AStA Advances in Statistical Analysis
https://doi.org/10.1007/s10182-026-00566-5

ORIGINAL PAPER

Modeling with sensitive variables

Felix Chan[1] · **László Mátyás**[2,3] · **Ágoston Reguly**[4,5]

Received: 16 September 2025 / Accepted: 15 June 2026

Abstract
The paper deals with models in which the dependent variable, some explanatory variables, or both represent unobserved sensitive data. We introduce a novel discretization method that reveals sufficient information from the sensitive variable to approximate the parameter(s) of interest. Multiple discretization schemes are employed, and we show convergence in distribution for the unobserved variable. The asymptotic properties of the OLS estimator for linear models are derived and discussed. Monte Carlo simulations support our theoretical findings and demonstrate finite-sample properties. Finally, we contrast our method with other alternative methods for estimating the Australian gender wage gap.

Keywords Sensitive variable · Discretization · Interval censored variables

1 Introduction

Over the last decade, there has been an explosion in the amount of data collected on individuals, firms, the economy, and society. Governments and the private sector gather large amounts of information that researchers and analysts use on a daily basis. Privacy constraints to share these data or willingness to respond in case of surveys,

Felix Chan and László Mátyás contributed equally to this work.

✉ Ágoston Reguly
agoston.reguly@uni-corvinus.hu

Felix Chan
felix.chan@cbs.curtin.edu.au

László Mátyás
matyas@ceu.edu

1 School of Accounting, Economics and Finance, Curtin University, Kent Street, 6102 Bentley, WA, Australia

2 Department of Economics, Central European University, Quellenstraße 51, 1100 Vienna, Austria

3 Department of Economics, Central European University, Nador u. 15, Budapest 1051, Hungary

4 Institute of Economics, Corvinus University of Budapest, Fovam ter 8, Budapest 1093, Hungary

5 Scheller College of Business, Georgia Institute of Technology, 800 W Peachtree St NW, Atlanta, GA 30308, USA

Published online: 13 July 2026

often prevent access to sensitive data in their original form. We introduce here a novel sampling method that naturally increases the willingness to answer through discretization, while minimizes information loss from an estimation perspective.

Income is perhaps one of the most commonly used sensitive variable in economics. Instead of providing the sensitive data itself or asking directly for the actual weekly personal income, we work with discrete categories such as, for example, '*weekly personal income below* \$100, *between* \$100 *and* \$400, *or above* \$400'. It is well-known in the econometric literature that this simple discretization leads to modeling problems (see, e.g., Hsiao 1983, Manski and Tamer 2002, Beresteanu and Molinari 2008, Beresteanu et al. 2011, Pacini 2019, or Abrevaya and Muris 2020), which we solve by using multiple discretization schemes. In the context of income, instead of using only one discretization scheme, we use multiple discretizations by changing the interval boundaries, but keeping the number of intervals fixed. For example, discretization scheme (2): below \$50, between \$50 and \$350 or above \$350; discretization scheme (3): below \$150, between \$150 and \$450 or above \$450, etc. We call the method *split sampling*,[1] and show convergence in distribution for the discretized sensitive variable to the original unobserved variable as both the number of observations and the number of discretization schemes go to infinity.

Our solution is simple in the sense that each discretization scheme allows or enhances the willingness to respond,[2] while the properties of the underlying distribution are recovered by the combination of the resulting discretized variable. Building on our result of convergence in the distribution, we investigate three types of regression: i) the sensitive variable is an explanatory variable; ii) the sensitive variable is the outcome variable; and iii) sensitive variables are on both sides of the regression model. We use our split sampling method to identify parameters in such linear models and to derive the properties of the OLS estimator. The solution for identifying parameters in such a context is threefold: i) our method leverages the additional information from the split samples to estimate the underlying distribution of the latent variables. This allows us to change the unit of observations from individual responses to a function of the number of the split samples; ii) the way we create the (split) samples to work with allows us to mitigate random sampling, and iii) we rewrite the linear regression model as the conditional expectation of the known discretization intervals, which allows us to identify the parameters of interest. The definition of the conditional expectation depends on where the sensitive variable(s) appear in the regression. We show that this general method yields point identified parameters, with a consistent,[3] asymptotically normal estimator in linear models when both the number of observations and the number of split samples go to infinity. We emphasize that our estimator has favorable

[1] The term *split sampling* in this paper is not related to the technique occasionally used in chromatography (see e.g., Schomburg et al., 1977) or cross-validation methods in machine learning, which splits the initial sample into folds.

[2] Evidence across several studies supports the idea that reducing choice complexity increases survey response rates. Studies on questionnaire design find that simpler, integrated question formats yield more complete responses, further illustrating that reduced cognitive burden enhances survey participation, see e.g. Ziegenfuss et al. (2021)

[3] This consistency is nonstandard as we require both the number of observations and the number of split samples to go to infinity.

theoretical and finite-sample properties, and even with a few discretization schemes (5 or 10), one can obtain a meaningful estimate(s). We provide multiple Monte Carlo simulations for this purpose and share practical guidance for designing the discretization. Finally, we apply our procedure to wage data from the Australian Tax Office (ATO) to investigate the gender wage gap in Australia.

Our approach relates to the literature on partially identified parameters (e.g., see Manski 2003, Manski and Tamer 2002, Tamer 2010, Molinari et al. 2020, Abrevaya and Muris 2020, Pacini 2019, and Wang and Chen 2022). This literature starts from the fact that the variable of interest is discretized (also called interval censored) for some reason (e.g., Pacini, 2019 discuss income surveys). Manski and Tamer (2002) show that the conditional expectation function and the regression parameters in such context cannot be point identified in general. They show how to derive set identification when the discretized variable is on the right-hand side. Beresteanu and Molinari (2008), Bontemps et al. (2012) discuss the case when the outcome variable is discretized, while Beresteanu et al. (2011) cover the case when both outcome and covariate(s) are discretized. The main toolkit of partially identified parameters in such context is moment (in)equalities.[4] The main advantage of these methods is that they do not require adding any noise to the variables or further distributional assumptions and allow valid statistical inferences on the parameters of interest. The magnitudes and signs of the estimated parameter vector can be interpreted in the same way as the classical regression coefficients. However, the main drawback is that the estimated parameters are not point identified, but rather up to a set to which the parameter vector belongs. In many empirical applications, this set may be too wide to be meaningful. Our method circumvents the problem of using interval censored variables and shows that if a multiple discretization scheme is applicable, this additional information can be used to get point estimates.

Our approach is also related to statistical work on combining multiple discretization schemes or 'binnings' to estimate the density of a random variable. In nonparametric density estimation, Scott (1985) proposes averaged shifted histograms (ASH), which average histograms computed on the same data under multiple bin origins, thereby reducing sensitivity to bin placement and improving density recovery. Jones (1989) further studies discretization effects in kernel density estimation and connects discretized representations to ASH-type constructions. In contrast to these post-processing methods, in our setting the underlying continuous values are not observed.

This connects to the coarse data framework of Heitjan and Rubin (1991) or Heitjan (1997) and the multiple imputation approach for heaping or rounded data in Heitjan and Rubin (1990), which deals with imprecise survey respondence that naturally result in discretized values. Recent work on inference under grouping underscores both the inferential challenges created by coarse measurement and the need for methods that

[4] Moment (in)equality models generalize the problems to multiple equations and/or inequalities, see for examples, Chernozhukov et al. (2007) or Andrews and Soares (2010). Beresteanu and Molinari (2008) show the asymptotic properties of such partially identified parameters. Imbens and Manski (2004), Chernozhukov et al. (2007), and Kaido et al. (2019), among others, derive confidence intervals for these set-identified parameters. These methods are feasible ways to estimate parameter sets for a given conditional expectation function without any further assumptions.

better recover distributional features from coarsened observations, including density estimation directly from grouped data (Jang and Loh 2010) and modeling/reporting mechanisms for heaped self-reports (Crawford et al. 2015). In addition, identifiability considerations under interval censoring highlight that a fixed discretization structure may be insufficient for full recovery without additional information (Oller et al. 2007), motivating our use of deliberately designed, overlapping discretizations to obtain point identification and consistent estimation without releasing the underlying sensitive values.

The paper is organized as follows: Sect. 2 introduces the discretization problem, the parameters of interest, and the notation[5]. Section 3 describes the proposed split sampling approach. We discuss the *shifting* method in detail and show convergence in distribution to the unknown underlying distribution. In Sect. 4, we identify parameters in a multivariate linear regressions and derive the asymptotic properties of the least square estimator for the cases of discretized variables on the right or left hand side. We also outline the case of discretization on both side of a regression model. Section 5 presents Monte Carlo evidence, practical guidance along with an empirical application on the Australian gender wage gap. Section 6 concludes. Finally, let us note that the paper uses heavy notations and readers are referred to Tables A24-A25 in the online supplement, for a comprehensive list.

2 The discretization problem

Consider $Z \sim f_Z(z; a_l, a_u)$ an i.i.d. random variable, where $f_Z(z; a_l, a_u)$ denotes the probability density function (pdf) with support $[a_l, a_u]$, where $a_l, a_u \in \mathbb{R}$, $a_l < a_u$. We assume that $f_Z(\cdot)$ is unknown and continuous. Let Z_i be $i = 1, \ldots, N$, realizations of Z. The sensitive variable Z_i is **not** observed by the modeler, but one can work with a discretized version Z_i^*, through the following discretization process:

$$Z_i^* = \begin{cases} v_1 & \text{if } c_0 \leq Z_i < c_1 \quad \text{or} \quad Z_i \in \mathcal{C}_1 = [c_0, c_1) \quad \text{1st interval,} \\ \vdots & \vdots \\ v_m & \text{if } c_{m-1} \leq Z_i < c_m \quad \text{or} \quad Z_i \in \mathcal{C}_m = \left[c_{m-1}, c_m\right), \\ \vdots & \vdots \\ v_M & \text{if } c_{M-1} \leq Z_i < c_M \quad \text{or} \quad Z_i \in \mathcal{C}_M = \left[c_{M-1}, c_M\right) \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{last interval,} \end{cases} \tag{1}$$

where $v_m \in \mathcal{C}_m$, $m = 1, \ldots, M$ is the assigned value for each interval. The value v_m can be a measure of centrality (e.g., mid-point) or an arbitrarily assigned value within its interval. The central point for the paper is that M which denotes the number of intervals is **fixed** and **known**. Also, the discretization intervals $\mathcal{C}_m$, are set by the survey maker, and we take them as given. It is helpful to see the discretization intervals

[5] In the online supplement, we provide notation tables under Section A12.

as independent from the random variable Z, but this is not a necessary requirement in our case. For simplicity, we use the terms interval and class interchangeably for $\mathcal{C}_m$.

