, one has to make an assumption about the joint distribution of the errors to ensure identification of β. For example, [Heckman, 1979] assumed joint normality, which implies that $f(\cdot)$ is the inverse Mills ratio. This is a non-linear function ensuring identification of β. However, the inverse Mills ratio is nearly linear on a large part of its domain so an exclusion restriction is recommended to reduce the variance of the estimator of β.

Absent restrictions on the distribution of the errors (restrictions on the shape of $f(\cdot)$), identification instead requires that $m(W)$ is restricted to satisfy the identifying assumption above. If there is a variable satisfying an exclusion restriction, this will be satisfied: $m(W)$ depends on elements in W not included in X, so no function $f(\cdot)$ exists such that $f(m(W)) = X'\gamma$ for any γ. This is because individuals with similar X's have different probability of selection, breaking multicollinearity and enabling identification.

Even without an exclusion restriction, so $W = X$, $m(\cdot)$ could be sufficiently nonlinear to prevent multicollinearity. To see this, consider the case where X is a scalar random variable and $m(X) = X - X^2$. Then there exists no function $f(\cdot)$ such that $f(m(X)) = \gamma X$. Formally, $m(\cdot)$ is not allowed to be an invertible single-index function of X.

Our contribution is an estimator that exploits both nonlinearities and potential exclusion restrictions without making assumptions about the distribution of the errors. The estimator extends the semiparametric approach of [Newey, 2009]. In the first step, machine learning

[2]Note that some other identification arguments exist, such as 'identification at infinity,' but these are seldom used in applied work.

(ML) techniques are used to estimate $m(W)$. Typically, ML techniques are used to improve prediction or to deal with high-dimensional estimation problems; we instead use ML for its ability to elicit nonlinearities in the data generating process. In the second step, B-spline functions are evaluated at $\hat{m}(w_i)$ and added as control variables in the outcome equation to approximate $f(m(W))$. Finally, OLS is used to estimate β.

The intuition for the ability of ML techniques to uncover nonlinearities is perhaps best illustrated by decision trees, which are used in random forests. Decision trees partition the support of w_i into, say, J regions labeled $R_1, R_2, \ldots, R_J$, and assign a constant c_j to each,[3] giving an estimator of the form $\hat{m}(w_i) = \sum_{j=1}^{J} c_j 1\{w_i \in R_j\}$. This estimator is a 'simple function' in the measure-theoretic sense. Any measurable function, no matter how nonlinear or discontinuous, taking values on the real line, can be approximated arbitrarily well using such simple functions as $J \to \infty$. In contrast, conventional nonparametric estimation methods impose smoothness on the approximating function, which may be particularly restrictive with a multidimensional covariate.

In finite samples, an empirical Monte Carlo study[4] shows the substantial gains of the SPML estimator compared to existing methods. The study is also used to provide evidence of the suitability of estimated standard errors. Finally, we modify the Monte Carlo to ideally suit the Heckman two-step estimator and show that the SPML estimator continues to perform admirably. In this setting, $f(m(W))$ is the inverse Mills ratio and the SPML estimator correctly approximates this term without imposing a normality assumption, demonstrating the robustness and flexibility of the estimator. We do not attempt to answer any substantial empirical questions using the simulated data, but only to evaluate the performance of the SPML estimator.

The rest of this paper proceeds as follows. In Section 2, we present the SPML estimator, its implementation, and discuss inference procedures. Section 3 outlines the data on which

[3]The choice of c_j depends on the ML method used. c_j may be an average of the outcomes associated with the w_i within region j, or something more sophisticated.

[4]By 'empirical' we mean merely that the study is based on a real dataset.

the empirical Monte Carlo is based, how the simulated data is constructed, and presents and discusses the results. Section 4 concludes.

2 Estimation and Implementation

We restate the identifying assumption pertaining to equation 3 above:

Assumption. *There exists no parameter vector* $\gamma \in \mathbb{R}^p$ *such that* $f(m(W)) = X'\gamma$

2.1 Estimator

In the first step, we compute an ML estimate of $m(W)$. By predicting s_i using w_i, ML methods typically produce an estimate for $E[S|W]$ rather than $m(W)$ directly. However, a simple relationship exists between $E[S|W]$ and $m(W)$, namely $m(W) = F_V^{-1}(E[S|W])$, where F_V^{-1} is the inverse of the CDF of V. Since $m(W)$ only influences the outcome equation as an input to $E[U|m(W), S = 1]$, it is equivalent to use $E[S|W]$ directly in place of $m(W)$ if F_V is strictly increasing. This is the case if V is continuously distributed with connected support, as we assume. In practice, when estimating nuisance parameters in sample selection models via nonparametric methods, the assumed distribution of V has little impact on the final estimate (see, for example, [Newey et al., 1990] and [Vella, 1998]).

The SPML estimator for the sample selection model is constructed as follows:

1. Randomly partition the data into two disjoint sets, call these A and B. On set A, estimate $\hat{m}(W) = \hat{E}[S|W]$ by predicting s_i given w_i using ML techniques, denote the resulting function $\hat{m}^A(\cdot)$. Calculate $\hat{m}^A(w_i)$ for $i \in B$, and save this as an extra variable. Likewise, estimate $\hat{m}^B(\cdot)$ and calculate $\hat{m}^B(w_i)$ for $i \in A$, and save this as well. We refer to this new variable as $\hat{m}(w_i)$.
2. Construct B-spline control variables for $f(m(W))$ by evaluating L B-splines on $\hat{m}(w_i)$ from step 1:
   $\varphi_1(\hat{m}(w_i)), \varphi_2(\hat{m}(w_i)), \ldots, \varphi_L(\hat{m}(w_i))$.

3. Run OLS of y_i on $x_i, \varphi_1(\hat{m}(w_i)), \varphi_2(\hat{m}(w_i)), \ldots, \varphi_L(\hat{m}(w_i))$ to obtain $\hat{\beta}$.

This procedure is equivalent to that of [Newey, 2009] but uses ML methods in the first step, rather than semiparametric techniques, and data partitioning in the form of cross-fitting.

If the nuisance function estimated in the first step is estimated via either parametric methods or traditional nonparametric approaches, $\hat{\beta}$ would, in general, be $\sqrt{n}$-consistent without the need to partition the data. If, however, these functions are estimated via ML methods without partitioning the data, $\hat{\beta}$ is not $\sqrt{n}$-consistent; see, for example, [Chernozhukov et al., 2018] and [Chang, 2020] for a detailed discussion of this phenomenon.

