Modelling with Discretized Continuous Dependent Variable*

Felix Chan[†1], Ágoston Reguly[2], and László Mátyás[2]

[1]Curtin University
[2]Central European University

April 23, 2021

Abstract

The paper deals with econometric models where the dependent variable is continuous but cannot be observed directly. Instead, it is observed through intervals or discretized ordered choice windows. Manski and Tamer (2002) show that the parameters in the conditional expectation cannot be point-identified using these discretized observations. Here we introduce a new sampling design, the so called *split sampling*, which makes the point-identification of the parameters in regression models feasible. Split sampling yields point-identification through the way information is collected. The target sample set is split into multiple parts and data is collected in a differentiated way. We explore how split sampling affects statistical inference, and further Monte Carlo evidence is provided about it's effect on estimation. Finally, we propose a simple formulation to deal with an eventual perception effect.

JEL: C01, C13, C21, C25, C81, C83

Keywords: Discretized outcome variable, interval data, identification, split sampling, survey design, perception effect.

*The authors would like to thank William Greene for his helpful comments on the earlier draft of this paper.

†Corresponding author: F.Chan@curtin.edu.au

1 Introduction

Recently there has been increasing use of econometric models where the dependent variable is continuous but cannot be observed directly. Instead, it is observed through a discretization process. Paper or internet-based survey questions are common examples for such discretizations. These questions are usually asked in the following way e.g., *'Is your weekly personal income below* 100\$*, between* 100 *and* 400\$ *or above* 400\$, where specific intervals are given for each option. This discretization leads to interval data, where respondents typically need to pick one option from a pre-set list, creating discrete ordinal observations from an underlying continuous variable. These responses are qualitative values, but the choices are ordered, and this order is the only quantitative information available in the resulting variable.

In the empirical literature, income is a typical example for interval data. Income is usually discretized in surveys because it improves response rate dramatically when the question is asked in the form of income categories rather than as an exact amount. Another related reason for discretization is data confidentiality: e.g., statistical offices are not allowed to give exact information on personal income. (For more details on these practices, see e.g., Duncan et al. (2001)) Just to give a few examples, Bhat (1994) shows the effects of age, employment and other socio-economic variables on income where income is observed through three different categories; Micklewright and Schnepf (2010) compare individual and household income distributions controlling for age, gender and employment. The income is discretized and observed through single question surveys.

Modelling the conditional expectation for such discretized dependent variable with interpretable parameters is challenging as the regression parameters are generally not point-identified. Here we depart from the classical econometric approach to identification: we do not assume that the sample is given and the outcome variable is observed as interval values, but we propose a new sampling method which we called *split sampling*[1] that re-creates the continuous unobserved outcome variable rather than its discretized version. In other words, we investigate how to gather the data in order to point-identify and estimate the conditional expectation and to be

[1]The term *split sampling* in this paper is not related to the technique occasionally used in chromatography (Schomburg et al., 1977, Schomburg et al., 1981) or methods in machine learning, which splits the initial sample into folds.

able to rely on simple least squares regression techniques.

This paper deals with linear regression models where the dependent variable is observed through a discretization process, resulting in an interval variable. Here, let us note that, there is a substantial difference between *interval* and *ordinal* variables. Interval variables have known lower and upper boundaries for each choice interval and numeric intervals can be assigned for each observation. In principle, an interval variable has a (conditional) expectation, but it cannot be estimated directly using discretized data. The only information provided by this type of data is the lower and upper bounds of each of the categories and their frequencies. An ordinal variable embodies an even more severe information loss relative to the underlying continuous variable. The observed values have only a relational connection to each other (e.g., they are higher or lower), they are sorted into classes and numerical values cannot be assigned to the observations. Data from these qualitative variables cannot be used to estimate any conditional moments of the underlying random variables and can only be used to obtain the frequencies of each class. In this paper we only deal with interval variables.

Manski and Tamer (2002) show that the parameters of a regression model cannot be point-identified without any further restrictive assumption, when interval variable is used as the dependent variable. To circumvent this identification problem, there are two known solutions in the literature, both taking the discretization process as given. This paper adds a third possible solution by revisiting the discretization process itself.

The first and more popular solution among applied papers relies on the so-called ordered discrete choice models (Greene and Hensher, 2010), such as the ordered logit or probit. These models can handle the interval and ordinal variables as well and they aim for prediction or categorisation instead of (parameter) interpretation. The key for these models is that instead of modelling the conditional expectation, they focus on the conditional probabilities for each interval or class (e.g., the probability for an observation to fall into a certain class given a set of covariates). Ordinal choice models use a priori distributional assumptions to create the mapping between the outcome variable and the explanatory variables. This can be a strong assumption, especially in the absence of any probabilistic justification. These models tend to have their names based on the assumed distribution (e.g., ordered logit uses a (standardized) logistic, while ordered probit uses (standard) normal distribution). Under the

assumed distribution along with some mild conditions, the parameters are generally point-identifiable up to scaling. The main disadvantage is that it does not model the conditional mean directly, but rather, it provides the conditional probabilities. Therefore, the interpretation of the estimated parameters is markedly different than in a (linear) regression model. Generally, *'neither the sign nor the magnitude of the coefficient is informative [...], so the direct interpretation of the coefficients is fundamentally ambiguous.'* (Greene and Hensher, 2010, p. 142). To get meaningful interpretations, one can calculate the partial effects on the probabilities with the use of the assumed distribution. Greene and Hensher (2010) give a thorough overview of ordered choice models estimated via maximum likelihood. We should also mention the case when ordered choice models are augmented with the information on the observed interval boundaries. In practice, it is common to use this additional information and incorporate it into the model, but the rather strong distributional assumption remains for point-identification. The difficulties and drawbacks of this approach are nicely summarized by Greene and Hensher (2010, p. 133).

The second solution comes from the literature of partially identified parameters (see e.g., Manski and Tamer, 2002, Manski, 2003, Tamer, 2010). This approach focuses on interval data, and assigns numerical *intervals* for each observation. The main advantage of this method is that it does not require any distributional assumptions and still allows valid statistical inference on the conditional expectation. The magnitudes and the signs of the estimated parameter vector can be interpreted in the same way as the classical regression coefficients. The drawback is that the estimated parameters are not point-identified, but rather, it identifies a set in which the parameter vector may belong. In other words, it only obtains a lower and upper bound for each of the unidentified point estimates. Empirical applications are rare because the estimation method is complex and, in our experience, the estimated parameter intervals are too wide for them to be useful empirically.

This paper adopts the framework of Manski and Tamer (2002) and proposes a new solution for the point identification of the parameters of a linear regression model where the dependent variable is discretized into interval data. By revisiting the discretization process, a (survey) method is put forward which collects enough information for the point-identification without any additional (e.g., distributional) assumption. Intuitively, the parameters can be point-identified when the discretization process is designed in such a way that the lower and upper bounds for each

interval converge to each other. Then any linear regression models can be estimated in the usual way, e.g., by least squares (LS). The resulting point-estimates are then consistent and can be interpreted as in the classical regression framework.