2.1 The parameter of interest

We are interested in the regression parameter β, in simple linear regressions,

$$\mathbb{E}[Y|X] = X'\beta. \tag{2}$$

For the sake of simplicity, we introduce our notation in scalar terms; β stands for a scalar parameter belonging to a subset of a compact finite-dimensional space ($\mathcal{B} \subset \mathbb{R}$), while Y, X are scalars. Facing the discretization of a variable can affect the identification of β in *three* ways, that is well-known in the partial identification literature. In the first and simplest case, the explanatory variable (X^*) is discretized; the second possibility is when the outcome variable (Y^*) is discretized; and lastly, where both variables are discretized (Y^*, X^*).

To show an example for the identification problem, let us consider the following model:

$$Y = X\beta + u. \tag{3}$$

Now, instead of observing X, we observe the discretized version of X^* and one plugs in the mid-point values, resulting in,

$$Y = X^*\beta + \left(X - X^*\right)\beta + u. \tag{4}$$

Under the usual assumptions, the *naive* OLS estimator for β results in $\hat{\beta}^n = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}Y = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}X\beta + \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}u$, that is in general $\hat{\beta}^n \nrightarrow \beta$ as $X^* \neq X$. Furthermore, the effect of the discretization is contagious; thus, if one uses further variable(s) W with parameter(s) γ, that are correlated with X, then $\hat{\gamma}^n \nrightarrow \gamma$. This result implies that the practice of assigning arbitrary values (e.g., mid-points) to the discretized variable leads to biases in the parameter estimates even when the discretized variables are controls. Note that the size of the bias depends on the underlying distribution of X and the discretization process.[6]

As mentioned, this phenomenon is known in the partial identification literature, and Manski and Tamer (2002) were the first to show that with discretized variables, only *set identification* is possible. In this paper, we offer an alternative approach in which the discretization process approximates the β parameters, and in the limiting case—where one has an infinite discretization scheme—point identifies β. For completeness, however, based on Manski and Tamer (2002) and Molinari et al. (2020), we work out the set-identified regions for the parameter of interest in the online supplement under Section A4 to draw similarities between the two approaches.

[6] We provide more detailed discussion on this topic with proofs in the online supplement, Section A6.

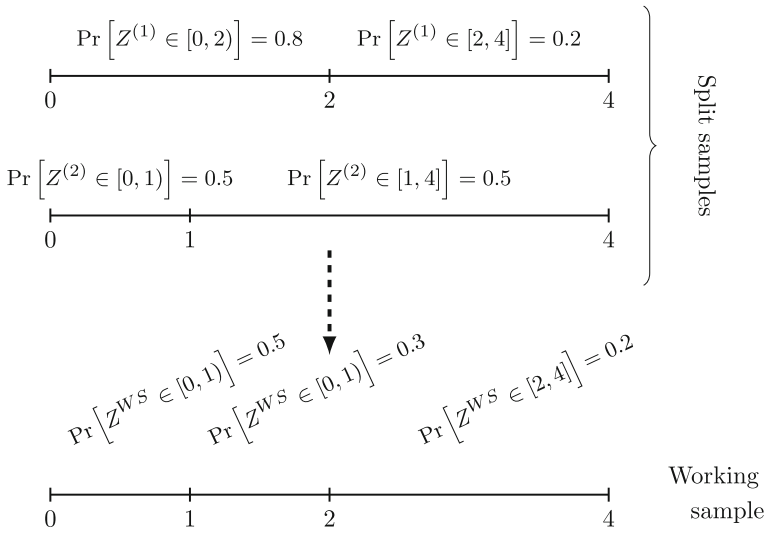

Fig. 1 Illustrative example for split sampling

3 Split sampling

The key to our method is the use of split sampling, which recovers the marginal distribution of the *unobserved* Z. Split sampling consists of two steps: 1) discretize the sensitive variable into multiple samples; and 2) *randomly* combine the samples by using the information on the known interval boundaries. Although the idea is simple, it requires some heavy notation for rigorous proofs. Thus first, let us take a very simple illustrative example and then discuss the univariate case. As a last step, we extend our approach to the multivariate models.

Let Z be a random variable with the following pdf, with support boundaries $a_l = 0$ and $a_u = 4$,

$$\Pr(Z \in [a,b)) = \begin{cases} 0.5, & \text{if } a = 0,\ b = 1 \\ 0.3, & \text{if } a = 1,\ b = 2 \\ 0.2, & \text{if } a = 2,\ b = 4. \end{cases} \tag{5}$$

As the first step, let us discretize Z with two intervals $M = 2$ and let us use two discretization schemes or split samples. The first split sample discretizes the Z as [0, 2) and [2, 4], while in the second split sample [0, 1) and [1, 4]. The top section of Fig. 1 referred to as 'split samples' visualizes this setup. By taking these two separate sets of information, we cannot recover the true probabilities for Z.

However, in the second step, we combine information from the two split samples into the 'working sample', which is represented in the bottom part of Fig. 1. In this simple example, the working sample has 3 intervals, with boundaries [0, 1), [1, 2), and Andrews and Soares (2010); Bontemps et al. (2012). By combining the information from the split samples to the working sample, we can deduce the probabilities of the

non-observed intervals. Let us emphasize the each observation i is assigned to one split sample s, that is known by the modeler. The assignment of observations to split samples can be random, but it is not needed to be. The only requirement is that it is known which observation belongs to which split sample.

Now, let us generalize our method.

First, create multiple samples ($s = 1, \ldots, S$) and discretize them through different schemes while *fixing* the number of intervals (M) within each split sample.

Discretization for the s-th split sample looks exactly as the problem introduced in Equation (1), with the only difference being that the interval boundaries are different,[7]

$$Z_i^{(s)} = \begin{cases} v_1^{(s)} & \text{if } Z_i \in \mathcal{C}_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ \vdots & \vdots \qquad \text{1st interval for split sample } s, \\ v_m^{(s)} & \text{if } Z_i \in \mathcal{C}_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ v_M^{(s)} & \text{if } Z_i \in \mathcal{C}_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last interval for split sample } s. \end{cases} \tag{6}$$

The number of observations across split samples ($N^{(s)}$) can be the same or, more likely, different.

As a second step, we combine information into the *working sample*. The construction of the working sample's interval is simply given by the union of unique interval boundaries within each split sample,

$$\bigcup_{b=0}^{B} \mathcal{C}_b^{WS} = \bigcup_{s=1}^{S} \bigcup_{m=0}^{M} \mathcal{C}_m^{(s)}. \tag{7}$$

One can generalize the approach with any type of distribution. Let us write the probability for a random observation to be in a given class of a split sample,

$$\Pr\left(Z \in \mathcal{C}_m^{(s)}\right) = \Pr(Z \in \mathcal{S}_s) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f_Z(z)\mathrm{d}z, \tag{8}$$

where $\mathcal{S}_s$ denotes the s-th split sample. We can express, the probability of an observation falling into the working sample's $\mathcal{C}_b^{WS}$ interval as

$$\Pr\left(Z \in \mathcal{C}_b^{WS}\right) = \sum_{s=1}^{S} \Pr(Z \in \mathcal{S}_s) \sum_{m=1}^{M} \Pr\left(Z \in \mathcal{C}_b^{WS} \mid Z \in \mathcal{C}_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f_Z(z)\mathrm{d}z. \tag{9}$$

[7] To simplify the notation, we use instead of $Z^{*(s)}$ simply $Z^{(s)}$.

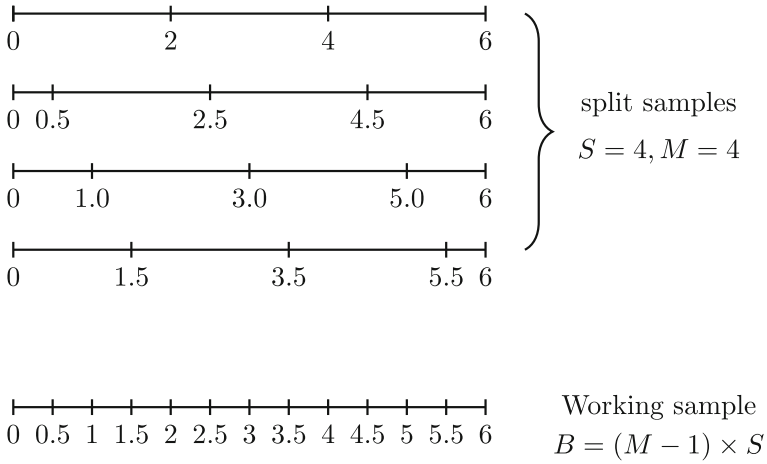

Fig. 2 Example for the shifting method

In practice, we recommend splitting the sample size into equal parts, thus $\Pr(Z \in \mathcal{S}_s) = 1/S$ and use uniform probability for $\Pr\left(Z \in \mathcal{C}_b^{WS} \mid Z \in \mathcal{C}_m^{(s)}\right)$. Then Equation (9) simplifies to

$$\Pr\left(Z \in \mathcal{C}_b^{WS}\right) = \frac{1}{S} \sum_{s=1}^{S} \sum_{\substack{m \\ \text{if } \mathcal{C}_b^{WS} \subset \mathcal{C}_m^{(s)}}} \frac{c_b^{WS} - c_{b-1}^{WS}}{c_m^{(s)} - c_{m-1}^{(s)}} \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f_Z(z) \mathrm{d}z. \tag{10}$$

3.1 The shifting method

The shifting method is *an* application of split sampling. It uses an equal distance discretization scheme as the benchmark and shifts the boundaries of the intervals with a certain value. The interval widths for the different split samples remain the same, except for the boundary intervals. As one increases the number of split samples, the size of the shift becomes smaller and smaller. This enables the mapping of the marginal probability distribution at the limit. First, we introduce the discretization process proposed for the shifting method, and then, as a second step, we show how to use the realizations of this discretization to create a synthetic variable that maps the original variable's marginal distribution.