To obtain the asymptotic properties of the SPML estimator, we can use the results of [Robinson, 1988] if it can be shown that the error resulting from using $\hat{m}$ in place of m is asymptotically negligible. To this end, we appeal to Theorems 1 and 3 of [Foster and Syrgkanis, 2019]. These theorems states that under general conditions for the target object, the algorithms used to estimate the target object, and the nuisance function, the estimation error from the nuisance function has a second order impact on the target estimation error if data partitioning is used. The key to their result is a functional analogue of Neyman orthogonality which holds under mild regularity conditions in the setting of this paper. Furthermore, the generality of the results in [Foster and Syrgkanis, 2019] allows for the majority of popular ML algorithms, or ensembles thereof, to be used in the first step; we opt for gradient boosted regression trees and random forests in the empirical Monte Carlo. The appendix contains details of the requirements needed for consistency and asymptotic normality of the SPML estimator.

2.2 Inference

As shown in [Ackerberg et al., 2012], for inference purposes, many semiparametric estimators - including the SPML estimator of this paper - can be treated as if they were fully parametric, allowing standard parametric two-step estimator inference procedures to be used. Furthermore, Theorems 1 and 3 of [Foster and Syrgkanis, 2019] can be used to show that

the estimation error resulting from the first-stage ML estimation of $m(W)$ is asymptotically negligible due to sample-splitting. Thus, the asymptotic variance of the SPML estimator is unaffected by replacing the nuisance function by its estimated counterpart, and the textbook OLS variance estimator can be used (or a heteroskedasticity robust or cluster robust version).[5]

Finally, although the asymptotic properties of the SPML estimator do not depend on the random partitions used, this does generate additional variance in finite samples. To account for this variability, we follow [Chernozhukov et al., 2018] and repeat the SPML procedure R times to give $\hat{\beta}_{(r)}$ for $r = 1, \ldots, R$. The adjusted variance estimator is then given by

$$\hat{\sigma}^2 = \frac{1}{R}\sum_{r=1}^{R} \hat{\sigma}^2_{(r)} + \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{(r)} - \frac{1}{s}\sum_{s=1}^{R}\hat{\beta}_{(s)}\right)^2.$$

3 Empirical Monte Carlo

We consider a sample selection model that explains the determinants of criminal sentencing. In particular, the selection equation characterizes the court's decision to convict or acquit a defendant at trial, and the outcome equation depicts the sentencing decision for those defendants who are convicted. A plausible exclusion restriction exists in this setting, using the random assignment of judges to cases (discussed below), which allows a comparison of estimators both with and without an exclusion restriction.

Subsection 3.1 discusses the data and provides summary statistics. Subsections 3.2 and 3.2.1 detail how the simulated data for the empirical Monte Carlo is constructed, and the subsequent modifications for the parametric setting designed for the Heckman two-step estimator. Results are presented and discussed in subsections 3.3 and 3.3.1.

[5]The use of sample-splitting to obtain such a result has long been known. Indeed, the split-sample IV and jackknife IV of [Angrist and Krueger, 1995] and [Angrist et al., 1999], respectively, use this idea, as does [Dufour and Jasiak, 2001] to improve inference in the presence of generated regressors.

3.1 Data

The dataset which forms the basis of the Monte Carlo is obtained from the North Carolina Administrative Office of the Courts. Trial courts in North Carolina can be broadly split into Superior Courts and District Courts. Only data from the District Courts is used in this paper.[6] These courts hear the majority of trial-level cases including almost all misdemeanors and infractions. There are currently 41 District Courts in North Carolina with 273 court judges who serve four-year terms. These judges decide the outcome of the trial and the resulting sentence, and are randomly assigned to cases by the chief District Court judge for their respective district.

The data covers the period 1995-2010, and the unit of observation is a single case. We make several restrictions to the sample to ensure a coherent analysis, leaving us with only single-charge misdemeanor cases settled by judges who hear at least 50 cases, and where the defendant is either black or white. Online Appendix B contains details of the restrictions made. The final dataset contains 14 705 observations. Summary statistics are given in Table 1.

In the notation of the sample selection model given in (1) and (2), y_i is the sentence given to a convicted defendant measured in number of days in prison (non-prison sentences are set at zero), s_i indicates conviction, and x_i includes the remaining variables in the table and a measure of the sentencing severity of the judge randomly assigned to the case. w_i includes all variables within x_i but excludes the judge sentencing severity and includes a measure of the conviction tendency of the judge. These two measures of judge leniency are explained fully in the next section.

For completeness, in Table A.1 in Online Appendix A, we report the outcome equation coefficients and standard errors related to this model using this original sample. However, the focus of this paper is on introducing the SPML estimator, therefore we do not discuss

[6] [Dyke, 2007] and [Silveira, 2017] use the same data source, but consider only Superior Courts, to analyze prosecutors' political incentives and plea bargaining, respectively.

Table 1: Descriptive Statistics

Status	Acquitted		Convicted	
Observations	2 432		12 273	
	Mean	Std. Dev.	Mean	Std. Dev.
Defendant Characteristics				
Age	30.8	10.2	31.4	10.8
Age Squared	1056.9	723.2	1099.7	790.2
Black	0.44	0.50	0.47	0.50
Female	0.15	0.35	0.17	0.38
Previous Criminality	0.00	0.00	2.23	2.47
Private Attorney	0.27	0.44	0.32	0.47
Court-Appointed Attorney	0.67	0.47	0.62	0.49
Public Defender	0.06	0.24	0.06	0.23
Case Characteristics				
Crime Severity [1]	0.17	0.37	0.07	0.25
Crime Severity [2]	0.28	0.45	0.12	0.33
Crime Severity [3]	0.46	0.50	0.52	0.50
Crime Severity [4]	0.09	0.29	0.29	0.45
Drug Crime	0.22	0.41	0.18	0.39
Property Crime	0.22	0.42	0.24	0.43
Peace Crime	0.40	0.49	0.22	0.41
Violent Crime	0.16	0.36	0.36	0.48
Sentence (Days)	-	-	55.8	51.1

these empirical results further.