The above discretization does not deviate substantially from the typical methods, but it allows to obtain additional information on the distribution of the dependent variable with the use of split sampling. In the context of surveys, it means to use multiple questionnaires. These questionnaires have the same set of questions but the choices (possible answers) of each questions are different. The term *split sample* is referred to the fact that the sample is *split* between these questionnaires. In general, the idea is to collect the data for the same set of questions with each question contains heterogenous sets of possible outcomes in different split samples.

Furthermore, the *perception effect* or survey heterogeneity – due to the use of multiple surveys – can also be estimated through fixed effects type estimator. Another useful property of the proposed approach is that it maintains the privacy considerations through the discretization process; therefore, the data provider can safely use this method as the individuals behind the answers cannot be identified.

The paper is organized as follows: Section 2 introduces the identification problem and justifies the proposed split sampling approach. Section 3 describes our split sampling approach with two easy to implement methods: *magnifying* and *shifting*. We also derive consistent estimates via least squares, and present some Monte Carlo simulations. Section 4 extends the simple framework in two ways. First it proposes a method to estimate and test perception effects. Second, it looks at non-linear models. Section 5 concludes.

2 Identification Problem

This section discusses the identification problems associated with the discretization of the data and justifies the split sampling approach by using the results from Manski and Tamer (2002).

Consider $y_i \sim f(a_l, a_u)$ an i.i.d. random variable, where $f(a_l, a_u)$ denotes the parent distribution function (pdf) with support $[a_l, a_u]$, where $a_l, a_u \in \mathbb{R}$, $a_l < a_u$ and $i = 1, \ldots, N$. We assume that $f(\cdot)$ is unknown and can be continuous, discrete or mixed. Instead of observing the outcomes of y_i, we observe y_i^* through a

discretization process:

$$y_i^* = \begin{cases} z_1 & \text{if } c_0 \leq y_i < c_1 \quad \text{or} \quad y_i \in C_1 = [c_0, c_1) \quad \text{1st choice} \\ z_2 & \text{if } c_1 \leq y_i < c_2 \quad \text{or} \quad y_i \in C_2 = [c_1, c_2) \\ \vdots & \vdots \\ z_m & \text{if } c_{m-1} \leq y_i < c_m \quad \text{or} \quad y_i \in C_m = [c_{m-1}, c_m) \\ \vdots & \vdots \\ z_M & \text{if } c_{M-1} \leq y_i < c_M \quad \text{or} \quad y_i \in C_M = [c_{M-1}, c_M) \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{last choice,} \end{cases} \tag{1}$$

where, $z_m \in C_m$, $m = 1, \ldots, M$ is the assigned value for each choice. It can be a measure of centrality (e.g., mid-point), or an arbitrarily assigned value within its interval. M denotes the number of choices, which is known. For simplicity, we refer to each choice or choice interval as a class.
We examine identification of $\mathbb{E}[y|x]$ when

$$\mathbb{E}[y|x] = h(x; \beta), \tag{2}$$

where $h(\cdot)$ is a known function, β is a parameter vector belonging to a subset of a compact finite-dimensional space ($\mathcal{B}$), and x denotes the vector of covariates.

Observing y_i^* instead of y_i leads to an identification problem. Following Manski and Tamer (2002) and Lewbel (2019), we show the conditions for the partial or set identification of β, which then point to the cases when point-identification is possible.
Let $\underline{y_i^*}$ and $\bar{y}_i^*$ denote the random variables that take the lower and upper bounds as the choice value in a given interval, respectively. In other words, $z_m = c_{m-1}, m = 1, \ldots, M$ for $\underline{y_i^*}$ and $z_m = c_m, m = 1, \ldots, M$ for $\bar{y}_i^*$. By the design of the discretization, the unobserved values lie between these lower and upper bounds, $\underline{y_i^*} \leq y_i \leq \bar{y}_i^*, \forall i$. Furthermore, as these refer to the same random variable, the *unknown* conditional probabilities are the same $\Pr\left[\underline{y}^* \in C_m|x\right] = \Pr[y \in C_m|x] = \Pr\left[\bar{y}^* \in C_m|x\right], \forall m$. Using the law of total expectation, it is easy to show that

$$\mathbb{E}[y|x] = \sum_m \left[\int_{c_{m-1}}^{c_m} y \Pr[y|x]\, dy\right] \Pr[y \in C_m|x], \tag{3}$$

where $\mathbb{E}(y \in C_m|x) = \int_{c_{m-1}}^{c_m} y \Pr[y|x]\, dy$, and by design $c_{m-1} \leq \mathbb{E}(y \in C_m|x) \leq c_m$, $\forall m$. Now, for the conditional expectations we get,[2]

$$\mathbb{E}\left[\underline{y}^*|x\right] \leq \mathbb{E}[y|x] \leq \mathbb{E}[\bar{y}^*|x]. \tag{4}$$

This bound reduces to a point in the limit when y_i is measured (or observed) precisely. However if y_i is only observed through an interval, we have a set of conditional expectations, which leads to the set identification for β. That is, under Equation 2, any $\mathrm{b} \in \mathcal{B}$ that satisfies $\mathbb{E}\left[\underline{y}^*|x\right] \leq h(x;\mathrm{b}) \leq \mathbb{E}[\bar{y}^*|x]$ is said to be observationally equivalent to β.[3]

Remark 1: β cannot be point-identified when $\mathbb{E}\left[\underline{y}^*|x\right] < \mathbb{E}[\bar{y}^*|x]$ or $\Pr[y|x]$ is unknown.

Remark 2: If the density of the conditional probability ($\Pr[y|x]$) is known, β is point-identified, which leads to the special cases of ordered choice models.

Following Manski and Tamer (2002), point identification of β can be achieved by using the equality condition in Equation (4) and by reconstructing the conditional probability of y_i. We maintain the assumption that y_i cannot be directly observed only through a limited number of choices/classes and we are not making any further (distributional) assumptions. The key to our approach is the use of split sampling, which uses different thresholds for the choices in each split sample. As we increase the number of split samples we achieve point-identification of $\mathbb{E}[y|x]$, and thus β. This split sampling method can be viewed as a non-parametric estimator on $\Pr[y|x]$. Split sampling method works in two ways:

1. The bounds are made narrower as we increase the number of split samples $(c_m - c_{m-1}) \to 0$. This leads to $\mathbb{E}\left[\underline{y_i^*}|x\right] \to \mathbb{E}[y_i|x]$ and $\mathbb{E}[\bar{y}_i^*|x] \to \mathbb{E}[y_i|x]$, without the need of $\underline{y_i^*} \to y_i$ and $\bar{y}_i^* \to y_i$.

2. It gives a better mapping of $\Pr[y|x]$ as we increase the number of different questions, therefore provides additional knowledge on $\mathbb{E}[y_i|x]$

Let us mention the implementations of the two other possible solutions. In set

[2]The same result can be found in Manski (1989) or Manski and Tamer (2002).