Fig. 2 shows an illustrative example for split samples with $S = 4$ and $M = 4$ classes.[8]

[8] We can interpret this method in two ways: 1) We can consider the original discretization and take the number of classes as fixed. Then we shift the boundaries which results in adding an extra class for each split sample. Thus the original number of intervals is M and we add one for each (shifted) split sample. 2) We can look at each (shifted) split sample and take M as given by the shift and consider the first split sample as a special case where the size of the shift is zero, thus there are only $M - 1$ intervals. For the convenience of mathematical derivations, we use the second interpretation. As Fig. 2 shows, for the first split sample

To derive the properties of the shifting method, we define the uniform class widths for the first split sample as $\frac{a_u - a_l}{M-1}$. We split these intervals into S parts, thus we can shift the class boundaries S times. This defines the size of the shift as

$$h = \frac{a_u - a_l}{S(M-1)}. \tag{11}$$

This implies that the number of intervals in the working sample is $B = S(M-1)$. The boundary points for each split sample are given by,[9]

$$c_m^{(s)} = \begin{cases} a_l, & \text{if } m = 0, \\ a_l + (s-1)\frac{a_u - a_l}{S(M-1)} + (m-1)\frac{a_u - a_l}{M-1} & \text{if } 0 < m < M, \\ a_u, & \text{if } m = M. \end{cases} \tag{12}$$

As the second step, we introduce a synthetic random variable $Z^\dagger$ that maps the underlying marginal distribution for Z. Observations from a particular interval in the split sample s can end up in several candidate intervals from the working sample $\mathcal{C}_b^{WS}$. To make it more transparent, we define the discretization intervals $\mathcal{C}_b^{(s)}$ as sets of $\mathcal{C}_b^{WS}$,

$$\mathcal{C}_m^{(s)} = \begin{cases} \{\emptyset\}, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1}\{\mathcal{C}_b^{WS}\}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-1)(M-1)}^{s-1+(m-2)(M-1)}\{\mathcal{C}_b^{WS}\}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B}\{\mathcal{C}_b^{WS}\}, & \text{if } m = M. \end{cases} \tag{13}$$

We create the synthetic variable $Z^\dagger$ by assigning each discretized observation from its split sample $\mathcal{C}_m^{(s)}$ to one of its corresponding working sample's intervals $\mathcal{C}_b^{WS}$ with uniform probability. Note that this assignment mechanism does not assume any distributional form for Z itself, but it randomly assigns each value of $Z^{(s)}$ to the working sample's related interval with uniform probability to get $Z^\dagger$.

The random assignment mechanism can be written as[10]

$$\Pr\left(Z^\dagger \in \mathcal{C}_b^{WS} | Z^{(s)} \in \mathcal{C}_m^{(s)}\right) = \begin{cases} 1, & \text{if } s = 1 \text{ and } m = 1, \\ 1/(s-1), & \text{if } s \neq 1 \text{ and } m = 1, \\ 1/S, & \text{if } 1 < m < M, \text{ or} \\ 1/(S-s+1), & \text{if } m = M. \end{cases} \tag{14}$$

$(s = 1)$ there is one class less $(M-1)$ (or if we prefer, we can look at the benchmark where we shifted the boundaries with zero).

[9] In the online supplement under Section A1, Algorithm A1 summarizes how to construct split samples using the shifting method.

[10] An alternative notation would be to sample uniformly the corresponding intervals of the working sample conditional on the split sample interval. This would imply the same results.

Let us emphasize the importance of the random assignment. Even with a fixed S, the assigned values to the working sample are random. This means that with the same S, one gets different realizations of the synthetic variable $Z^\dagger$. In Sect. 4 we exploit this fact to identify the appropriate conditional expectations.

The unconditional probability of $Z^\dagger \in \mathcal{C}_b^{WS}$ for the shifting method, following Equation (9) and assuming $\Pr(Z \in \mathcal{S}_s) = 1/S$,

$$\Pr\left(Z^\dagger \in \mathcal{C}_b^{WS}\right) = \begin{cases} 0, & \text{if } s = 1 \text{ and } m = 1, \\ \frac{1}{S}\sum_{s=2}^{S} \frac{1}{s-1} \int_{\mathcal{C}_1^{(s)} | \mathcal{C}_b^{WS} \subset \mathcal{C}_1^{(s)}} f_Z(z)\mathrm{d}z, & \text{if } s \neq 1 \text{ and } m = 1, \\ \frac{1}{S^2}\sum_{s=1}^{S} \int_{\mathcal{C}_m^{(s)} | \mathcal{C}_b^{WS} \subset \mathcal{C}_m^{(s)}} f_Z(z)\mathrm{d}z, & \text{if } 1 < m < M, \\ \frac{1}{S}\sum_{s=1}^{S} \frac{1}{S-s+1} \int_{\mathcal{C}_M^{(s)} | \mathcal{C}_b^{WS} \subset \mathcal{C}_M^{(s)}} f_Z(z)\mathrm{d}z, & \text{if } m = M. \end{cases} \tag{15}$$

Next, we outline the assumptions under which $Z^\dagger \xrightarrow{d} Z$ as the number of split samples (S) and observations (N) go to infinity.

Assumption 1 *Let Z be a continuous random variable with probability density function $f_Z(z; a_l, a_u)$, where a_l, a_u, S, N and $\mathcal{C}_m^{(s)}$ are as defined above, then*

1.a) $\frac{S}{N} \to c$ *with* $c \in (0, 1)$ *as* $N \to \infty$.

1.b) $\int_a^b f_Z(z)dz > 0$ *for any* $(a, b) \subset [a_l, a_u]$.

Assumption 1.a) ensures that the number of observations is always higher than the number of split samples.[11] This is required to identify each point of $f_Z(z; a_l, a_u)$ through the mapping of split samples to the working set. Assumption 1.b) imposes a mild assumption on the underlying distribution: The support of the random variable is not disjoint within the working samples' interval, thus $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z)\mathrm{d}z > 0, \ \forall c_b^{WS}$.

Proposition 1 *Under Assumptions 1.a), 1.b) and* $\Pr(Z \in \mathcal{S}_s) = 1/S$,

$$\lim_{N,S\to\infty} F_{Z^\dagger}(a) = F_Z(a), \qquad \forall a \in (a_l, a_u). \tag{16}$$

Proof Recall the probability of Z falling into the working sample's interval $\mathcal{C}_b^{WS}$,

$$\Pr\left(Z \in \mathcal{C}_b^{WS}\right) = \sum_{s=1}^{S} \Pr(Z \in \mathcal{S}_s) \sum_{m=1}^{M} \Pr\left(Z \in \mathcal{C}_b^{WS} \mid Z \in \mathcal{C}_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f_Z(z)\mathrm{d}z. \tag{17}$$

For any $c_b^{WS}, \exists l \in [1, S], m \in [1, M]$ such that $c_b^{WS} = c_m^{(l)}$ and note that as $S \to \infty$, $N \to \infty$ in such a way that the number of participants in each s is greater than 0. Now consider $\Pr(Z^\dagger < c_b^{WS}) = \Pr(Z^\dagger < c_m^{(l)})$, given $\Pr(Z \in \mathcal{S}_s) = 1/S$ and using Equation (17) gives

[11] Note, that we can decrease c to be arbitrarily close to 0.

$$\Pr(Z^{\dagger} < c_m^{(l)}) = \frac{1}{S}\sum_{s=1}^{S} \Pr(Z < c_m^{(l)} | Z < c_m^{(s)}) \Pr(Z < c_m^{(s)}). \tag{18}$$

The summation over the different classes in Equation (17) is being replaced by the cumulative probabilities, where we use the fact that no value greater than $c_m^{(l)}$ is represented in the working sample c_b^{WS}. Under the shifting method, $c_m^{(s)} \leq c_m^{(l)}$ for $s < l$ and using the definition of conditional probability, gives

$$\begin{aligned}
\Pr(Z^{\dagger} < c_m^{(l)}) =& \frac{1}{S}\sum_{s=1}^{S} \Pr(Z < c_m^{(l)}, Z < c_m^{(s)}) \\
=& \frac{1}{S}\sum_{s=1}^{l} \Pr(Z < c_m^{(l)}, Z < c_m^{(s)}) + \frac{1}{S}\sum_{s=l+1}^{S} \Pr(Z < c_m^{(l)}, Z < c_m^{(s)}) \\
=& \frac{1}{S}\sum_{s=1}^{l} \Pr(Z < c_m^{(s)}) + \frac{1}{S}\sum_{s=l+1}^{S} \Pr(Z < c_m^{(l)}).
\end{aligned} \tag{19}$$