3.1.1 Judge Leniency Measures

We wish to compare the performance of several estimators in settings where a plausible exclusion restriction exists and also in settings where it does not. As such, an exclusion restriction is required. We argue that the conviction tendency of the randomly assigned judge should be included in the selection equation - it affects the likelihood that a defendant is convicted - but, conditional on the sentencing severity of the judge, should not be included in the outcome equation. As previously mentioned, District Court judges in North Carolina are randomly assigned to cases by the chief District Court judge. Since the judge hearing

the trial decides both the outcome and the severity of the punishment, there is a strong correlation between their conviction tendency and their sentencing severity (0.28 with a p-value below 1%). Thus, it is only reasonable to exclude conviction tendency from the outcome equation if sentencing severity is included as a control and vice versa.

We follow the previous literature (for example, [Dobbie et al., 2018]) and construct a residualized leave-one-out measure of conviction tendency as follows. First, estimate the linear regression

$$C_{ijt} = \gamma\alpha_{jt} + \varepsilon_{ijt}, \tag{4}$$

where C_{ijt} denotes whether case i heard at courthouse j in year t resulted in a conviction, and α_{jt} are year $\times$ courthouse fixed-effects. The residualized conviction tendency is then given by

$$\eta_{ikt} = \frac{1}{n_k - 1} \sum_{l \neq i} \hat{\varepsilon}_{ljt},$$

where $\hat{\varepsilon}_{ljt}$ denotes the residual conviction decision from (4), and the sum is taken over all cases heard by judge k (excluding case i), with n_k being the total number of cases heard by judge k. A measure of sentencing severity is calculated analogously by replacing the left-hand side in 4 with the sentence for case i heard at courthouse j in year t.

There are 120 unique judges in the final dataset, with the average judge hearing 123 cases. A judge at the 75% quantile of conviction tendency convicts defendants at a rate 2.9 percentage points higher than a judge at the 25% quantile; a judge at the 75% quantile of sentencing severity sentences defendants to 18.5 days longer in prison than a judge at the 25% quantile. In Online Appendix A, Table A.2 reports results for a regression of the conviction tendency of the randomized judge on case and defendant characteristics, as well as results for an analogous regression using sentencing severity. Together, these provide evidence to the random assignment of judges to cases and, therefore, the suitability of the

exclusion restriction.

3.2 Method

The data is randomly partitioned in two, and we use the subscripts (1) and (2) to denote variables and estimates created from these respective datasets. Using dataset (1), parameter estimates and functional relationships are obtained, which, when applied to dataset (2), generate a set of outcomes mimicking that obtained in dataset (1). We employ ML methods to generate the functional relationships for the Monte Carlo. However, to avoid favoring the approaches of this paper, we use a random forest algorithm in the estimation of the parameters and functions from the first dataset (which then act as the population parameters for the Monte Carlo study), but use gradient boosted regression trees when applying the SPML estimator to the synthetic data generated using the aforementioned population parameters. In all cases, the ML tuning parameters are selected using 5-fold cross-validation.

Denote by θ_l the regression coefficient associated with $\varphi_l\left(\hat{m}\left(w_i\right)\right)$ and define $\theta = (\theta_1, \ldots, \theta_L)'$ and $\varphi\left(\hat{m}\left(w_i\right)\right) = (\varphi_1(\hat{m}(w_i)), \ldots, \varphi_L(\hat{m}(w_i)))'$, with $\hat{\theta}$ the corresponding estimate of θ.

With dataset (1), calculate $\hat{\beta}_{(1)}$ using the SPML estimator with the random forest ML algorithm; this also gives $\hat{m}_{(1)}(\cdot)$ and $\hat{\theta}_{(1)}$. Construct fitted values for dataset (2) as $\dot{y}_{i,(2)} = x'_{i,(2)}\hat{\beta}_{(1)} + \varphi\left(\hat{m}_{(1)}\left(w_{i,(2)}\right)\right)'\hat{\theta}_{(1)}$. This represents the expected sentence length for every observation in dataset (2) conditional on that individual being found guilty. No values are available for the counterfactual of the individual being acquitted, but these are never observed in real data in any case.

The outcome residuals can be calculated for the selected observations in dataset (2) as $\hat{u}_{i,s=1,(2)} = y_{i,(2)} - \dot{y}_{i,(2)}$. The outcome regression error term can then be constructed by taking a smoothed bootstrap resample of $\hat{u}_{i,s=1,(2)}$ of the same size as the sample size of dataset (2), denoted $u^*_{i,(2)}$. Note that because $\hat{u}_{i,s=1,(2)}$ has been purged of the selection bias, $\hat{u}_{i,s=1,(2)}$ has the same distribution as $\hat{u}_{i,s=0,(2)}$. Thus, it is sufficient to resample only from $\hat{u}_{i,s=1,(2)}$ to give error terms for both selected an un-selected observations. The final outcome variable is

then given by $y^*_{i,(2)} = \dot{y}_{i,(2)} + u^*_{i,(2)}$.

To determine which outcomes are censored, a distribution for the selection equation error, v_i, must be chosen. As discussed at the beginning of Section 2, if the distribution of the errors is continuous with connected support, using $\hat{m}(w_i) = \hat{E}[s_i|w_i]$ is sufficient for identification. We consider four different zero-mean distributions, namely: (1) the normal distribution, (2) the t-distribution with four degrees-of-freedom, (3) the skew-normal distribution with shape parameter five (the third central moment being equal to 2.5), and (4) a mixture of two normal distributions with means of -1.5 and 1.5 (giving a bimodal distribution). Each distribution is standardized to have the same variance. We denote the CDF of these simulated draws by F_{V^*}.

The choice of this variance for the error term should be informed by the data. To that end, we employ the 'semi-nonparametric' method of [Gallant and Nychka, 1987] to the binary choice regression of $s_{i,(1)}$ on $\Phi^{-1}\left(\hat{m}_{(1)}(w_{i,(1)})\right)$, where Φ^{-1} is the quantile function of the standard normal distribution.[7] By using $\Phi^{-1}\left(\hat{m}_{(1)}(w_{i,(1)})\right)$, we make the initial assumption that $v_i \sim N(0,1)$. The approach of [Gallant and Nychka, 1987] produces an updated estimate of the true density for V, f_V, that can then be used to estimate the variance of V.