[3]See more on the terminology of observational equivalence in Chesher and Rosen (2017) or Lewbel (2019).

identification, Manski and Tamer (2002) propose a modified minimum-distance estimator, where the lower and upper bounds on the conditional expectation are also estimated along with the parameter set. Moment (in)equality models generalize Manski and Tamer (2002) for cases where there are multiple equations and/or inequalities (i.e., Chernozhukov et al. (2007) or Andrews and Soares (2010)). Beresteanu and Molinari (2008) shows 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 assumption. However, these methods do not produce point-estimates for β, only estimated lower and upper bounds.

On the other hand, ordered choice models point-identifies β up to a scale, by a particular distributional assumption on $F(c_m - \beta' x) = \Pr[y_i^* \leq m \mid x]$. Here $F(\cdot)$ is the assumed cumulative distribution function, usually Gaussian or logit and m is the m'th interval value (mid point or arbitrarily chosen ordinal value). β is identified through the assumption on $F(\cdot)$, which creates the mapping between the conditional probabilities and x. Point-identification of β up to a scale as any estimator for β is dependent on the distributional assumption, yielding different values for different distributions.

Ordered choice models aims to produce *conditional probabilities* rather than a proper interpretation for coefficients. Although there are many papers which tries to interpret the resulting parameters, which encourages us to emphasise the following properties of these models:

1. The interpretation of the parameters in an ordered choice model are different from the models where identification is based on the conditional expectation function.

2. This approach can be used to deal with interval and ordered variables as well, however these models (by default) handle the interval data as ordinal data.

3. If $F(\cdot)$ is wrongly assumed, parameter estimates will not provide consistent estimators of β.

4. In the case of the interval variable, it is possible to use the information on the intervals and fix the c_m parameters in the model. If one is interested in the

conditional expectation, it is possible to use an EM algorithm to compute the maximum likelihood estimator for β along with estimators for the expected values based on the assumed distribution. (Greene and Hensher, 2010, p. 133) We are going to refer to this method as 'interval regression' in our comparison study.

3 The Split Sampling Approach

The split sampling approach investigates how to gather the data (what is a good discretization process) in order to estimate the conditional expectation and to be able to rely on simple least squares regression techniques. The main idea is to create different questionnaires by using choices with different boundaries in each question, while fixing the number of choices (M). The term 'split sample' is referred to the fact that while the questions in each of these questionnaires are the same, the boundaries on their choices are different and therefore each questionnaire will have its own *split sample*. Due to human cognitive capacities, usually, a very limited number of choices is the only feasible way to construct such questionnaires.[4] The use of S split samples enables us to collect the answer of the same question in S different ways, which eliminates the discretization problem. We achieve this through changing the class boundaries (c_m) between each split sample.

The intuition behind the approach is that this leads to a better mapping of the unknown distribution of y and, in principle, to a complete mapping of the focus model. By merging the different split samples into one data set (let us call this the *working sample*), we get $b = 1, \ldots, B$ overall number of choice classes across the merged split samples, where B is much larger than M. In a given split sample, each respondent (i) is given one questionnaire. The set of respondents who fill in a questionnaire with the same class boundaries defines a split sample. Each split sample has $N^{(s)}$ number of observations ($s = 1, \ldots, S$ and $\sum_s N^{(s)} = N$). In this setup, the discretization of a split sample looks exactly as the problem introduced above in Equation (1); the only difference across split samples is that the class boundaries are different. Note that the number of observations across split samples

[4]Typically, the optimal number of choices for a survey is relatively small, $M = 3, 5, 7$ or at most $M = 10$. There is a large literature about the optimal number of choices (or 'scale points') in a survey, see e.g., Givon and Shapira (1984), Srinivasan and Basu (1989) or Alwin (1992).

can be the same or, more likely, different. Now a split sample is as follows,

$$y_i^{(s)} = \begin{cases} z_1^{(s)} & \text{if } y_i \in C_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ & \qquad \text{1st choice for split sample } s, \\ z_2^{(s)} & \text{if } y_i \in C_2^{(s)} = [c_1^{(s)}, c_2^{(s)}), \\ \vdots & \vdots \\ z_m^{(s)} & \text{if } y_i \in C_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ z_M^{(s)} & \text{if } y_i \in C_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last choice for split sample } s. \end{cases} \tag{5}$$

The observed values $z_m^{(s)}$ are set to a numeric value between $c_{m-1}^{(s)}$ and $c_m^{(s)}$, typically to the mid-point. In the second step, we merge all the split samples and create a 'working-sample' used to estimate the parameter(s) of interest. The working-sample is an artificial construction created in such a way that the working class boundaries (c_b^{WS}) are the union of the class boundaries of split samples.[5]

$$\bigcup_{b=0}^{B} c_b^{WS} = \bigcup_{s=1}^{S} \bigcup_{m=0}^{M} c_m^{(s)} \,. \tag{6}$$

With proper re-distribution of the observations to the working sample, we can re-construct the underlying unobserved continuous variable's distribution. To be more specific, we show two different ways to carry out split sampling.

3.1 The Magnifying Method

The magnifying method magnifies parts of the domain in each questionnaire by one equally sized choice class. The size of classes depends on the number of split samples (S) and number of choices (M). As the number of split samples increases, class sizes

[5]Here, we discuss the cases where the domain (a_l, a_u) for y_i is known and the working sample's class boundaries maps the domain of y_i, $c_0^{WS} = a_l, c_B^{WS} = a_u$. Our proof holds for cases where (a_l, a_u) are unknown and $c_0^{WS} \neq a_l$ and/or $c_B^{WS} \neq a_u$. In these cases, one might drop observations which are outside the domain of the survey(s) e.g., $a_l < c_0^{WS}$, or there is a known censoring in the survey, $a_l = -\infty$ and/or $a_u = \infty$. Note that in these cases the sample properties (e.g., speed of convergence) can be different.

decrease, which uncovers the unknown underlying distribution. Figure 1 shows the main idea of the magnifying method with the individual questionnaires for the case $M = 3$ and $S = 4$. The last line shows the working sample.

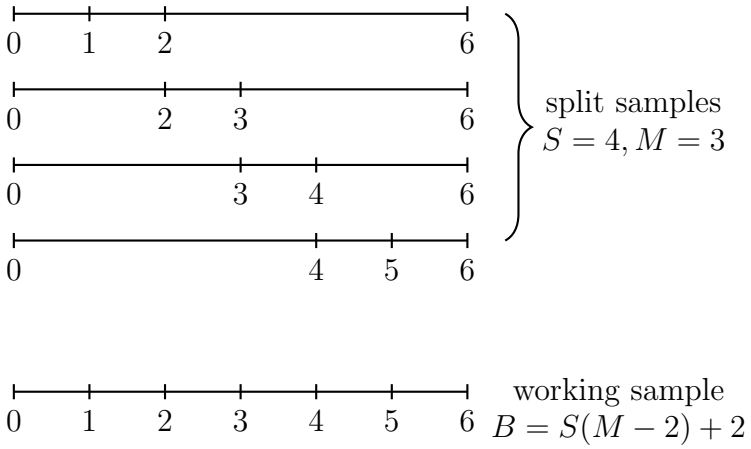

Figure 1: The magnifying method

The first and last split samples are slightly different from the split samples in-between. They have one extra class with the same class width, while split samples in-between have $M-2$ classes with the same class width. Observations which fall into these classes are called *directly transferable observations* (DTOs). The connection between the number of magnified classes in the working sample (B), and the number of split samples (S) and choices (M) is given by

$$B = S(M-2) + 2\,.$$

Given the fact that there are B classes in the working sample, we get the widths of these classes,

$$h = \frac{a_u - a_l}{S(M-2)+2}\,.$$

Fixing the upper and lower bounds on the domain[6] for the split samples ($a_l = c_0^{WS} = c_0^{(s)}$; $a_u = c_B^{WS} = c_M^{(s)}$, $\forall s$), we can reduce the class size $h \to 0$ as $S \to \infty$, which enables us to ensure convergence in distribution. This can also be seen through the