The last line follows from the fact that $\Pr(Z < a_1, Z < a_2) = \Pr(Z < a_1)$ if $a_1 < a_2$, and the construction of the shifting method allows us to always disentangle the two cases. Since l is fixed,

$$\begin{aligned}
\Pr(Z^{\dagger} < c_m^{(l)}) =& \frac{S-l-1}{S} \Pr(Z < c_m^{(l)}) + \frac{1}{S}\sum_{s=1}^{l} \Pr(Z < c_m^{(s)}), \\
\lim_{N,S\to\infty} \Pr(Z^{\dagger} < c_m^{(l)}) =& \Pr(Z < c_m^{(l)}),
\end{aligned} \tag{20}$$

where $\Pr(Z^{\dagger} < a) = F_{Z^{\dagger}}(a)$ and $\Pr(Z < a) = F_Z(a)$. □

3.2 Multivariate case

Let us generalize our results to the multivariate case. Let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_p, \ldots, \mathbf{z}_P)$ distributed as $\mathbf{Z} \sim \mathbf{f_Z}(\mathbf{z}; \mathbf{a}_l, \mathbf{a}_u)$, i.i.d. random variables with P dimensions, where $\mathbf{f_Z}(\cdot)$ denotes the (joint) pdf with $\mathbf{z} = (z_1, \ldots, z_p, \ldots, z_P)$,[12] and known support $\mathbb{A}^P = [a_{1,l}, a_{1,u}] \times \ldots \times [a_{P,l}, a_{P,u}] \subseteq \mathbb{R}^P$. We assume that $\mathbf{f_Z}(\cdot)$ is unknown and continuous. Instead of observing $\mathbf{Z}$, we observe $\mathbf{Z}^*$ through a (multivariate) discretization process similar to Equation (1). The difference is that instead of intervals, we use a multi-dimensional grid,[13] where each grid is defined as $\mathcal{C}_m =$

[12] Note here that $\mathbf{z}$ is not a row from $\mathbf{Z}$, but represents a vector that is used with the distribution $\mathbf{f_Z}(\cdot)$

[13] One-by-one discretization does not work in our case as it will not be able to recover the joint distribution of $\mathbf{f_Z}(\cdot)$. See more in online supplement, Section A1.3.

$[c_{1,m-1}, c_{1,m}) \times \ldots \times [c_{P,m-1}, c_{P,m})$ with $c_{m,p}$ being the m'th grid point for variable p. Note that $c_{p,0} = a_{p,l}$ and $c_{p,M} = a_{p,u}$, for $p = 1, \ldots, P$, thus the first boundary vector and last boundary vector contains the boundaries of the distribution's support. $\mathbf{v}_m \in \mathcal{C}_\mathbf{m}$ is the assigned vector for each grid, that can be a measure of centrality (e.g., focus point) or other point within the grid. In the case of $P = 2$, we have a 2D grid that defines the discretization scheme, and each $\mathcal{C}_\mathbf{m}$ stands for a rectangle with M^2 different partitions.

The shifting method in the multivariate case utilizes the initial discretization scheme with $\mathcal{C}_\mathbf{m}$ and shifts the boundary points of the grids with $h_p = \frac{a_{p,u} - a_{p,l}}{S(M-1)}$ for each $p = 1, \ldots, P$. This results in $(S(M-1))^P$ different grids in the working sample, which is much larger than the initial M^P. Changing the grids' boundary points allows for mapping the underlying distribution using the same inductive logic to prove convergence in distribution. More formally, first consider the following assumptions:

Assumption 2 *Let* $\mathbf{Z}$ *be a* $P \times 1$ *vector of continuous random variables with joint probability density function* $f_\mathbf{Z}(\mathbf{z}; \mathbb{A}^P)$*, where* $\mathbb{A}^P$*,* S*,* N *and* $\mathcal{C}_\mathbf{m}^{(s)}$ *are as defined above, then*

2.a) $\frac{S}{N} \to c$ *with* $c \in (0, 1)$ *as* $N \to \infty$*;*
2.b) $\int_{\mathbb{B}^P} f_\mathbf{Z}(\mathbf{z}) d\mathbf{z} > 0$ *for any* $\mathbb{B}^P \subseteq \mathbb{A}^P$.

Proposition 2 *Under Assumptions 2.a), 2.b) and* $\Pr(\mathbf{Z} \in \mathcal{S}_s) = 1/S$,

$$\lim_{N,S \to \infty} F_{\mathbf{Z}^\dagger}(\mathbf{a}) = F_\mathbf{Z}(\mathbf{a}), \qquad \forall \mathbf{a} \in \mathbb{A}^P. \tag{21}$$

See the complete proof in the online supplement under Section A1.1.

Let us mention that the extension to the multivariate case is theoretically appealing; however, in practice, this approach faces the curse of dimensionality, as the number of required grid cells increases exponentially, limiting its application.

3.3 Some further remarks

Speed of convergence

We investigated the speed of convergence for $F_{Z^\dagger}(\cdot) \to F_Z(\cdot)$ as a function of the number of split samples (S) to get insight into the choice of S in finite samples for the univariate case. In general, the rate of convergence is $1/S$, except around the boundaries of the support. Boundaries of the support are defined as: $c_1^{(1)} + (c_2^{(S)} - c_1^{(1)})$ for the lower bound, and $c_M^{(1)} + (c_{M-1}^{(1)} - c_M^{(1)})$ for the upper bound. In these regions, the speed of convergence is $\frac{\log S}{S}$. This result means that even with small values of S, one can get a decent mapping of the underlying marginal distribution of Z. We show the detailed derivations for the speed of convergence in online supplement, under Section A1.2.

Unbounded support

Let us consider the case when Z has unbounded support: $a_l = -\infty$ or $a_u = \infty$ or both. If we set the first and last interval's boundary points to infinity, we face censoring. Here we propose a simple solution: truncate observations that fall into the first or last interval: $\mathcal{C}_1^{(s)}, \mathcal{C}_M^{(s)}, \ \forall s$. Although this method loses observations and important units, the flip side is that, for the truncated distribution, all previous results now apply. In the online supplement, under Section A8, we discuss how to extend our approach for censoring using the continuous mapping theorem, which allows for a complete and satisfactory handling of the matter.

Other split sampling methods

It is possible to work out other discretization schemes that can map the underlying marginal distribution of Z. In fact, in the online supplement, under Section A5, we introduce the *magnifying method*, which offers an alternative method to uncover $f_Z(z; a_l, a_u)$. The key to all possible split sampling methods is to show convergence in distribution.

4 Linear models with discretized variable(s)

In this section we discuss how to use the synthetic variable $Z^\dagger$ to identify and estimate β in a linear regression model. As linear models and the OLS have well-known properties, we directly address modeling with multiple regressors.

First, we discuss the identification of β for each model, depending on where the discretization happens. This step helps us to pin down which theoretical expressions to quantify. To do so, we will first assume that $S \to \infty$, so the parameter of interest can be point identified. However, in practice, S is fixed, $Z^\dagger$ is constructed with some error,[14] and one only approximates the conditional expectation based on the (appropriate) class. We use these approximations of the conditional expectations to estimate the regression parameter(s) and show that the approximation error vanishes with $O(1/S^2)$.

Let $\mathbf{y} = (Y_1, \ldots, Y_N)'$ a $(N \times 1)$ vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_K)$ with dimensions $(N \times K)$ where $\mathbf{x}_k = (\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,N})'$. The $(K \times 1)$ parameter vector of interest is $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. We allow further conditioning variables that are not discretized denoted by $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_j, \ldots, \mathbf{w}_J)$ i.e., a $(N \times J)$ matrix with $\mathbf{w}_j = (\mathbf{w}_{j,1}, \ldots, \mathbf{x}_{j,N})'$ and parameter vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$, as this mimics standard practices better. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ belong to a subset of a compact finite-dimensional space $(\mathcal{B}, \mathcal{G})$.

Let the unknown data-generating process (DGP) be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \mathbf{u}, \qquad \mathbf{u} \sim iid\left(\mathbf{0}, \sigma_{\mathbf{u}}^2 \mathbf{I}\right), \tag{22}$$

[14] The difference between the true conditional expectation and the randomly assigned $Z^\dagger$ value.

where $\mathbf{X}$ and $\mathbf{W}$ are linearly separable. $\mathbf{u}$ is homoskedastic (for simplicity), and we use the usual OLS assumptions: $\mathbb{E}[\mathbf{u}|\mathbf{X}, \mathbf{W}] = 0$ and $\text{plim}_{N\to\infty}\left[\frac{1}{N}\mathbf{X}\mathbf{M}_{\mathbf{W}}\mathbf{X}\right]^{-1} = \mathbf{Q}$, positive definite matrix, with the usual residual maker $\mathbf{M}_{\mathbf{W}} = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$. In all three discretization cases, we follow a three-step procedure:

Step 1. Point identify $\boldsymbol{\beta}$, through defining conditional expectations using the known discretization intervals $\mathcal{C}_m^{(s)}$.
Step 2. Use the synthetic variable ($\mathbf{Z}^\dagger$) to estimate these conditional expectations and then discuss the OLS estimator and its asymptotic properties.
Step 3. Use the estimator from Step 2 and replace variable(s) from the linear model with their estimated conditional means. We derive an OLS estimator for $\boldsymbol{\beta}$ and show the asymptotic properties of the estimator in the modified regression, depending on where discretization happened.

4.1 Explanatory variable

First, let us discuss the case when we observe $\mathbf{X}^*$, instead of $\mathbf{X}$. We utilize the result of Proposition 2, hence $\mathbf{X}^\dagger \xrightarrow{d} \mathbf{X}$, where $\mathbf{X}^\dagger = \left(\mathbf{x}_1^\dagger, \ldots, \mathbf{x}_k^\dagger, \ldots, \mathbf{x}_K^\dagger\right)$ and $\mathbf{x}_k^\dagger = (\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,N})'$.

Step 1: Identification with X*

To identify $\boldsymbol{\beta}$, from the DGP defined by Equation (22), we need to use the conditional expectations given $\mathbf{X}^* \in \mathcal{C}_m^{(s)}$. Applying the conditional expectation operator on our model yields

$$\begin{aligned}\mathbb{E}\left[\mathbf{y}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}, \mathbf{W}\right] &= \mathbb{E}\left[\mathbf{X}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}\right]\boldsymbol{\beta} + \mathbb{E}\left[\mathbf{W}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}\right]\boldsymbol{\gamma} \\ &\quad + \mathbb{E}\left[\mathbf{u}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}\right]. \end{aligned} \tag{23}$$

Note that the classes, $\mathcal{C}_{\mathbf{m}}^{(s)}$ are mutually exclusive for given s along m and k.