Next, we obtain $v^*_{i,(2)}$ by sampling from F_{V^*}, and construct

$$s^*_{i,(2)} = 1\left\{a - F^{-1}_{V^*}\left(1 - \hat{m}_{(1)}(w_{i,(2)})\right) + v^*_{i,(2)} \geq 0\right\},$$

where a is a constant chosen such that the probability of selection is equal to the desired level; we consider 84% (equal to the proportion censored in the original data), 60%, and 40%. The outcome $y^*_{i,(2)}$ is then censored if $s^*_{i,(2)} = 0$.

From $\left(y^*_{i,(2)}, s^*_{i,(2)}, w_{i,(2)}\right)$, β is estimated using five different methods. The first uses the conventional Heckman two-step estimator. The second uses the semiparametric binary choice estimator of [Klein and Spady, 1993] to estimate the probability of selection and includes

[7] The method of [Gallant and Nychka, 1987] proceeds by replacing the unknown distribution of the error term in the binary choice likelihood function with a series approximation.

the inverse Mills ratio of the probability of selection in the outcome equation. The method of [Klein and Spady, 1993] avoids assuming that the selection equation error is normally distributed. However, since the inverse Mills ratio is used in the second stage, the normality assumption in the outcome equation error is still imposed. For the third approach, we use a probit first stage but follow [Newey, 2009] in the second stage by approximating the selection bias using a cubic B-spline of the inverse Mills ratio of the estimated selection probability. Using splines in the second stage provides robustness against the normality assumption. The fourth estimator uses the [Klein and Spady, 1993] estimator in the first stage, and the spline method of [Newey, 2009] in the second. Finally, results for the SPML method are given.

Having computed the first set of estimates, we draw a new sample of $u^*_{i,(2)}$ and $v^*_{i,(2)}$, creating a new $y^*_{i,(2)}$ and $s^*_{i,(2)}$, and the five estimators are computed on this new dataset. This process is repeated R times.

Three measures of estimator performance are then calculated. For $j = 1, \ldots, 5$ and $p = 1, \ldots, P$, let $\hat{\beta}^{(j)}_{r,p}$ denote the p^{th} slope coefficient estimated using the j^{th} estimator in the r^{th} simulated dataset, where P is the number of regressors in the outcome equation (excluding the constant and any coefficients related to selection bias terms). Also, let $\hat{\beta}_{(1),p}$ denote the p^{th} 'population' slope coefficient for $p = 1, \ldots, P$. To evaluate the performance of each estimator, the mean squared error (MSE) over the Monte Carlo simulations is calculated as

$$MSE^{(j)} = \frac{1}{P}\sum_{p=1}^{P}\left\{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}^{(j)}_{r,p} - \hat{\beta}_{(1),p}\right)^2\right\}.$$

In Tables A.3 and A.4 in Online Appendix A, we also report measures of the absolute bias and standard deviation of each estimate, given by

$$\begin{aligned} Bias^{(j)} =& \frac{1}{P}\sum_{p=1}^{P}\left\{\left|\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}^{(j)}_{r,p} - \hat{\beta}_{(1),p}\right|\right\}, \\ SD^{(j)} =& \frac{1}{P}\sum_{p=1}^{P}\left\{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}^{(j)}_{r,p} - \bar{\hat{\beta}}^{(j)}_{p}\right)^2\right\}^{1/2}, \end{aligned}$$

where $\bar{\hat{\beta}}_p^{(j)} = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{r,p}^{(j)}$.

Finally, to determine the sensitivity of the results to the presence of an exclusion restriction, we consider both the setting where the conviction tendency of the judge to convict is available, and the setting where it is not. However, note that to create $\hat{\beta}_{(1)}$, the conviction tendency is always used, i.e. the population objects do not change in these two settings.

To summarize, the empirical Monte Carlo is carried out as follows:

1. Randomly partition the data into dataset (1) and (2). With dataset (1), calculate $\hat{\beta}_{(1)}$ using the SPML estimator with random forest as the ML method; this also gives $\hat{m}_{(1)}(\cdot)$ and $\hat{\theta}_{(1)}$.
2. Construct $\dot{y}_{i,(2)} = x'_{i,(2)}\hat{\beta}_{(1)} + \varphi\left(\hat{m}_{(1)}\left(w_{i,(2)}\right)\right)'\hat{\theta}_{(1)}$.
3. Calculate the outcome equation residuals as $\hat{u}_{i,s=1,(2)} = y_{i,(1)} - \dot{y}_{i,(2)}$.
4. Take a smoothed bootstrap sample of $\hat{u}_{i,s=1,(2)}$ of size equal to that of the sample size of dataset (2), denote this $u^*_{i,(2)}$. Compute $y^*_{i,(2)} = \dot{y}_{i,(2)} + u^*_{i,(2)}$.
5. Obtain $v^*_{i,(2)}$ by sampling from the chosen CDF of v_i and construct $s^*_{i,(2)} = 1\left(a - F_V^{-1}\left(1 - \hat{m}_{(1)}(w_{i,(2)})\right) + v^*_{i,(2)} \geq 0\right)$. Mask all values of $y^*_{i,(2)}$ for which $s^*_{i,(2)} = 0$.
6. From $\left(y^*_{i,(2)}, s^*_{i,(2)}, w_{i,(2)}\right)$, estimate β using the five different methods.
7. Repeat steps (4) - (6) R times and calculate performance measures over these R samples.

3.2.1 Parametric Setting

In addition to the above described empirical Monte Carlo, we conduct a second study designed to be ideally suited to the Heckman two-step parametric estimator. Our aim is to show that even in such a setting, the SPML estimator performs better than existing methods. This second empirical Monte Carlo is designed as follows:

1. Randomly partition the data into dataset (1) and (2). With dataset (1), calculate $\hat{\beta}_{(1)}$ using the Heckman two-step estimator; this also gives the coefficient on the inverse

Mills ratio term, $\hat{\theta}_{(1)}$, and the coefficients from the probit estimation of the selection equation, denoted $\hat{\gamma}_{(1)}$.