[6]In case of infinite support for the domain, one needs to set c_1^{WS} and/or c_{B-1}^{WS} into a reasonable value on the support. We discuss later the these truncated cases.

working sample's boundary points, which have the following simple form

$$c_b^{WS} = a_l + bh\,.$$

Algorithm 1 Magnifying method – creation of the split samples

1: For any given S and M, set

$$\begin{aligned} B =& S(M-2)+2 \\ h =& \frac{a_u - a_l}{B} \\ s =& 1. \end{aligned}$$

2: Set $c_0^{(s)} = a_l$ and $c_M^{(s)} = a_u$.
3: If $s = 1$, then set

$$c_1^{(s)} = c_0^{(s)} + h,$$

else set

$$c_1^{(s)} = c_{M-1}^{(s-1)}.$$

4: Set $c_m^{(s)} = c_{m-1}^{(s)} + h$ for $m = 2, \ldots, M-1$.
5: If $s < S$, then $s := s + 1$ and go to Step 2.

With the magnifying method we can separate two types of observations. The first is the already mentioned directly transferable observations. Formally, $y_i^{(s)} \in \zeta$, where ζ is the set of choice intervals of $\zeta = C_m^{(s)}$, $\forall$ pair of $(1 < s < S, 1 < m < M)$, and $(s = 1, m = 1), (s = S, m = M)$. Here, $\lim_{S\to\infty} ||C_m^{(s)}|| = 0$, which means that at the limit we observe responses without any discretization. Moreover, these observations have the same class width as the working sample's classes and each can be directly linked to a certain working sample class, by design, hence the name '*directly transferable observations*'. These observations are denoted by $y_{i,DTO}^{WS}$, with $i = 1, \ldots, N_{DTO}^{WS}$. Algorithm 2 describes how to construct in practice the working sample, using the directly transferable observations.

Algorithm 2 The magnifying method - creation of the 'DTO' working sample

1: Set $m = 1, s = 1$ and $y^{WS}_{i,DTO} = \emptyset$

2: If $C^{(s)}_m \in \zeta$, add observations from class $C^{(s)}_m$ to the working sample:

$$y^{WS}_{i,DTO} := \left\{ y^{WS}_{i,DTO}, \bigcup_{j=1}^{N} \left(y^{(s)}_j \in C^{(s)}_m \right) \right\},$$

3: If $s < S$, then $s := s + 1$ and go to Step 2.

4: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 2.

To achieve point-identification of β, the first step is to have a consistent mapping of the unconditional distribution of the unknown y variable by using $y^{WS}_{i,DTO}$. To show $\lim_{S\to\infty} \Pr\left(y^{WS}_{DTO}\right) = \Pr(y)$, we need the following assumptions,

Assumptions 1.

1. $\frac{S}{N} \to c$ with $c \in (0, 1)$ as $N \to \infty$.
2. $\Pr(y \in s) > 0$.
3. $\int_a^b f(y)dy > 0$ for any $(a, b) \subset [a_l, a_u]$.

Assumption 1.1 ensures that the number of respondents will always be higher than the number of split samples. Assumption 1.2 provides utilisation of all split samples, i.e. each split sample will have non-zero respondents. Assumption 1.3 imposes a mild assumption on the underlying distribution. That is, the support of the random variable is not disjoint, which implies $\int_{c^{WS}_{b-1}}^{c^{WS}_b} f(y)\mathrm{d}y > 0$.

Remarks: we can decrease c as close to 0 as we would like to. This means that there is an equal or higher number of observations than split samples. On the other hand, we exclude by assumption the case when c $\geq$ 1, which means that there is an equal or higher number of split samples than observations. In this case, we most certainly would not observe values for each working sample class.

These assumptions allow us to claim the following proposition,

Proposition 1. *Under Assumptions 1.1 - 3,*

$$\Pr\left(y^{WS}_{DTO} < c\right) = \Pr(y < c) \text{ for any } c \in [a_l, a_u]$$

.

See the proof of Proposition 1 in the online appendix[7]. The proposition establishes convergence in distribution which allows point-identification for the parameter of interest, discussed at Section 3.3.

Next, let us consider the second type of observations, which are all the other observations which fall into choice classes at the boundaries. We call them '*non–directly transferable observations*' (NDTOs) as for these observations $\lim_{S\to\infty} ||C_m^{(s)}|| = a_u - a_l$ for split sample and choice value pairs of $1 < s < S$, $m = \{1, M\}, s = 1$, $m = M$ or $s = S$, $m = 1$. This means that there is no systematic reduction in the measurement error for these responses. One way to proceed is to drop them completely so they would not appear in the working sample (thus, only using $y_{i,DTO}^{WS}$ for estimation purposes). However, in practice it seems that too many could fall into this category, resulting in a large efficiency loss during the estimation.
Another approach is to use DTOs to proxy the measurement error for the NDTOs. We can utilise the information from y_{DTO}^{WS} to calculate specific interval means for the underlying distribution and use these to replace the non–directly transferable observations. The simplest way to get the estimators for these conditional means is to regress the directly transferable observations on a vector of indicator variables referring to the NDTOs' intervals defined as follows,

$$y_{i,DTO}^{WS} = \boldsymbol{\pi}\, \mathbf{1}_{\{y_{i,DTO}^{WS} \in \zeta^{\complement}\}} + \eta_i$$

where, $\boldsymbol{\pi}$ is the replacement estimator for the non-directly transferable observations and

$$\mathbf{1}_{\{y_{i,DTO}^{WS} \in \zeta^{\complement}\}} = \begin{cases} 1\,, & \text{if } y_{i,DTO}^{WS} \in \zeta^{\complement} := \bigcup_{s,m} C_m^{(s)} \bigcap_b C_b^{WS} \\ 0\,, & \text{otherwise} \end{cases}$$

$\boldsymbol{\pi}$ can be used to replace the NDTOs, as they reflect the interval means: $\mathbb{E}(y|y \in \zeta^{\complement})$. The LS estimator for $\boldsymbol{\pi}$ has the asymptotic properties,

$$\sqrt{N_{DTO}^{WS}} \left(\hat{\boldsymbol{\pi}} - \mathbb{E}(y|y \in \zeta^{\complement}) \right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}\right),$$

[7] https://github.com/regulyagoston/Split-sampling/blob/master/Dependent/doc/Supplementary%20Materials_LHS.pdf

where the asymptotic variance of the LS estimator is

$$\mathbf{\Omega} = \mathrm{V}\left(\eta_i\right)\left(\mathbf{1}'_{\{y^{WS}_{i,DTO}\in\zeta^{\complement}\}}\mathbf{1}_{\{y^{WS}_{i,DTO}\in\zeta^{\complement}\}}\right)^{-1}.$$

See the derivations in the online appendix. As it turns out, this method is quite handy, while we can utilise to some extent the responses for NDTOs as well. The decision to replace the observations with the replacement estimator or not can be based on these asymptotic variances.