Let us define $\kappa(s, \mathbf{m})$ a function[15] that maps the conditional expectations of $\mathbf{X}$ for given grid $\mathcal{C}_{\mathbf{m}}^{(s)}$. Proposition 2 ensures equality between the conditional expectation of the observable $\mathbf{X}^\dagger$, and the conditional expectation of the unknown $\mathbf{X}$ given $\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}$, which is needed to identify $\boldsymbol{\beta}$,

$$\kappa(s, \mathbf{m}) := \mathbb{E}\left[\mathbf{X}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}\right] = \mathbb{E}\left[\mathbf{X}^\dagger|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}\right] + O(1/S^2). \tag{24}$$

Note that the conditional expectations are identified even with fixed S, but with a finite S, the estimator we discuss below is biased.

[15] Note that $\mathbf{m} = (m_1, \ldots, m_K)$ is a vector, defining the grid points in $\mathcal{C}_{\mathbf{m}}^{(s)}$ for each variable k.

Step 2: OLS estimator for $\mathbb{E}\left[\mathbf{X}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(s)}\right]$

Let $\boldsymbol{\kappa} = (\kappa(1,\mathbf{1}),\ldots,\kappa(S,\mathbf{1}),\ldots,\kappa(S,\mathbf{m}),\ldots,\kappa(S,\mathbf{M}))'$ be a vectorized version of $\kappa(s,\mathbf{m})$ with $(SM^K \times 1)$ dimension. We propose a joint estimation for $\boldsymbol{\kappa}$ using OLS,[16]

$$\hat{\boldsymbol{\kappa}} = \left(\mathbf{1}'_{\{\mathbf{X}^\dagger \in \mathcal{C}_{\mathbf{m}}^{(s)}\}}\mathbf{1}_{\{\mathbf{X}^\dagger \in \mathcal{C}_{\mathbf{m}}^{(s)}\}}\right)^{-1}\mathbf{1}'_{\{\mathbf{X}^\dagger \in \mathcal{C}_{\mathbf{m}}^{(s)}\}}\mathbf{X}^\dagger, \tag{25}$$

where $\mathbf{1}_{\{\cdot\}}$ stands for the indicator function and results in a matrix with dimensions $(N \times SM^K)$ with the same ordering as $\boldsymbol{\kappa}$. Let us note, $\hat{\boldsymbol{\kappa}}$ does not require the actual values of $\mathbf{X}$, only $\mathbf{X}^\dagger$ and $\mathbf{X}^\dagger \in \mathcal{C}_{\mathbf{m}}^{(s)}$. Under the standard OLS assumptions, with convergence in *both* $N, S \to \infty$,

$\sqrt{N}\left(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}\right) \overset{a}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\kappa}})$,

where $\boldsymbol{\Omega}_{\boldsymbol{\kappa}}$ is the variance–covariance matrix. See the derivations for the proposed estimator in the online supplement, under Section A2.1. We add further derivations under the condition that S is fixed and show: i) the exact upper boundary, and ii) we prove that $\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa} = O(1/S^2)$. Furthermore, if one is concerned about the remaining bias, we provide a bias-reduction for $\hat{\boldsymbol{\kappa}}$ with an $O(1/S^4)$ rate. See these results in the online supplement, under Section A2.2. and A2.3.

Step 3: OLS estimator for $\boldsymbol{\beta}$

To get a consistent estimator for $\boldsymbol{\beta}$, let us follow Equation (23), and define $\ddot{\mathbf{X}}$, which replaces $\mathbf{X}^*$ with the corresponding conditional expectations via $\hat{\boldsymbol{\kappa}}$. Furthermore, let $\ddot{\mathbf{y}} = \hat{\mathbb{E}}\left[\mathbf{y}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}\right]$ and $\ddot{\mathbf{W}} = \hat{\mathbb{E}}\left[\mathbf{W}|\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}\right]$ the respective conditional averages.

$$\ddot{\mathbf{y}} = \ddot{\mathbf{X}}\boldsymbol{\beta} + \ddot{\mathbf{W}}\boldsymbol{\gamma} + \mathbf{e}, \tag{26}$$

while $\mathbf{e} = (e_1,\ldots,e_N)$. Now the OLS estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{y}}, \tag{27}$$

where $\mathbf{M}_{\ddot{\mathbf{W}}}$ is the usual residual maker using $\ddot{\mathbf{W}}$. Note that $\mathbb{E}[e_i] = 0$ for all i since $\ddot{\mathbf{X}}$ is a consistent estimate. Moreover, $\mathbb{E}[e_i e_j] = 0$ for $i \neq j$ due to $\mathcal{C}_{\mathbf{m}}^{(s)}$ being mutually exclusive in m and k. As long as the discretization is mean independent from the error term: $\mathbb{E}\left[\mathbf{u}|\mathbf{X},\mathbf{W},\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}\right] = \mathbb{E}\left[\mathbf{u}|\mathbf{X},\mathbf{W}\right] = 0$, the OLS is a consistent estimator.

[16] We can estimate all conditional expectations jointly via OLS. Otherwise simple conditional means for a given interval $\mathcal{C}_{\mathbf{m}}^{(s)}$ would yield the same result. Joint estimation, however, allows to establish asymptotic properties more easily.

Proof of consistency

$$\begin{aligned}
\operatorname*{plim}_{N,S\to\infty} \hat{\boldsymbol{\beta}} &= \operatorname*{plim}_{N,S\to\infty} \left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{y}} \\
&= \operatorname*{plim}_{N,S\to\infty} \left[\left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\left(\ddot{\mathbf{X}}\boldsymbol{\beta}+\ddot{\mathbf{W}}\boldsymbol{\gamma}+\mathbf{e}\right)\right] \\
&= \boldsymbol{\beta}+\mathbf{0}+\operatorname*{plim}_{\mathbf{N},\mathbf{S}\to\infty}\left[\left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}\mathbf{M}_{\ddot{\mathbf{W}}}\mathbf{e}\right] \\
&= \boldsymbol{\beta}+\operatorname*{plim}_{N,S\to\infty}\left[\left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\mathbb{E}\left[\mathbf{u}|\mathbf{X},\mathbf{W},\mathbf{X}^*\in\mathcal{C}_{\mathbf{m}}\right]\right] \\
&= \boldsymbol{\beta}.
\end{aligned}$$

□

Note that if S is fixed, then

$$\operatorname*{plim}_{N\to\infty}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\operatorname*{plim}_{N\to\infty}\left[\left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\left\{\mathbb{E}\left[\mathbf{X}|\mathbf{X}^*\in\mathcal{C}_m^{(s)}\right]-\ddot{\mathbf{X}}\right\}\right],$$

where $\mathbb{E}\left[\mathbf{X}|\mathbf{X}^*\in\mathcal{C}_m^{(s)}\right]\neq\ddot{\mathbf{X}}$, hence it is biased with finite S; therefore, our estimator is consistent in a nonstandard way and requires $S\to\infty$. However, the bias vanishes asymptotically as $S\to\infty$ at the same rate (see online supplement, Section A2.4). The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ can be derived similarly under the assumption $N, S\to\infty$. As $\mathbf{u}\sim iid\left(\mathbf{0},\sigma_{\mathbf{u}}^2\mathbf{I}\right)$ and as $\ddot{\mathbf{X}}$ and $\ddot{\mathbf{W}}$ are consistent estimates of the conditional means, $\mathbb{E}\left[\mathbf{ee}'|\ddot{\mathbf{X}},\ddot{\mathbf{W}}\right]=\sigma_e^2\mathbf{I}$. Under assumptions $N, S\to\infty$ and $\mathbb{E}\left[\mathbf{u}|\mathbf{X},\mathbf{W},\mathbf{X}^*\in\mathcal{C}_{\mathbf{m}}\right]=\mathbb{E}\left[\mathbf{u}|\mathbf{X},\mathbf{W}\right]$, we get $\sqrt{N}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\overset{\text{a}}{\sim}\mathcal{N}\left(\mathbf{0},\sigma_e^2\left(\ddot{\mathbf{X}}'\mathbf{M}_{\ddot{\mathbf{W}}}\ddot{\mathbf{X}}\right)^{-1}\right)$.

4.2 Outcome variable

The second case is when the outcome variable is discretized, thus $\mathbf{y}$ is not observed, only $\mathbf{y}^*$. To simplify our analysis let us neglect $\mathbf{W}$ along with the parameter vector $\boldsymbol{\gamma}$ as they can be seen as part of $\mathbf{X}$ and $\boldsymbol{\beta}$, respectively.

Step 1: Identification with y*

To identify $\boldsymbol{\beta}$ with discretized $\mathbf{y}^*$, let us partition the domain of $\mathbf{X}$'s into mutually exclusive partitions, denoted by $\mathcal{D}_{\mathbf{l}}$ and defined by the researcher, such that it is (mean) independent from the error term $\mathbf{u}$. We use the same logic as derived in Equation (23), but instead of conditioning on $\mathbf{X}^*\in\mathcal{C}_m$, we use $\mathbf{X}\in\mathcal{D}_{\mathbf{l}}$. Similar derivation can be done,

$$\begin{aligned}
\mathbb{E}\left[\mathbf{y}^*|\mathbf{X}\in\mathcal{D}_{\mathbf{l}}\right] &= \mathbb{E}\left[\mathbf{X}|\mathbf{X}\in\mathcal{D}_{\mathbf{l}}\right]\boldsymbol{\beta} \\
\mathbb{E}\left[\mathbb{E}[\mathbf{y}|\mathbf{y}^*\in\mathcal{C}_m,\mathbf{X}\in\mathcal{D}_{\mathbf{l}}]|\mathbf{X}\in D_l\right] &= \mathbb{E}\left[\mathbf{X}|\mathbf{X}\in\mathcal{D}_{\mathbf{l}}\right]\boldsymbol{\beta} \\
\sum_l\sum_m \mathbb{E}[\mathbf{y}|\mathbf{y}^*\in\mathcal{C}_m,\mathbf{X}\in\mathcal{D}_{\mathbf{l}}]\Pr[\mathbf{y}^*\in\mathcal{C}_m|\mathbf{X}\in\mathcal{D}_{\mathbf{l}}] &= \mathbb{E}\left[\mathbf{X}|\mathbf{X}\in\mathcal{D}_{\mathbf{l}}\right]\boldsymbol{\beta}.
\end{aligned} \tag{28}$$

$\Pr[\mathbf{y}^* \in \mathcal{C}_m|\mathbf{X} \in \mathcal{D}_\mathbf{l}], \mathbb{E}[\mathbf{X}|\mathbf{X} \in \mathcal{D}_\mathbf{l}]$ are known quantities and we aim to learn $\mathbb{E}[\mathbf{y}|\mathbf{y}^* \in \mathcal{C}_m, \mathbf{X} \in \mathcal{D}_\mathbf{l}]$. Note that $\mathbb{E}[\mathbf{u}|\mathbf{X} \in \mathcal{D}_\mathbf{l}] = 0$ by the definition of $\mathcal{D}_\mathbf{l}$.