2. Construct $\dot{y}_{i,(2)} = x'_{i,(2)}\hat{\beta}_{(1)} + \hat{\theta}_{(1)}\lambda\left(w'_{i,(2)}\hat{\gamma}_{(1)}\right)$, where λ denotes the inverse Mills ratio.
3. Calculate the outcome equation residuals as $\hat{u}_{i,s=1,(2)} = y_{i,(1)} - \dot{y}_{i,(2)}$.
4. Take a smoothed bootstrap sample of $\hat{u}_{i,s=1,(2)}$ of size equal to that of the sample size of dataset (2), denote this $u^*_{i,(2)}$. Compute $y^*_{i,(2)} = \dot{y}_{i,(2)} + u^*_{i,(2)}$.
5. Generate $v^*_{i,(2)} \sim N(0,1)$ and construct $s^*_{i,(2)} = 1\left\{a + w'_{i,(2)}\hat{\gamma}_{(1)} + v^*_{i,(2)} \geq 0\right\}$. Mask all values of $y^*_{i,(2)}$ for which $s^*_{i,(2)} = 0$.
6. From $\left(y^*_{i,(2)}, s^*_{i,(2)}, w_{i,(2)}\right)$, estimate β using the five different methods.
7. Repeat steps (4) - (6) R times and calculate performance measures over these R samples.

3.3 Results

To ease comparison, the MSE results in Table 2 and 3 are multiplied by 100, as are the bias and standard deviation results in Tables A.3, A.4, A.5, and A.6 in Online Appendix A. All results are based on 500 Monte Carlo simulations and a sample size of 14 705. For each parameter setting in Tables 2 and 3, the lowest MSE is highlighted in bold.

As would be anticipated, all five estimators exhibit better performance as a larger fraction of the sample is selected and when an exclusion restriction is present. The SPML method outperforms the other four estimators, regardless of proportion selected, presence of an exclusion restriction, or distribution of the errors. In some cases, in particular when the selected sample is smaller and in the absence of an exclusion restriction, the reduction in MSE from using the SPML methods is dramatic. Tables A.3 and A.4 make it evident that the lower MSE is a result of both lower bias and lower variance.

It is clear that both the Probit and the Klein and Spady estimators using spline methods suffer badly when an exclusion restriction is missing. The reason for this result is the severe multicollinearity that is caused by an absence of an exclusion restriction and using a flexible

Table 2: MSE Results (Study 1)

Exclusion Restriction	**Yes**			**No**		
Proportion Selected	**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
			$U \sim Norm$			
Probit - Linear	0.62	0.33	0.18	34.3	7.52	1.46
Klein and Spady - Linear	0.57	0.31	0.17	55.4	11.7	1.45
Probit - Spline	0.70	0.35	0.19	9660	822	28.0
Klein and Spady - Spline	0.66	0.34	0.18	19600	2150	39.7
SPML	**0.23**	**0.14**	**0.09**	**0.25**	**0.14**	**0.09**
			$U \sim t_4$			
Probit - Linear	0.72	0.35	0.18	17.9	5.78	1.33
Klein and Spady - Linear	0.76	0.35	0.17	57.0	12.4	1.36
Probit - Spline	0.79	0.38	0.19	2400	22.1	18.0
Klein and Spady - Spline	0.83	0.38	0.18	8460	102	23.3
SPML	**0.28**	**0.15**	**0.09**	**0.29**	**0.15**	**0.09**
			$U \sim SkewNorm$			
Probit - Linear	0.65	0.37	0.19	20.9	5.02	1.01
Klein and Spady - Linear	0.67	0.37	0.22	60.6	11.2	1.47
Probit - Spline	0.70	0.38	0.20	2670	130	3.32
Klein and Spady - Spline	0.73	0.39	0.23	7820	639	8.98
SPML	**0.28**	**0.16**	**0.10**	**0.29**	**0.16**	**0.11**
			$U \sim MixNorm$			
Probit - Linear	0.48	0.29	0.19	107	30.4	1.65
Klein and Spady - Linear	0.35	0.25	0.19	53.9	31.8	2.25
Probit - Spline	0.71	0.34	0.20	74400	67100	24.1
Klein and Spady - Spline	0.42	0.34	0.20	94700	74300	82.4
SPML	**0.22**	**0.11**	**0.09**	**0.24**	**0.11**	**0.09**

function (in the form of cubic B-splines) of the fitted values from the first stage. This provides greater opportunity for near-perfect multicollinearity between the regressors of interest and the selection bias terms. When an exclusion restriction is present, this multicollinearity does not materialize and the spline methods perform similarly to the linear estimators. Note that the SPML method does not suffer a similar fate despite also using splines in the second stage. This is because the first-stage ML estimation elicits nonlinearities in $\hat{m}(w_i)$, which act to break the multicollinearity.

Table 3 shows the MSE results for the second study, where the setup is modified to favor

Table 3: MSE Results (Study 2)

Exclusion Restriction	**Yes**			**No**		
Proportion Selected	**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
Probit - Linear	0.44	0.29	0.17	10.1	2.69	0.72
Klein and Spady - Linear	0.44	0.29	0.17	9.51	2.53	0.70
Probit - Spline	0.46	0.29	0.19	1210	131	7.96
Klein and Spady - Spline	0.45	0.29	0.19	948	111	6.87
SPML	**0.24**	**0.16**	**0.12**	**0.25**	**0.16**	**0.12**

the Heckman Estimator (here called 'Probit - Linear'). Notably, the linear estimators have similar performance across the board, and while they fare slightly worse than the SPML estimator under the most favorable conditions, they are significantly outperformed in the absence of an exclusion restriction. The spline methods show reasonable performance when an exclusion restriction is present, being only marginally worse than the correctly specified parametric (linear) estimators. However, as was the case in the previous study, these spline methods perform poorly when an exclusion restriction is missing.

Based on the results in tables 2 and 3, we find the SPML estimator to be a promising alternative to existing methods. The performance of the SPML estimator is persistently good even in the absence of an exclusion restriction, reducing mean squared error drastically in some cases and giving both low bias and low variance across all parameter settings. 3 shows that the SPML estimator successfully approximates the inverse Mills ratio, without the estimator imposing any distributional assumption on the error term, while 2 shows that SPML is more robust to alternative distributions of the error terms and to the fraction of the sample selected compared to existing methods.