As a last step we need to consider the censoring case for the magnifying method. A straightforward solution is to remove those observations which have infinite class boundary. In the magnifying method, this means to remove observations in the class(es) C_1^{WS} if we have $a_l = -\infty$ and/or C_B^{WS} if $a_u = \infty$. This solution means we artificially truncate both $y \to y^{tr}$ and $x \to x^{tr}$. For the truncated distribution, we can use all the derivations presented above, and we end up with convergence in distribution. That is, $f(y^{WS}_{DTO} \in \zeta^{tr}) \xrightarrow{d} f(y^{tr})$.[8]

The magnifying method can be seen as the simplest theoretical design for split sampling, which shows how the method works, but its use is limited in practice. It is mostly applicable when the survey design deal with small number of S, while with large number of split samples, the creation of the questions are infeasible. At last, let us remark this method does not preserve data confidentiality. It uses the fact that some individuals are correctly observed and y is i.i.d., therefore the generalization of those observations is correct.

3.2 The Shifting Method

The shifting method is an alternative to the magnifying method. It takes the original class width as given, with fixed class widths, and shifts the boundaries of each choice with a given fixed value. Increasing the split sample size does not affect the boundary widths in-between the domain, only the size of the shift. As we shift the boundaries, we add an extra class[9] for each split sample at the boundary where, due to the shift,

[8] ζ^{tr} is the set of intervals, which do not contains C_1^{WS} and/or C_B^{WS} depending on the support. Furthermore, note that truncation implies that we cannot replace the observations from the split samples with infinite boundaries, and also that the replacement estimator does not converge to the conditional expectation due to the truncation.

[9] To be more specific, we in fact reveal a hidden class.

the class width has increased. Figure 2 shows the split samples in this approach with $S = 4$ and with $M = 4$ classes.

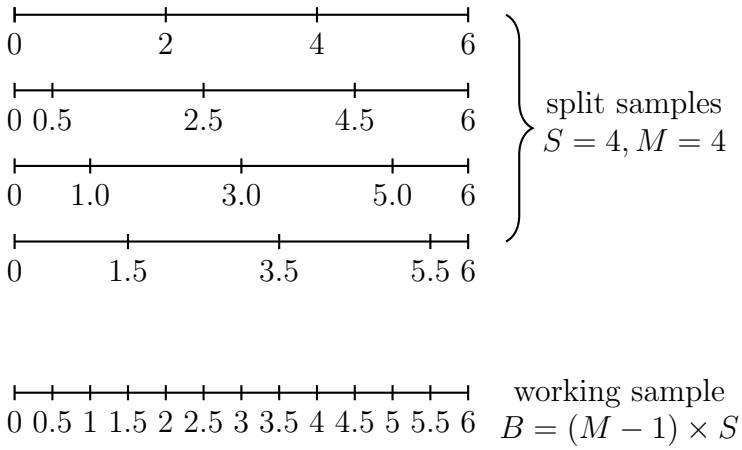

Figure 2: The shifting method

As Figure 2 shows, there is one split sample (the benchmark $s = 1$) where there is one class less, otherwise everywhere there is always M classes. The number of intervals in the working sample is

$$B = S \times (M - 1).$$

The boundary points for each split sample are

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

For the working sample, we get $c_b^{WS} = a_l + b\frac{a_u - a_l}{S(M-1)}$. Intuitively, we achieve complete mapping by reducing the shifting size to 0, thus we are be able to identify observations which lie in these small intervals, using the information available in all the other split samples.

Algorithm 3 describes how to create in practice split samples using the shifting method.

Algorithm 3 The shifting method - creation of split samples ($\psi^{(s)}(\cdot)$)

1: For any given S and M, set

$$\begin{aligned} B =& S(M-1) \\ h =& \frac{a_u - a_l}{B} \\ \Delta =& \frac{a_u - a_l}{M-1} \\ s =& 1. \end{aligned}$$

2: Set $c_0^{(s)} = a_l$ and $c_M^{(s)} = a_u$.

3: If $s = 1$, set

$$c_m^{(s)} = c_{m-1}^{(s)} + \Delta, \qquad m = 2, \ldots, M-1$$

else

$$c_m^{(s)} = c_m^{(s-1)} + h, \qquad m = 1, \ldots, M-1.$$

Note: $c_1^{(1)}$ does not exist.

4: If $s < S$, then $s := s + 1$ and go to Step 2.

Merging the split samples into the working sample is somewhat cumbersome, but works efficiently. The main idea is to uniformly assign each split sample's observations to the working sample's choice values, whose intervals are congruent with the split sample's class interval. Algorithm 4 describes how to create an artificial variable $y_i^\dagger$, which is the realisation of the random variable $y^\dagger$.

Algorithm 4 The shifting method – creation of artificial variable ($y_i^\dagger$)

1: Set $s := 1, m := 1, y_i^\dagger = \emptyset$.

2: Create the set of observations from the defined split sample class:

$$\mathcal{A}_m^{(s)} := \{y_i^{(s)} \in C_m^{(s)}\}, \forall i.$$

Let n_a be the number of observations in $\mathcal{A}_m^{(s)}$

3: Create set $\mathcal{Z}_m^{(s)}$, with possible working sample choice values, defined by $\mathcal{Z}_m^{(s)} = \bigcup \left(z_b \in [c_{m-1}^{(s)}, c_m^{(s)}) \right)$ Let $n_{\mathcal{Z}}$ be the number of choice values in $\mathcal{Z}$

4: Draw from $\mathcal{Z}_m^{(s)}$, with $n_{\mathcal{Z}}^{-1}$ uniform probabilities n_a times and assign each value as $\mathcal{Y}_j : \mathcal{A}_m^{(s)}(j) \to \mathcal{Z}_m^{(s)}$
Example: Let $C_3^{(2)} = [2.5, 4.5]$, $\mathcal{A}_m^{(s)} = \{3.5, 3.5, 3.5\}, n_a = 3$, $\mathcal{Z} = \{2.75, 3.25, 3.75, 4.25\}$, the uniform probabilities are $1/4$ for each choice value. Then we pick values with the defined probability from the set of $\mathcal{Z}_m^{(s)}$, 3 times with repetition, resulting in $\bigcup_{j=1}^{n_a} \mathcal{Y}_j = \{2.75, 3.25, 3.25\}$

5: Add these new values to $y_i^\dagger$,

$$y_i^\dagger := \left\{ y_i^\dagger, \bigcup_{j=1}^{n_a} \mathcal{Y}_j \right\}.$$

6: If $s < S$, then $s := s + 1$ and go to Step 2.

7: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 2.