Let us define $\pi(s, m, \mathbf{l})$ a function that maps the conditional expectations of $\mathbf{y}$ for a given interval $\mathbf{y} \in \mathcal{C}_m^{(s)}$ and partition $\mathbf{X} \in \mathcal{D}_\mathbf{l}$. Then

$$\pi(s, m, \mathbf{l}) := \mathbb{E}\left[\mathbf{y}|\mathbf{y}^* \in \mathcal{C}_m^{(s)}, \mathbf{X} \in \mathcal{D}_\mathbf{l}\right] = \mathbb{E}\left[\mathbf{y}^\dagger|\mathbf{y}^\dagger \in \mathcal{C}_m^{(s)}, \mathbf{X} \in \mathcal{D}_\mathbf{l}\right] + O(1/S^2) \quad (29)$$

where $\mathbf{y}^\dagger = (Y_1^\dagger, \ldots, Y_N^\dagger)$ and we used the results of the shifting method to show convergence.

Step 2: OLS estimator for $\mathbb{E}\left[\mathbf{y}|\mathbf{y}^* \in \mathcal{C}_m^{(s)}, \mathbf{X} \in \mathcal{D}_\mathbf{l}\right]$

Let $\boldsymbol{\pi} = (\pi(1, 1, \mathbf{1}), \ldots, \pi(S, 1, \mathbf{1}), \ldots, \pi(S, M, \mathbf{1}), \ldots, \pi(S, M, \mathbf{l}), \ldots, \pi(S, M, \mathbf{L}))'$ be the vectorized version of $\pi(s, m, \mathbf{l})$. Estimating $\boldsymbol{\pi}$, via OLS yields,

$$\hat{\boldsymbol{\pi}} = \left(\mathbf{1}'_{\{\mathbf{y}^\dagger \in \mathcal{C}_m^{(s)}, \mathbf{X} \in \mathcal{D}_\mathbf{l}\}} \mathbf{1}_{\{\mathbf{y}^\dagger \in \mathcal{C}_m^{(s)}, \mathbf{X} \in \mathcal{D}_\mathbf{l}\}}\right)^{-1} \mathbf{1}'_{\{\mathbf{y}^\dagger \in \mathcal{C}_m^{(s)}, \mathbf{X} \in \mathcal{D}_\mathbf{l}\}} \mathbf{y}^\dagger. \quad (30)$$

The vector $\hat{\boldsymbol{\pi}}$ has dimensions of $(SML^K \times 1)$, $\mathbf{1}_{\{\cdot\}}$ is the corresponding indicator matrix, and estimation of $\boldsymbol{\pi}$ does not require the actual values of $\mathbf{y}$, only $\mathbf{y}^\dagger$, and observing $\mathbf{y}^\dagger \in \mathcal{C}_m^{(s)}$ and $\mathbf{X} \in \mathcal{D}_\mathbf{l}$. Under the standard OLS assumptions, and both $N, S \to \infty$,

$\sqrt{N}\left(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\right) \overset{a}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\pi}})$,

where $\boldsymbol{\Omega}_{\boldsymbol{\pi}}$ is the corresponding variance–covariance matrix. Similarly to the case of discretized explanatory variable we show the exact upper boundary and got the same convergence results that is $O(1/S^2)$, when S is fixed. See these results in the online supplement, under Section A2.6.

Step 3: OLS estimator for $\boldsymbol{\beta}$

As shown in our identification strategy, the standard regression model can be rewritten in the form of Equation (28). Let $\tilde{\mathbf{y}}$ be a $(N \times 1)$ vector, where we replace the discretized elements of $\mathbf{y}^*$ with the consistent estimates of $\sum_l \sum_m \mathbb{E}[\mathbf{y}|\mathbf{y}^* \in \mathcal{C}_m, \mathbf{X} \in \mathcal{D}_\mathbf{l}] \Pr[\mathbf{y}^* \in \mathcal{C}_m|\mathbf{X} \in \mathcal{D}_\mathbf{l}]$, using $\hat{\boldsymbol{\pi}}$ and sample analogs for $\Pr[\mathbf{y}^* \in \mathcal{C}_m|\mathbf{X} \in \mathcal{D}_\mathbf{l}]$. Let us redefine $\tilde{\mathbf{X}} = \hat{\mathbb{E}}[\mathbf{X}|\mathbf{X} \in \mathcal{D}_\mathbf{l}]$, as the conditional sample means for $\mathbf{X}$ in partition $\mathcal{D}_\mathbf{l}$. With the replaced measures, we get

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\nu}, \quad (31)$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)'$. Note that $\mathbb{E}[\nu_i] = 0$ for all i since $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ are consistent estimates, if $N, S \to \infty$. Moreover, $\mathbb{E}[\nu_i \nu_j] = 0$ for $i \neq j$ due to that $\mathcal{D}_\mathbf{l}$ are mutually exclusive $\forall l$, and $\mathbb{E}[\nu_i|\tilde{\mathbf{X}}] = 0$ since the partitioning is (mean) independent from the

error term and does not affect the sampling error. Let $\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$, then under similar argument, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$.

The asymptotic distribution of $\hat{\boldsymbol{\beta}}$ can be derived in the same spirit.[17] As $\mathbf{u} \sim iid\left(\mathbf{0}, \sigma_{\mathbf{u}}^2\mathbf{I}\right)$ and $\tilde{\mathbf{X}}$ is a consistent estimate of the conditional means, $\mathbb{E}\left[\boldsymbol{\nu}\boldsymbol{\nu}'|\tilde{\mathbf{X}}\right] = \sigma_{\nu}^2\mathbf{I}$. The proposed estimator is asymptotically distributed as $\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \overset{a}{\sim} \mathcal{N}\left(\mathbf{0}, \sigma_{\nu}^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right)$.

4.3 Discretization on both sides

Our last case is when the discretization happens with both outcome $\mathbf{y}^*$, and with one or more explanatory variables $\mathbf{X}^*$. In this case, we do not need to partition the domain of $\mathbf{X}$, but can use the discretization grids $\mathcal{C}_{\mathbf{m}}^{(s)}$ for the explanatory variable. As $\mathbf{y}$ is also discretized, we need to partition $\mathbf{W}$ with $\mathcal{D}_{\mathbf{l}}$, similarly to the case discussed in Sect. 4.2. Identification is the same as with discretized outcome variable and the sample estimator for the conditional expectations follows the same logic as the OLS estimator for the $\boldsymbol{\beta}$ parameter. This case is discussed in details in the online supplement, under Section A3.

4.4 Discussion of the asymptotic approximation

Let us emphasize that both N and S must go to infinity at the rate given by Assumption 1. This is a key assumption, which enables us to estimate $\boldsymbol{\beta}$ consistently. In practice, S is fixed, and one faces bias in the estimator that depends *locally* on the second derivative of the underlying pdf. This means that if *within* the working sample bin there is excess curvature, and it is true for all bins, then bias is a problem. However, this is an unlikely scenario, and we argue that with narrow bin-widths in the working sample, a uniform approximation of the density function gives a reasonable result in most cases. We show that this approximation error converges at a rate of $1/S^2$ in the discussed linear regression cases. This means that the reduction in the approximation error compared to $S = 1$ is 96% with $S = 5$ and 99% with $S = 10$.[18] In Sect. 5 we show that this is indeed the case, and even a small number of S can lead to a significant reduction in the estimation bias relative to the true value. This makes our method appealing for empirical applications. However, let us caution that inference based on asymptotic normality may perform poorly in finite samples with fixed S, even when the sample is large, if the approximation bias is not reduced sufficiently.

Lastly, let us compare our results with the properties of the set identification approach, which uses only the boundary points from all other split samples. (See details in the online supplement, under Section A4.) First, we show that under the

[17] With fixed S, similar results hold as before, see online supplement, under Section A2.7.

[18] Furthermore, if one is worried about the size of the approximation error, then one can use further bias-reduction that achieves $1/S^4$ convergence rate, hence an even larger reduction in the approximation error.

same assumptions, the identified set regions collapse to the same point. We also derive the convergence rate, which is only $1/S$. This theoretical result aligns with our empirical analysis, in which the identified region shrinks much more slowly (see Table A23 in the online supplement).

4.5 Extensions

In case of panel data, the extension of this methodology is relatively straightforward. The most important difference is in the identification. If the interval of an individual does not change over the time periods covered, the individual effects in the panel and the parameter associated with the choice variable cannot be identified separately. The *Within* transformation would wipe out the interval variable as well. When the interval does change over time, but not much, then we are facing weak identification, i.e., in fact very little information is available for identification, so the parameter estimates are going to be highly unreliable.[19]

Using a panel setup is a good opportunity to raise awareness about a phenomenon that we call the *perception effect*. There is much evidence in the behavioral literature that the answers to a question may depend on the way the question is asked (see, e.g., Haisley et al., 2008). Note, that this is present regardless of whether split sampling has been performed or not. However, with split sampling with panel data there is a way to tackle this issue, by using split sample fixed effects. In the online supplement, Section A7, we extend our method toward fixed effect type of estimators.