3.3.1 Inference Results

Table 4 displays coverage results for the SPML estimator based on the standard error estimator given in Section 2.2. The parameter settings are identical to those used in study 1 of the empirical Monte Carlo. The 'Mean' measure is constructed by first calculating the coverage

Table 4: Coverage Results

Exclusion Restriction		Yes			No		
Proportion Selected		**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
Nominal Level	Measure			$U \sim Norm$			
90%	Mean	90.1	89.4	89.1	90.2	89.4	89.1
	MAD	0.83	1.17	1.11	0.86	1.26	1.38
95%	Mean	95.4	94.8	94.6	95.4	94.9	94.6
	MAD	0.63	0.95	0.71	0.66	0.74	0.74
				$U \sim t_4$			
90%	Mean	89.9	89.8	89.3	89.9	89.9	89.4
	MAD	1.14	1.40	1.20	1.09	0.63	1.35
95%	Mean	95.0	95.0	94.9	95.2	94.9	94.8
	MAD	0.57	0.58	1.08	0.55	0.63	1.18
				$U \sim SkewNorm$			
90%	Mean	89.6	89.5	89.8	89.8	89.4	89.8
	MAD	1.22	1.23	1.22	1.22	0.98	0.89
95%	Mean	94.8	94.6	94.9	94.8	94.8	95.1
	MAD	1.18	1.14	0.72	1.17	0.94	0.71
				$U \sim MixNorm$			
90%	Mean	89.4	88.6	89.2	89.4	88.7	89.1
	MAD	0.86	1.21	1.55	0.92	1.61	1.51
95%	Mean	94.5	94.3	94.4	94.7	94.5	94.7
	MAD	0.82	1.02	1.03	0.69	1.17	0.80

probability over Monte Carlo samples for each slope coefficient of the outcome equation (excluding any coefficients related to bias terms) and then taking the mean of these coverage probabilities over each of the slope coefficients. The mean absolute deviation, 'MAD', is calculated by taking the mean of the absolute difference between each slope coefficient coverage probability and the nominal level.

The results in Table 4 show that the standard error estimator is able to reflect the uncertainty in the SPML estimator with reasonable accuracy. It is encouraging to see that there are no major discrepancies across the error distributions or across the proportion of observations selected. From the 'Mean' results, it appears that the confidence intervals are well centered at the nominal level; however, the 'MAD' results show that there is some degree of dispersion around this level across the different slope coefficients. While this is to be expec-

ted, it does raise the question of whether alternative approaches may yield more accurate measures of uncertainty. For example, a bootstrap-based method may provide more accurate finite-sample coverage results. Unfortunately, no such schemes for semiparametric estimators using ML estimators currently exist. However, [Cattaneo and Jansson, 2019] show the consistency of a 'cross-fit bootstrap' for semiparametric two-step estimators where conventional nonparametric methods are used in the first step. Although their results cannot be generalized to our setting, they do provide hope for the suitability of a bootstrap procedure here.

4 Conclusion

This paper provides a novel method for rectifying the bias in canonical sample selection models using a semiparametric machine learning estimator. Identification is driven by nonlinearities in the selection function along with plausible exclusion restrictions; however, in the Monte Carlo study, the estimator is shown to be nearly unaffected by the presence, or absence, of an exclusion restriction. The performance improvement from using the SPML estimator is significant: the mean squared error is reduced in all settings in an empirical Monte Carlo - in some cases, radically so - and is even found to outperform the Heckman two-step estimator in an environment otherwise suited to this parametric estimator. Simple results on the asymptotic distribution of the SPML estimator allow for the easy calculation of standard errors rendering inference straightforward. The practical takeaway of this paper is that a new and easily implementable method for sample selection models is readily available that is robust and persistently well-performing in a range of realistic data settings.

Appendix

To show the $\sqrt{n}$-consistency and asymptotic normality of the SPML estimator, we use the results of [Newey, 1997] together with [Foster and Syrgkanis, 2019]. Throughout, we let (Y, W, U, V) be a random vector whose law satisfies (1) and (2) in Section 1.

Ignoring the error from the first-stage estimation of $m(w_i)$, the estimator is a standard partially-linear model, as considered in [Newey, 1997]. Theorem 9 of that paper proves the $\sqrt{n}$-consistency and asymptotic normality for a partially-linear model based on regression splines under a set of standard assumptions (given below as Assumptions 1 - 4). However, we must also show that the first-stage ML estimation error of $m(w_i)$ is asymptotically negligible when sample-splitting is used. To this end, we appeal to Theorems 1 and 3 of [Foster and Syrgkanis, 2019].

To make the dependence of $\hat{\beta}$ on $\hat{m}$ explicit, we write the SPML estimator as $\hat{\beta}(\hat{m})$ and denote the estimator of β if m is known as $\hat{\beta}(m)$. Taken together, Theorems 1 and 3 of [Foster and Syrgkanis, 2019] show

$$\begin{aligned} E\left[\left(\hat{\beta}\left(\hat{m}\left(W\right)\right)-\beta\left(m\left(W\right)\right)\right)^{2}\right] =& O\left(E\left[\left(\hat{\beta}\left(m\left(W\right)\right)-\beta\left(m\left(W\right)\right)\right)^{2}\right]\right) \\ &+ O\left(E\left[\left(\hat{m}\left(W\right)-m\left(W\right)\right)^{2}\right]\right)^{4}. \end{aligned}$$

Therefore, if $\hat{m}$ converges sufficiently quickly relative to the convergence rate of $\hat{\beta}$, the estimation error from the first stage is asymptotically negligible. Since the first term is of order n^{-1}, we require $E\left[\left(\hat{m}\left(W\right)-m\left(W\right)\right)^{2}\right]=o\left(n^{-1/4}\right)$, which is satisfied by the majority of off-the-shelf ML algorithms.

To use the results of [Foster and Syrgkanis, 2019], we must verify Assumptions 1 - 4 of that paper in this setting. By the Frisch-Waugh-Lovell theorem, the target population loss function can be written in the [Robinson, 1988] partialled-out form:

$$E\left[\left(\left(Y-\varphi\left(m\left(W\right)\right)\theta_{Y}\right)-\left(X'-\varphi\left(m\left(W\right)\right)'\theta_{X}\right)\beta\right)^{2}\right]$$

where θ_Y and θ_X are coefficients from a series-based regression of Y on $m(W)$ and of X on $m(W)$, respectively. It is straightforward to show that this loss function satisfies the Neyman orthogonality condition imposed by Assumption 1 of [Foster and Syrgkanis, 2019]. Furthermore, as discussed in [Foster and Syrgkanis, 2019], since the target population loss function is the squared loss, Assumptions 2, 3, and 4, which concern first order optimality of the target parameter, strong convexity of the target loss, and higher-order smoothness of the target loss, respectively, are also satisfied.