The shifting method works similarly as $y^\dagger$ converges in distribution to the underlying continuous variable with the following mild assumptions:

Proposition 2. *Under Assumptions 1.1, 1.3 and* $\Pr(y \in s) = 1/S$,

$$\lim_{S \to \infty} \Pr(y^\dagger < c) = \Pr(y < c) \qquad \forall c \in (a_l, a_u)$$

See the proof in the online appendix. Furthermore with shifting method we can investigate the speed of convergence, as we increase the number of split samples

(S). The main result from the exercise is that on the boundaries of the support[10], the method converges slower, with $\frac{\log S}{S}$, while for the rest it converges with $1/S$. See the derivations in the online appendix.

A caveat is we cannot directly use $y_i^\dagger$ for estimation, while by design each individual observation only represents the conditional mean for the given split sample's class, and not the underlying variable's conditional expectation. However, while $\lim_{S\to\infty} F_S(y^\dagger) = F(y)$, we can use these values to calculate the specific conditional means – similarly to the NDTOs. Here I would not add the asymptotic distributions for these replacement estimators. They are similar to NDTOs, but slightly different as they are conditioned on D_l as well. Furthermore there is no decision here to replace or not as this is the only solution. Is it okey if I leave it like this?
Algorithm 5 describes how to create in practice a working sample with the shifting method.

Algorithm 5 The shifting method – creation of working sample

1: Set $s := 1, m := 1, y_i^{WS} = \emptyset$.
2: Calculate the sample conditional mean $\hat{\pi}_\tau$ using $y_i^\dagger$ conditioning on D_l class. D_l denotes a set containing L mutually exclusive partitions of the domain of x_i's.
3: Add the conditional mean $\hat{\pi}_\tau$ and the observed values to the working sample,

$$y_i^{WS} := \left\{ y_i^{WS}, \bigcup_{j=1}^{N} \hat{\pi}_\tau \mid (y_j \in D_l) \right\}$$

4: If $s < S$, then $s := s + 1$ and go to Step 2.
5: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 2.

We need to track the individual observations to be able to pair them with the right-hand side variables. Note that this pairing only applies to the conditional expected values not to the actual (un)observed value.
Some remarks: 1) The shifting method enables a more flexible survey design in practice, while the choice class widths are approximately the same. 2) The shifting method ensures data privacy considerations: creating artificial observations, and

[10]Which is given by the maximum distance from the support given by the split samples. For the lower bound: $c_1^{(1)} + (c_2^{(S)} - c_1^{(1)})$ and for the higher bound: $c_M^{(1)} + (c_{M-1}^{(1)} - c_M^{(1)})$.

calculating their conditional averages given the covariates will make the individuals' real value intractable.[11]

3.3 OLS Estimation

The proposed split sampling methods lead to two possible ways to obtain a consistent estimate of β via the least squares estimator. The first approach uses only the DTOs, while the second relies on all observations.

3.3.1 LS Estimation Based on DTOs

Let N^D denote the number of DTOs and let

$$\begin{aligned} \mathbf{y} =& \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim iid\left(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I}\right), & (7) \\ =& \hat{\mathbf{y}} + \boldsymbol{\varepsilon}, & (8) \end{aligned}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{N^D})'$. We make the following assumptions:

Assumptions 2.

1. $\mathbb{E}\left[|\mathbf{y}|\right]$ exists and $\mathbb{E}\left[\mathbf{X}\boldsymbol{\varepsilon}\right] = 0$.
2. There exists a bounded matrix $\mathbf{Q}$ such that $\dfrac{\mathbf{X}'\mathbf{X}}{N^D} - \mathbf{Q} = o_p(1)$.
3. $\forall \xi > 0\ \exists S$ such that $\Pr\left(|\varepsilon_i| > ||C_m||\right) = 1 - \xi$ given $\hat{y}_i \in C_m$.

The first two assumptions are common for LS estimator. Assumption 2.3. implies that ξ can be made arbitrarily small by choosing an appropriate S. This is due to the fact that $||C_m|| \to 0$ as $S \to \infty$.

Consider that the discretized version of $\mathbf{y}$ denotes $\mathbf{y}^*$ and write $\mathbf{y}^* = \mathbf{y} + \mathbf{u}$, where $\mathbf{u}$ represents the 'measurement errors' due to discretization. Consider the set

$$\mathbf{A} = \{i : |\varepsilon_i| > ||C_m|| \wedge \hat{y}_i \in C_m\}. \tag{9}$$

[11] One needs to pay special attention to the responses near the support $(C_1^{(s)}, C_M^{(s)})$ of the questionnaire as those can reveal some information about the respondent. In the limit $S \to \infty$ $C_2^{(1)}$ and $C_M^{(S)}$ identifies the respondents, however this is rather a theoretical case. This problem does not emerges if the support of y is infinite or these observations are dropped.

Set $\mathbf{A}$ allows us to distinguish between two sets of DTOs. Those belonging to $\mathbf{A}$ when the unobserved value, y_i, is in a different class than its corresponding $\hat{y}_i$. In this case, $|\varepsilon_i| > ||C_m||$ given $\hat{y}_i \in C_m$, which implies $x_i \perp u_i$. Those not belonging to $\mathbf{A}$ when both y_i and $\hat{y}_i$ belong to the same class, and therefore $Cov(x_i, u_i) \neq 0$. As we shall demonstrate below, these two sets of observations affect the properties of OLS differently.

Partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ such that $\mathbf{X}_1$ contains those observations whose indexes belong to $\mathbf{A}$ and $\mathbf{X}_2$ contains observations whose indexes do not belong to $\mathbf{A}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} =& (\mathbf{X}'\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y}^* \\
=& (\mathbf{X}'\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{u} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2'\mathbf{u} \\
=& \boldsymbol{\beta} + o_p(1) + \xi\left(\frac{\mathbf{X}'\mathbf{X}}{N^D}\right)^{-1}\frac{\mathbf{X}_2'\mathbf{u}}{\xi N^D} \qquad (10) \\
=& \boldsymbol{\beta} + o_p(1) + \xi O_p(1). \qquad (11)
\end{aligned}$$

The second last line follows from the fact that $X_1 \perp \mathbf{u}$. Under Assumption 2.3, as $||C_m|| \to 0$ for all m, $\xi \to 0$ and $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1)$.