Another possible extension is the use of nonlinear models ($\mathbf{y} = g(\mathbf{X}, \mathbf{W}; \boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{u}$) with one or more variables discretized. In this case the identification procedure is similar to Equations (23, 28), with the exception that $g(\mathbb{E}[\cdot; \boldsymbol{\beta}]) \neq \boldsymbol{\beta} g(\mathbb{E}[\cdot])$ in general. As we have shown with *split sampling* $\mathbf{Z}^{\dagger} \xrightarrow{p} \mathbf{Z}$, thus with discretized explanatory variable(s) by continuous mapping theorem $g(\mathbf{X}^{\dagger}; \boldsymbol{\beta}) \xrightarrow{p} g(\mathbf{X}; \boldsymbol{\beta})$. Similar argument applies when discretization happens on the left hand side ($\mathbf{y}^*$) or on both sides ($\mathbf{y}^*, \mathbf{X}^*$). In the online supplement, Section A8, we provide more details on the extension for nonlinear models.

Finally, let us mention that our discretization procedure may be used to protect data privacy. In the online supplement, Section A9, we discuss directions for extending the method to a privacy-preserving tool.

5 Simulation and empirical evidences

5.1 Monte Carlo evidence

For the simulation experiments, we consider a simple univariate linear regression model

$$Y_i = X_i' \beta + \epsilon_i \,, \tag{32}$$

[19] See more on this issue in the online supplement, Section A6.6.

where we discretize Y_i, or X_i or both. We set the parameter of interest $\beta = 0.5$. We report the Monte Carlo average bias and its standard deviation in parenthesis for 1, 000 repetitions and $N = 10,000$ observations. As the size of the bias depends on mapping the conditional expectations, we use multiple distributions for ϵ_i or X_i. To be more specific, we use *Normal* (standard normal distribution truncated at -1 and 3), *Logistic* (standard logistic distribution truncated at -1 and 3), *Log-Normal* (standard log-normal distribution truncated at 4 and subtracted 1 to adjust the support), *Uniform* (uniform distribution between -1 and 3), *Exponential* (exponential distribution with rate parameter 0.5, truncated at 4 and subtracted 1) and *Weibull* (Weibull distribution with shape parameter 1.5 and scale parameter 1, truncated at 4 and subtracted 1). All these distributions have known finite support between -1 and 3.

We use these distributions to generate ϵ_i when the outcome or both variables are discretized. In these cases $X_i \sim \mathcal{N}(0, 0.25, -1, 1)$ (truncated normal distribution between -1 and 1). When discretization happens with the explanatory variable X_i, we use the different distributions to generate X_i and use the same truncated normal distribution for ϵ_i. This setup allows us to control for the domain of Y_i in all cases. For the discretization of the variable(s) in all cases we use $M = 5$ and for split sampling $S = 10$.[20]

To estimate β, we use the *Shifting method*.[21] As alternatives, we use *Mid-point regression*, a commonly used naive model. This method uses mid-points for v_m with OLS for estimation. When the outcome variable is discretized, we use further popular estimation methods: *Set identification*, which provides estimates for the lower and upper bounds of the parameter set[22]; *Ordered probit and logit* models[23] and *Interval regression*[24] as alternative estimators.

We specify our models when applicable to estimate [25]as $Y_i = \alpha + X_i'\beta + \eta_i$, where $\epsilon_i = \alpha + \eta_i$ with $\mathbb{E}(\epsilon_i) = \alpha$, to adjust for those distributions, which do not have zero mean.

Table 1 shows our results. The shifting method provides smaller average bias than the mid-point regression by magnitudes between 5-50 almost everywhere. The two exceptions can be found in the case of discretized explanatory variable. When we use a uniform distribution, there is no bias for mid-point regression, as the conditional expectation is the mid-point of the interval. The second case is the exponential distribution, where the case is similar, as the curvature of the used distribution is rather flat.[26] The shifting method also outperforms the other competing methods, when discretization happens on the left hand side and provides significant reduction in the bias when both variables are discretized.

[20] We have experimented with several different setups; see more in the online supplement, Section A10.

[21] We used mid-values as observations for the working sample's values (v_b^{WS}).

[22] Estimation is based on Beresteanu and Molinari (2008), we use the package by Beresteanu et al. (2010).

[23] Note that the estimated maximum likelihood 'naive' parameters reported here are not designed to recover β and to be interpreted in the linear regression sense. Therefore, we call the difference from β *distortion* rather than bias.

[24] We assume a Gaussian conditional distribution to model the censored interval outcome.

[25] Ordered choice models' implementation in Stata removes the intercept parameter to identify β.

[26] This result is in line with the theoretical results shown in the online supplement, Sections A2.2 and A6.

Table 1 Monte Carlo average bias and standard deviations

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Discretized explanatory variable (X_i^*)						
Mid-point regression	−0.0252	−0.0101	−0.0174	0.0002	0.0005	−0.0422
	(0.0057)	(0.0046)	(0.0051)	(0.0040)	(0.0102)	(0.0073)
Shifting method ($S = 10$)	−0.0037	−0.0003	−0.0022	0.0002	0.0023	−0.0015
	(0.0060)	(0.0046)	(0.0050)	(0.0038)	(0.0094)	(0.0073)
Discretized outcome variable (Y_i^*)						
Set identification†	[−1.1, 1.15]	[−1.09, 1.15]	[−1.09, 1.16]	[−1.07, 1.17]	[−1.06, 1.19]	[−1.09, 1.15]
	[(0.02),(0.02)]	[(0.03),(0.03)]	[(0.02),(0.02)]	[(0.03),(0.03)]	[(0.03),(0.03)]	[(0.02),(0.02)]
Ordered probit*	0.1971	0.0688	0.2085	0.0158	0.0986	0.4461
	(0.0256)	(0.0253)	(0.0262)	(0.0234)	(0.0241)	(0.0295)
Ordered logit*	0.6509	0.3814	0.6862	0.2379	0.4338	1.2085
	(0.0464)	(0.0455)	(0.0499)	(0.0422)	(0.044)	(0.0546)
Interval regression	0.0268	0.0332	0.0371	0.0491	0.0663	0.0397
	(0.0198)	(0.0249)	(0.0221)	(0.0271)	(0.0249)	(0.0166)
Mid-point regression	0.0253	0.0322	0.0362	0.0490	0.2077	0.0314
	(0.0195)	(0.0236)	(0.0216)	(0.0273)	(0.0128)	(0.0157)
Shifting method	−0.0010	−0.0017	−0.0010	−0.0014	−0.0017	−0.0003
	(0.0211)	(0.0239)	(0.0215)	(0.0271)	(0.0125)	(0.0147)

Table 1 continued

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Both variables are discretized (Y_i^*, X_i^*)						
Mid-point regression	−0.0853	−0.0788	−0.0752	−0.0635	0.0797	−0.0759
	(0.0178)	(0.0213)	(0.0190)	(0.0243)	(0.0116)	(0.0137)
Shifting ($S = 10$)	−0.0027	0.0156	0.0104	0.0156	0.0006	0.0108
	(0.0235)	(0.0269)	(0.0243)	(0.0294)	(0.0132)	(0.0156)

Row names refers to X_i for 'discretized explanatory variable' and to ε_i when the outcome or both variables are discretized. Average bias and the standard deviations of the estimates in parenthesis are reported. $\beta = 0.5$. In the case of Y_i^*, we have used $L = 50$ with equal distances to partition X_i. When both variables are discretized, we used $M = M_Y = M_X = 5$ and $S = S_Y = S_X = 10$

† Set identification gives the lower and upper boundaries for the valid parameter set. We report these bounds subtracted from the true parameter; therefore, it should give a (close) interval around zero

* Distortion from the true β is reported. Ordered probit and logit models' maximum likelihood parameters do not aim to recover the true β parameter; therefore, we do not to call them biased

We have run several Monte Carlo experiments and the results are similar: the shifting method outperforms all alternatives. An important result is consistency, as we increase N along with S, we get smaller biases. This consistency does not hold for the competing methods. Also, when distribution has large mass around its boundary (e.g., log-normal or Weibull) our method performs similarly well. For a detailed discussion see the online supplement, Section A10.

Practical considerations

Based on the Monte Carlo experiments, we provide brief guidance on selecting S given N and the assumed underlying distribution.

1. As a rule of thumb, $S = 10$ eliminates most of the bias. $S = 5$ also performs well, especially when discretization is applied to the outcome.
2. If there are few observations ($N < 1,000$), then $S/N < 0.1$ should be used;
3. The curvature of the unknown pdf governs the bias; therefore, if there is a considerable amount of curvature in the pdf (e.g., multi-modal distribution), then $S = 10$ or even higher values may be used (but not exceeding $S/N < 0.1$). If feasible, an incremental increase in M reduces bias, so including one additional interval can in fact help a lot.

5.2 The Australian gender wage gap

We illustrate our approach with a short study of the Australian gender wage gap from 2017. We obtained wage and socio-economic data from the Australian Tax Office (ATO). The individuals' sample files record a 2% sample of the whole population,[27] including the actual wage values. This enabled us to estimate β when no discretization happens. In practice, researchers use interval censored data coming from surveys or privatized data. To mitigate this fact, we employ three different equally distanced discretization methods with $M = 3, 5$, and 10.

Table 2 shows regression estimates of the gender dummy (1 stands for female) on log yearly wage, while controlling for age, age square, marital status, occupation, and regional variables. The *'Directly observed/True'* row shows the estimated parameter when the actual values are known and used to estimate the parameter of interest. In the following rows we show the estimates when the discretization process is done in equal-sized intervals with $M = 3, 5, 10$. In all cases, the yearly wage was discretized, and then we took the log of the discretized mid-values as this represents the usual practice.[28] Mid-point regression does not use split samples, Set identification provides lower and upper bounds for the identified set based on Beresteanu and Molinari (2008), while with the shifting method we use $S = 10$ and report the asymptotic approximation for the point estimate. The point estimates are reported along with the standard errors for $\hat{\beta}$ in parenthesis.