Assumptions

Let P denote the dimension of X, L the number of B-splines used, and $\hat{m}$ the first-stage ML estimator of m. For the $\sqrt{n}$-consistency and asymptotic Normality of the SPML estimator, we require

1. $\{y_i, w_i\}$ are i.i.d., $0 < Var[Y|X, m(W)] < \infty$, and $E\left[(Y - E[Y|X, m(W)])^4 \middle| X, m(W)\right] < \infty$.
2. The support of $(X, m(W))$ is a known Cartesian product of compact connected intervals on which $(X, m(W))$ has a probability density function that is bounded away from zero.
3. $E[Y|X, m(W)]$ is continuously differentiable of order $s > P + 1$ on the support of $(X, m(W))$.
4. $nL^{-2s/(r+1)} \to 0$ and $nL^{-2} \to \infty$ as $n \to \infty$.
5. $E\left[(\hat{m}(W) - m(W))^2\right] = o\left(n^{-1/4}\right)$.

Online Appendix A

Table A.1: Full Sample Regression Results

	Dependent variable:
	Sentence
White	0.19^{***}
	(0.03)
Male	0.26^{***}
	(0.04)
Age	0.06^{***}
	(0.02)
Age^2	-0.03^{***}
	(0.01)
Previous Criminality [2]	0.22^{*}
	(0.12)
Previous Criminality [3]	1.43^{***}
	(0.12)
Public Defender	-0.004
	(0.06)
Private Attorney	-0.33^{***}
	(0.03)
Crime Type [Peace]	-0.39^{***}
	(0.05)
Crime Type [Property]	-0.33^{***}
	(0.03)
Crime Type [Violent]	-0.46^{***}
	(0.06)
Crime Severity [2]	-0.06
	(0.07)
Crime Severity [3]	0.48^{***}
	(0.04)
Crime Severity [4]	1.39^{***}
	(0.07)
Judge Sentencing Severity	0.13^{***}
	(0.01)
Observations	14 705
Note:	*p<0.1; **p<0.05; ***p<0.01

Notes: This table reports estimates of the outcome equation using the SPML estimator proposed in this paper and the data described in Section 3.1. Standard errors, calculated using the procedure outlined in Section 2.2, are reported in parentheses. All variables are standardised to have zero mean and unit variance.

Table A.2: Test of Randomisation

	Dependent variable:	
	Conviction Tendency	Sentencing Severity
White	0.003	−0.01
	(0.08)	(0.07)
Male	−0.04	−0.06
	(0.04)	(0.04)
Age	−0.01	−0.01
	(0.02)	(0.02)
Age^2	0.01*	0.01
	(0.01)	(0.01)
Previous Criminality [2]	0.10*	0.03
	(0.05)	(0.05)
Previous Criminality [3]	0.09	0.06
	(0.07)	(0.07)
Private Attorney	−0.02	0.03
	(0.05)	(0.05)
Public Defender	−0.05	−0.005
	(0.22)	(0.25)
Crime Type [Peace]	−0.03	0.01
	(0.13)	(0.09)
Crime Type [Property]	−0.03	0.01
	(0.09)	(0.06)
Crime Type [Violent]	0.06	−0.004
	(0.14)	(0.10)
Crime Severity [2]	0.10	−0.003
	(0.11)	(0.07)
Crime Severity [3]	0.04	−0.02
	(0.07)	(0.04)
Crime Severity [4]	−0.04	0.002
	(0.09)	(0.08)
Observations	14 705	14 705
Adjusted R^2	0.001	0.0001
F Statistic (df = 14; 14 690)	0.97	0.49
Note:		*p<0.1; **p<0.05; ***p<0.01

Notes: This table reports results from two linear regressions using the data described in Section 3.1. In column (1), the dependent variable is the residualised leave-one-out conviction tendency of the randomly assigned judge. In column (2), the dependent variable is the residualised leave-one-out sentencing severity of the randomly assigned judge. Standard errors for the estimated coefficients are reported in parentheses and are clustered at the judge level. All variables are standardised to have zero mean and unit variance.

Table A.3: Bias Results (Study 1)

Exclusion Restriction	**Yes**			**No**		
Proportion Selected	**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
			$U \sim Norm$			
Probit - Linear	2.29	1.75	1.26	17.3	8.46	3.86
Klein and Spady - Linear	2.21	1.70	1.22	21.5	9.92	3.74
Probit - Spline	2.43	1.77	1.28	283	86.5	16.5
Klein and Spady - Spline	2.37	1.73	1.24	365	112	16.1
SPML	1.25	0.94	0.72	1.31	0.98	0.75
			$U \sim t_4$			
Probit - Linear	2.42	1.68	1.23	11.9	7.30	3.77
Klein and Spady - Linear	2.48	1.67	1.21	20.4	9.89	3.64
Probit - Spline	2.49	1.71	1.24	135	44.6	14.1
Klein and Spady - Spline	2.58	1.72	1.24	228	71.3	14.1
SPML	1.39	0.99	0.69	1.39	0.99	0.70
			$U \sim SkewNorm$			
Probit - Linear	2.30	1.80	1.32	13.0	6.83	3.25
Klein and Spady - Linear	2.35	1.81	1.43	21.6	9.86	3.79
Probit - Spline	2.38	1.82	1.33	146	34.1	5.84
Klein and Spady - Spline	2.47	1.86	1.44	231	60.7	8.30
SPML	1.39	1.03	0.84	1.44	1.07	0.84
			$U \sim MixNorm$			
Probit - Linear	2.15	1.55	1.28	32.4	16.6	4.06
Klein and Spady - Linear	1.78	1.41	1.27	22.9	16.4	4.45
Probit - Spline	2.68	1.71	1.28	625	724	15.2
Klein and Spady - Spline	2.00	1.58	1.29	885	856	21.0
SPML	1.23	0.78	0.71	1.29	0.78	0.73

Table A.4: SD Results (Study 1)