The argument above relies on the assumption that the number of DTOs approaches infinity. In order to show that this can be the case, at least theoretically, we need to derive the relation between S and N^D. Let $\mathbf{B}$ denote the set of DTOs. From the proof of Proposition 1 we have,

$$\begin{aligned}
\Pr(y \in \mathbf{B}) =& \frac{1}{S}\left[\sum_{m=2}^{M}\Pr(y \in C_m^{(S)}) + \sum_{m=1}^{M-1}\Pr(y \in C_m^{(1)}) + \sum_{m=2}^{M-1}\sum_{s=2}^{S-1}\Pr(y \in C_m^{(s)})\right] \\
\leq& \frac{1}{S}.
\end{aligned}$$

Since y can only belong to one and only one class, all the events on the right-hand side are mutually exclusive and their sum must be less than or equal to 1. Note that $N^D = N\Pr(y \in \mathbf{B})$, and hence

$$N^D = O_p(\frac{S}{N}). \qquad (12)$$

3.3.2 $\hat{\beta}$ from the Conditional Expectation

The method above relies only on the DTOs. However, it is also possible to use all observations for the purpose of estimation. Consider the following standard regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim iid\left(\mathbf{0}, \sigma_{\varepsilon}^{2}\mathbf{I}\right), \tag{13}$$

where $\mathbf{y} = (y_1, \ldots, y_N)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ with $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN})'$, $i = 1, \ldots, k$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$. Let $\mathbf{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\}$ denote a set containing L mutually exclusive partitions of the domain of $\mathbf{X}$. Then

$$\mathbb{E}\left(\mathbf{y}|\mathbf{X} \in \mathbf{D}_l\right) = \mathbb{E}\left(\mathbf{X}|\mathbf{X} \in \mathbf{D}_l\right)\boldsymbol{\beta} \qquad l = 1, \ldots, L. \tag{14}$$

Let $\tilde{\mathbf{y}}_l$ and $\tilde{\mathbf{X}}_l$ denote consistent estimates of $\mathbb{E}\left(\mathbf{y}|\mathbf{X} \in \mathbf{D}_l\right)$ and $\mathbb{E}\left(\mathbf{X}|\mathbf{X} \in \mathbf{D}_l\right)$, respectively, $l = 1, \ldots, L$. Following from equation (14), we get

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u}, \tag{15}$$

where $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1 \ldots, \tilde{\mathbf{y}}_L)$, $\tilde{\mathbf{X}} = \left(\tilde{\mathbf{X}}_1', \ldots, \tilde{\mathbf{X}}_L'\right)'$ and $\mathbf{u} = (u_1, \ldots, u_L)'$. Note that $\mathbb{E}(u_l) = 0$ for all l since $\tilde{y}_l$ and $\tilde{\mathbf{X}}$ are consistent estimates. Moreover, $\mathbb{E}(u_l u_g) = 0$ for $l \neq g$ due to D_l are mutually exclusive $\forall l$ and $\mathbb{E}(u_l|\tilde{\mathbf{X}}_l) = 0$ since the partition does not affect the sampling error. Furthermore, assume $\mathbb{E}\left[\tilde{\mathbf{y}}\right]$ exists and $\mathbb{E}\left[\tilde{\mathbf{X}}\mathbf{u}\right] = 0$. Let $\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$, then under the usual argument of the classical OLS, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$.

The basic idea here is to obtain the averages of y_i conditional on different ranges of x_i. Such partitions preserve the correlation structure between y_i and x_i, as demonstrated above.

As we have shown that split sampling lead to $y_i^{\dagger} \xrightarrow{d} y_i$. Thus,

$$N_l^{-1} \sum_{\mathbf{x}_i \in \mathbf{D}_l} y_i^{\dagger} - \mathbb{E}\left(y_i|\mathbf{X} \in \mathbf{D}_l\right) = o_p(1). \tag{16}$$

Given this is a consistent estimator of $\mathbb{E}\left(y_i|\mathbf{X} \in \mathbf{D}_l\right)$, the above result holds and the consistency based on conditional expectation follows. It is worthwhile to point out that the above argument applies to any split sampling method that leads to convergence in distribution to the underlying dependent variable. Thus, its applicability

goes beyond the shifting and magnifying methods.

3.4 Monte Carlo Evidence

For the simulation experiments, we consider the following data generating process

$$y_i = x_i'\beta + \epsilon_i$$

and set $\beta = 0.5$. We focus on the case where the support of y_i is known, thus any error (bias) may come only from the discretization. We set the lower bound as $a_l = -2$, and the upper bound as $a_u = 4$. (We have experimented with different boundaries; the details are in the online appendix, Table 2.) The exogenous variable x_i is generated by a normal distribution, and to ensure the support of y_i, it is truncated at -1 and 1, with a variance of 0.25.[12] Our main concern is the disturbance term (ϵ_i), therefore we have visited several common types of distributions. To ensure the support of y_i is being met, we truncated/set ϵ_i such that it lies between $-1 \leq \epsilon_i \leq 3$.[13] We experiment with the following distributions:

- 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 (in order to adjust the mean).
- Uniform: Uniform distribution between -1 and 3.
- Exponential: Exponential distribution with rate parameter 0.5, truncated at 4 and subtracted 1.
- Weibull: Weibull distribution with shape parameter 1.5 and scale parameter 1, truncated at 4 and subtracted 1.

[12] This choice of variance ensures that even without truncation, 95% of the probability mass lies between -1 and 1.

[13] This creates an asymmetric distribution for ϵ_i in several cases, which favours distribution independent estimation methods rather than the maximum likelihood. In the online appendix, we show results with boundaries where ϵ_i was truncated in a symmetric way. The results and conclusions remain unchanged.

In the event that the distributions do not have a zero mean, we specified the conditional mean as $y_i = \alpha + x_i'\beta + \eta_i$, where $\epsilon_i = \alpha + \eta_i$ with $\mathbb{E}(\epsilon_i) = \alpha$.

For the discretization of y_i we use $M = 5$, $c_0 = a_l = -2$, $c_M = a_u = 4$ and equal distances for the thresholds between the boundaries.
To estimate β, we have used the following methods:

- *Set identification*: Estimates the lower and upper boundaries of the parameter set for β using y_i^* as interval data. Estimation is based on Beresteanu and Molinari (2008) and their published Stata package (Beresteanu et al., 2010)[14]. This method does not produces point-estimates for β, only lower and upper boundaries.

- *Ordered probit and logit*: Ordered choice models, where y_i^* values are ordinal data, and the model assumes a gaussian or logistic distribution (Greene and Hensher, 2010). 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* instead of bias. However, we find it important to report these values as (unfortunately) they are the most used and (mis-)interpreted estimates in applied work.

- *Interval regression*: A modification of the ordered choice model, where y_i^* values are interval data and the model assumes gaussian distribution in order to model the linear regression model. The maximum likelihood parameter estimates aim to recover β through the distributional assumption. For a detailed description, see Cameron and Trivedi (2010, p. 548-550) or Greene and Hensher (2010, p. 133).

- *Midpoint regression*: A simple linear regression using midpoints for y_i^* and OLS for estimation.

- *Magnifying*: The magnifying method with $S = 10$ split samples. We use only DTO observations.[15]

[14] https://molinari.economics.cornell.edu/programs.html

[15] We used mid-values as observations for the split samples' ($y_i^{(s)}$) and working sample's choice value. L is set to 50 equal distance partitions for x_i, where the conditioning was needed.