[27] For details, see ATO's website: https://www.ato.gov.au.

[28] This mitigates a survey design where first the researcher ask about wage intervals in levels, then as data are collected, it can decided to take log or not for percentage interpretation.

Table 2 Different estimates for gender wage gap based on different discretizations of yearly wages

Directly observed/True	−0.2261 (0.0023)		
	Mid-point	Set-identification†	Shifting ($S = 10$)
$M = 3$	−0.2747	[−8.0389, 5.3057]	−0.2214
	(0.0033)	[(0.0248), (0.0263)]	(0.0030)
$M = 5$	−0.2635	[−2.9793, 1.3292]	−0.2348
	(0.0031)	[(0.0206), (0.0149)]	(0.0027)
$M = 10$	−0.2364	[−0.6052, 0.1161]	−0.2280
	(0.0025)	[(0.0031), (0.0030)]	(0.0024)
N	125,995		

† Set identification gives the lower and upper boundaries for the valid parameter set. Estimation is based on Beresteanu and Molinari (2008) and standard errors are bootstrapped SEs

The results show clearly that the shifting method provides closer estimates than mid-point regression everywhere, whereas the latter statistically produces different results from the 'Directly observed/True' value. Also note that the set-identified regions are quite large and include zero everywhere. Although it does not hinges on asymptotic approximation, the resulting interval is too wide to be useful in this particular empirical application. Also note that the resulting bias (distance from directly observed parameter, shown in Table A.21) is a magnitude smaller for the shifting method than for Mid-point analysis. In addition, several different setups were investigated, including robustness to the choice of the number of split samples along with M. For more discussion on the empirical example and for other model setups, see online supplement, Section A.11., in particular Table A.21, A.22 and A.23.

6 Conclusion

This paper deals with linear models using sensitive variables. The use of discretization process is proposed to increase response rates in surveys. This results in discretized (also called interval censored) variables, which makes econometric modeling difficult as, in general, the conditional expectation cannot be point identified.

We propose the *split sampling* method that introduces multiple discretization schemes; thus, instead of using one set of intervals, the sensitive variable is discretized through multiple versions. These discretized realizations are combined in a way that, under some mild conditions, converge in distribution to the original (unknown) variable. We introduce the shifting method, an easy to implement algorithm for split sampling that satisfies convergence in distribution to the unobserved variable. With the help of this we can identify parameters in linear models. The necessary point identification assumptions require that both the number of observations and the number of discretization schemes go to infinity, and depend on where the discretization happens: i) the sensitive variable is an explanatory variable; ii) it is the outcome; or iii) discretization happens on both sides. Each case is analyzed and the appropriate OLS

estimators in a multivariate regression setup are derived. The asymptotic properties of these estimators are derived and extensions to panel data, nonlinear models and for data privacy are also discussed. Some Monte Carlo evidence is provided to show that our methods have superior finite-sample properties compared to the 'usual' ones. Finally, this method is applied to estimate the Australian gender wage gap. Small bias in the estimates shows that even when the number of intervals is small, this new approach is quite useful.

Acknowledgements Contribution by Balázs Kertész to earlier versions of this paper is kindly acknowledged. We are grateful for all the useful comments from the two referees and on different versions of this paper by William Greene, Tom Wansbeek, Botond Köszegi, Robert Lieli, Gábor Békés, Miklós Koren, Michael Knaus, and the participants at the Central European University's Seminar, the IAAE, and EEA conferences.

Author Contributions All authors contributed equally.

Funding Open access funding provided by Corvinus University of Budapest. Special thanks to the International Association for Applied Econometrics for the financial support of the 2019 conference presentation.

Data Availability Simulation data are available with the codes, see below. ATO's dataset is confidential; therefore, we are not allowed to publish. codes are available at: https://github.com/regulyagoston/Split-sampling

Materials Availability Not applicable.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

Abrevaya, J., Muris, C.: Interval censored regression with fixed effects. J. Appl. Economet. **35**, 198–216 (2020). https://doi.org/10.1002/jae.2737

Andrews, D.W.K., Soares, G.: Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica **78**, 119–157 (2010). https://doi.org/10.3982/ECTA7502

Beresteanu, A., Molchanov, I., Molinari, F.: Sharp identification regions in models with convex moment predictions. Econometrica **79**, 1785–1821 (2011). https://doi.org/10.3982/ECTA8680

Beresteanu, A., Molinari, F.: Asymptotic properties for a class of partially identified models. Econometrica **76**, 763–814 (2008). https://doi.org/10.1111/j.1468-0262.2008.00859.x
Beresteanu, A., Molinari, F., Darcy, M.,. Asymptotics for partially identified models in stata. (2010)https://molinari.economics.cornell.edu/programs.html. Accessed: 01.08.2025
Bontemps, C., Magnac, T., Maurin, E.: Set identified linear models. Econometrica **80**, 1129–1155 (2012). https://doi.org/10.3982/ECTA7637
Chan, F., Mátyás, L., Reguly, Á.: Online supplement: Modelling with Sensitive Variables.(2026) https://regulyagoston.github.io/papers/MSV_online_supplement.pdf. Accessed: 2026.03.05
Chernozhukov, V., Hong, H., Tamer, E.: Estimation and confidence regions for parameter sets in econometric models. Econometrica **75**, 1243–1284 (2007). https://doi.org/10.1111/j.1468-0262.2007.00794.x
Crawford, F.W., Weiss, R.E., Suchard, M.A.: Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth-death processes. Ann. Appl. Stat. **9**, 572–596 (2015). https://doi.org/10.1214/15-AOAS809
Haisley, E., Mostafa, R., Loewenstein, G.: Subjective relative income and lottery ticket purchases. J. Behav. Decis. Mak. **21**, 283–295 (2008). https://doi.org/10.1002/bdm.588
Heitjan, D.F., Rubin, D.B.: Inference from coarse data via multiple imputation with application to age heaping. J. Am. Stat. Assoc. **85**, 304–314 (1990). https://doi.org/10.1080/01621459.1990.10476202
Heitjan, D.F., Rubin, D.B.: Ignorability and Coarse Data. Ann. Stat. **19**, 2244–2253 (1991). https://doi.org/10.1214/aos/1176348396
Heitjan, D.F.: Ignorability, sufficiency and ancillarity. J Royal Stat. Soc. Ser. B (Stat. Methodol.) **59**, 375–381 (1997). https://doi.org/10.1111/1467-9868.00073
Hsiao, C.: Regression analysis with a categorized explanatory variable, in: Karlin, S., Amemiya, T., Goodman, A.L. (Eds.), Studies in Econometrics, Time Series, and Multivariate Statistics. Academic Press. chapter 5, pp. 93–129 (1983)
Imbens, G.W., Manski, C.F.: Confidence intervals for partially identified parameters. Econometrica **72**, 1845–1857 (2004). https://doi.org/10.1111/j.1468-0262.2004.00555.x
Jang, W., Loh, J.M.: Density estimation for grouped data with application to line transect sampling. Ann. Appl. Stat. **4**, 893–915 (2010). https://doi.org/10.1214/09-AOAS307
Jones, M.C.: Discretized and interpolated kernel density estimates. J. Am. Stat. Assoc. **84**, 733–741 (1989). https://doi.org/10.1080/01621459.1989.10478827
Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**, 255–260 (2015). https://doi.org/10.1126/science.aaa8415
Kaido, H., Molinari, F., Stoye, J.: Confidence intervals for projections of partially identified parameters. Econometrica **87**, 1397–1432 (2019). https://doi.org/10.3982/ECTA14075
Lewbel, A.: The identification zoo: meanings of identification in econometrics. J. Econom. Literat. **57**(4), 835–903 (2019)
Manski, C.F.: Partial identification of probability distributions. Sci. Bus. Media (2003). https://doi.org/10.1007/b97478
Manski, C.F., Tamer, E.: Inference on regressions with interval data on a regressor or outcome. Econometrica **70**, 519–546 (2002). https://doi.org/10.1111/1468-0262.00294
Molinari, F.: Chapter 5 - Microeconometrics with partial identification, in: Durlauf, S.N., Hansen, L.P., Heckman, J.J., Matzkin, R.L. (Eds.), Handbook of Econometrics, Volume 7A. Elsevier. volume 7 of *Handbook of Econometrics*, pp. 355–486.(2020) https://doi.org/10.1016/bs.hoe.2020.05.002
Oller, R., Gómez, G., Calle, M.L.: Interval censoring: identifiability and the constant-sum property. Biometrika **94**, 61–70 (2007). https://doi.org/10.1093/biomet/asm002
Pacini, D.: The two-sample linear regression model with interval-censored covariates. J. Appl. Economet. **34**, 66–81 (2019). https://doi.org/10.1002/jae.2654
Scott, D.W.: Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. Ann. Stat. **13**, 1024–1040 (1985). https://doi.org/10.1214/aos/1176349654
Schomburg, G., Behlau, H., Dielmann, R., Weeke, F., Husmann, H.: Sampling techniques in capillary gas chromatography. J. Chromatogr. A **142**, 87–102 (1977). https://doi.org/10.1016/S0021-9673(01)92028-X
Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 05, 557–570. World Scientific (2002)
Tamer, E.: Partial identification in econometrics. Ann. Rev. Econom. **2**, 167–195 (2010). https://doi.org/10.1146/annurev.economics.050708.143401

Wang, X., Chen, S.: Partial identification and estimation of semiparametric ordered response models with interval regressor data. Oxford Bull. Econ. Stat. **84**, 830–849 (2022). https://doi.org/10.1111/obes.12484

Ziegenfuss, J. Y., Easterday, C. A., Dinh, J. M., JaKa, M. M., Kottke, T. E., Canterbury, M.: Impact of Demographic Survey Questions on Response Rate and Measurement: A Randomized Experiment. Survey Practice 14: 1 (2021) https://doi.org/10.29115/SP-2021-0010