Exclusion Restriction	**Yes**			**No**		
Proportion Selected	**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
			$U \sim Norm$			
Probit - Linear	0.62	0.33	0.17	33.8	7.37	1.37
Klein and Spady - Linear	0.57	0.31	0.16	54.9	11.5	1.39
Probit - Spline	0.69	0.35	0.18	9600	812	27.9
Klein and Spady - Spline	0.65	0.33	0.17	19600	2130	39.6
SPML	0.22	0.13	0.08	0.23	0.14	0.08
			$U \sim t_4$			
Probit - Linear	0.72	0.35	0.17	17.6	5.50	1.20
Klein and Spady - Linear	0.76	0.35	0.17	55.7	11.8	1.27
Probit - Spline	0.78	0.37	0.18	2360	21.6	18.0
Klein and Spady - Spline	0.82	0.38	0.18	8430	102	23.3
SPML	0.28	0.14	0.08	0.29	0.15	0.08
			$U \sim SkewNorm$			
Probit - Linear	0.65	0.36	0.18	20.4	4.89	0.93
Klein and Spady - Linear	0.66	0.36	0.20	59.4	11.0	1.37
Probit - Spline	0.69	0.38	0.19	2630	129	3.32
Klein and Spady - Spline	0.72	0.38	0.21	7780	634	8.98
SPML	0.27	0.15	0.10	0.28	0.15	0.10
			$U \sim MixNorm$			
Probit - Linear	0.47	0.29	0.18	107	30.1	1.53
Klein and Spady - Linear	0.34	0.25	0.18	53.7	31.7	2.15
Probit - Spline	0.71	0.33	0.19	74400	67100	24.1
Klein and Spady - Spline	0.42	0.34	0.19	94500	74300	82.4
SPML	0.21	0.11	0.09	0.23	0.11	0.09

Table A.5: Bias Results (Study 2)

Exclusion Restriction	**Yes**			**No**		
Proportion Selected	**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
Probit - Linear	1.97	1.66	1.27	10.4	5.33	2.73
Klein and Spady - Linear	1.97	1.66	1.27	10.2	5.28	2.67
Probit - Spline	1.97	1.67	1.34	84.4	33.2	9.21
Klein and Spady - Spline	1.97	1.68	1.35	77.7	32.1	8.48
SPML	1.26	1.09	0.99	1.31	1.15	1.00

Table A.6: SD Results (Study 2)

Exclusion Restriction	**Yes**			**No**		
Proportion Selected	**40%**	**60%**	**84%**	**40%**	**60%**	**84%**
Probit - Linear	0.44	0.29	0.17	10.1	2.69	0.71
Klein and Spady - Linear	0.44	0.28	0.17	9.50	2.53	0.69
Probit - Spline	0.46	0.29	0.19	1210	130	7.94
Klein and Spady - Spline	0.45	0.29	0.19	948	110	6.85
SPML	0.21	0.12	0.08	0.22	0.13	0.08

Online Appendix B

The following restrictions are made for the data used in the empirical Monte Carlo. The initial dataset contains 1.2 million observations. First, all cases which are settled by a plea bargain are removed. For these cases, the sentence imposed is typically decided by the prosecutor and the defence counsel, not the judge, so are not relevant for the study. This results in an 87% reduction in the sample size. Only cases which reach a final trial are considered - that is cases that are not dismissed prior to trial - resulting in an additional 61% loss of observations. We further restrict the sample to ensure a level of homogeneity across cases. As such, we keep only single-charge cases where the offence is a misdemeanour and is classed as either a property crime, a disturbance of the peace, a violent crime, or is drug related. This removes 47% of the remaining observations.

As the conviction rate and sentencing severity of each judge are used in the empirical Monte Carlo analysis as exclusion restrictions, a sufficient number of observations per judge are needed to ensure the estimated measures of judge leniency are reliable. Consequently, the sample is restricted to cases which were overseen by a judge who hears at least 50 cases in the remaining sample; the same restriction is imposed by [Bhuller et al., 2020]. A further 41% of observations are lost due to this. Finally, only defendants with a reported age and race and who are either black or white are considered; resulting in a 22% reduction in sample size.

References

[Ackerberg et al., 2012] Ackerberg, D., Chen, X., and Hahn, J. (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94(2):481–498. Publisher: MIT Press.

[Angrist et al., 1999] Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67. Publisher: Wiley Online Library.

[Angrist and Krueger, 1995] Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235. Publisher: Taylor & Francis Group.

[Bhuller et al., 2020] Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324. Publisher: The University of Chicago Press Chicago, IL.

[Cattaneo and Jansson, 2019] Cattaneo, M. D. and Jansson, M. (2019). Average density estimators: Efficiency and bootstrap consistency. *arXiv preprint arXiv:1904.09372*.

[Chang, 2020] Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*.

[Chernozhukov et al., 2018] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. Publisher: Wiley Online Library.

[Dobbie et al., 2018] Dobbie, W., Goldin, J., and Yang, C. S. (2018). The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40.

[Dufour and Jasiak, 2001] Dufour, J.-M. and Jasiak, J. (2001). Finite Sample Limited Information Inference Methods for Structural Equations and Models With Generated Regressors. *International Economic Review*, 42(3):815–844.

[Dyke, 2007] Dyke, A. (2007). Electoral cycles in the administration of criminal justice. *Public Choice*, 133(3-4):417–437. Publisher: Springer.

[Foster and Syrgkanis, 2019] Foster, D. J. and Syrgkanis, V. (2019). Orthogonal Statistical Learning. *arXiv:1901.09036*. arXiv: 1901.09036.

[Gallant and Nychka, 1987] Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, pages 363–390. Publisher: JSTOR.

[Heckman, 1979] Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

[Klein and Spady, 1993] Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, pages 387–421. Publisher: JSTOR.

[Newey, 1997] Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168. Publisher: Elsevier.

[Newey, 2009] Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12:S217–S229. Publisher: Wiley Online Library.

[Newey et al., 1990] Newey, W. K., Powell, J. L., and Walker, J. R. (1990). Semiparametric Estimation of Selection Models: Some Empirical Results. *American Economic Review*, 80(2):324–328. Publisher: American Economic Association.

[Robinson, 1988] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, pages 931–954. Publisher: JSTOR.

[Silveira, 2017] Silveira, B. S. (2017). Bargaining with asymmetric information: An empirical study of plea negotiations. *Econometrica*, 85(2):419–452. Publisher: Wiley Online Library.

[Vella, 1998] Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources*, 33(1):127–169. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].