- *Shifting*: The shifting method with $S = 10$, all outcome observations are used and created as described in Algorithm 4.[15]

We have included an intercept wherever possible.[16] Finally, we used $N = 10,000$ observations and $1,000$ Monte Carlo repetitions. We report the Monte Carlo average bias or distortion from the true parameter along with the Monte Carlo standard deviation. Table 1 shows the results. The shifting method consistently provides

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.1, 1.15]$ (0.02),(0.02)	$[-1.09, 1.15]$ (0.03),(0.03)	$[-1.09, 1.16]$ (0.02),(0.02)	$[-1.07, 1.17]$ (0.03),(0.03)	$[-1.06, 1.19]$ (0.03),(0.03)	$[-1.09, 1.15]$ (0.02),(0.02)
Ordered probit*	0.1971 (0.0256)	0.0688 (0.0253)	0.2085 (0.0262)	0.0158 (0.0234)	0.0986 (0.0241)	0.4461 (0.0295)
Ordered logit*	0.6509 (0.0464)	0.3814 (0.0455)	0.6862 (0.0499)	0.2379 (0.0422)	0.4338 (0.044)	1.2085 (0.0546)
Interval regression	0.0268 (0.0198)	0.0332 (0.0249)	0.0371 (0.0221)	0.0491 (0.0271)	0.0663 (0.0249)	0.0397 (0.0166)
Midpoint regression	0.0253 (0.0195)	0.0322 (0.0236)	0.0362 (0.0216)	0.0490 (0.0273)	0.2077 (0.0128)	0.0314 (0.0157)
Magnifying ($S = 10$)	-0.0060 (0.0515)	-0.0205 (0.0674)	-0.0072 (0.0616)	-0.0332 (0.0781)	0.0213 (0.0333)	0.0066 (0.0417)
Shifting ($S = 10$)	-0.0034 (0.0205)	-0.0009 (0.023)	-0.0004 (0.0209)	-0.0023 (0.0278)	-0.0034 (0.0131)	-0.0008 (0.0153)

†: Set identification gives the lower and upper boundaries for the valid parameter set. We report these bounds subtracted with 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 it is not appropriate to call it bias.

Table 1: Monte Carlo average bias and standard deviation

the smallest average bias, while the magnifying method also outperforms the other procedures in general. Set identification gives such large intervals for the parameter set that it is unlikely to be useful in practice. Distortions of ordered probit and logit models are rather large. Interval regression and midpoint regressions perform poorly in the sense that both methods result in large biases. The Monte Carlo standard deviation is similar for all cases except for the magnifying method. This is due to the fact that the magnifying 'only DTO' method uses fewer observations, for the estimation, in our case $N/S \approx 1,000$ observations.

We have run several other Monte Carlo experiments to investigate the finite sample properties of our methods. With moderate sample size ($N = 1,000$), the results are similar: the shifting and magnifying methods outperform all alternatives. The

[16] Ordered choice models' implementation in Stata remove (restricts to zero) the intercept parameter to identify β.

magnifying method performs slightly more poorly in smaller samples, while the effective number of observations in this case is only around 100. Interestingly, in both the exponential and weibull setups, the magnifying method gives similar results as those from the interval and midpoint regressions. Naturally, if the distribution is well specified, methods with maximum likelihood estimation (ordered probit, logit or interval regression methods) produce even smaller biases. However, for the other (miss-specified) cases our split sampling methods work much better. For a smaller number of choices, $M = 3$, the biases are generally worse but the differences between the methods are similar. The shifting method still performs better, while the magnifying method still outperforms the alternatives in most cases. This suggests that our methods are robust to the underlying distributions.

Finally, we chose ϵ_i as a truncated standard normal and checked what happens if we increase the number of observations and the number of split samples. The simulation results – available in the online appendix – give evidence on the consistency of the estimator based on our split sampling approach. By contrast, for all the other alternative methods the same magnitude of bias remained as we increased the number of observations. This suggests that alternative methods provide not only biased but also inconsistent estimates for β in N. For a detailed discussion of these results see the online appendix.

4 Extensions

4.1 Perception Effect

There is some evidence in the behavioural literature that the answers to a question may depend on the way the question is asked (see, e.g., Diamond and Hausman (1994), Haisley et al. (2008) and Fox and Rottenstreich (2003)).[17] Let us call this the *perception effect*. This is present regardless whether split sampling has been performed or not. However, with split sampling there is a way to tackle this issue, much akin to the approach a similar problem has been dealt with in the panel data literature.

Let S be the total number of split samples and define two sets of discretization

[17] Comments by Botond Kőszegi on this section are highly appreciated.

of y_i namely,

$$y_i^* = \begin{cases} z_1 & \text{if} \quad c_0 < y_i < c_m \\ & \vdots \\ z_m & \text{if} \quad c_{m-1} < y_i < c_M \end{cases} \tag{17}$$

and

$$y^{**} = \begin{cases} z_1 & \text{if} \quad c_0 < y_i + B_s < c_1 \\ & \vdots \\ z_m & \text{if} \quad c_{m-1} < y_i + B_s < c_M, \end{cases} \tag{18}$$

where B_s denotes the perception effect for split sample s, $s = 1, \ldots, S$. Let $\tilde{y}_i^*$ and $\tilde{y}_i^{**}$ denote the observations in the working sample that derived from y_i^* and y_i^{**}, respectively. Following the construction of the working sample, it is straightforward to show that

$$\tilde{y}^{**} = \tilde{y}_i^* + B_s \tag{19}$$

given the corresponding y_i^* and y_i^{**} came from the split sample s. Thus, the regression

$$\tilde{y}_i^{**} = \beta x_i + u_i \tag{20}$$

is equivalent to

$$\tilde{y}_i^* + B_s = \beta x_i + u_i. \tag{21}$$

Writing the above in matrix form using the normal definition gives

$$\tilde{\mathbf{y}}^* + \mathbf{DB} = \mathbf{x}\beta + \mathbf{u}, \tag{22}$$

where $\mathbf{B} = (B_1, \ldots, B_S)'$ and $\mathbf{D}$ is a $N \times S$ zero-one matrix that extracts the appropriate elements from $\mathbf{B}$. So the estimation of β can be done in the spirit of a fixed effect estimator. Define the usual residual maker, $\mathbf{M_D} = \mathbf{I}_N - \mathbf{D}\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'$, then

$$\hat{\beta} = \left(\mathbf{x}'\mathbf{M_D}\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{M_D}\tilde{\mathbf{y}}^{**} \tag{23}$$

is a consistent estimator of β given the results presented in this paper and the similar argument for the consistency of standard fixed effect estimator in the panel data literature.

We also need to modify the estimator for $\mathbb{E}\left(y_i|\mathbf{X} \in \mathbf{D}_l\right)$ slightly in order for the

above to hold for the $\hat{\beta}$ based on conditional expectation. The main problem is to keep track of the perception effect. This means we need to keep track of which split sample each observation comes from when estimating the conditional averages. Specifically,

$$N_l^{-1} \sum_{\mathbf{x}_i \in D_l, \mathbf{x}_i \in s} \tilde{y}_i^{**} - \mathbb{E}\left(y_i | \mathbf{X} \in \mathbf{D}_l\right) + B_s = o_p(1). \tag{24}$$

While the above discussion focuses on one regressor, extension to K regressors is straightforward and requires no additional assumptions on B_s. This approach can also be extended to include interacting class and split sample effects, such as B_{sm} for $s = 1, \ldots, S$ and $m = 1, \ldots, M$, which hopefully would take care of all likely perception effects.

It is theoretically possible to test the impacts of the perception effects on the estimator. Since $\hat{\beta}$ as defined in equation (23) is consistent regardless of the presence of perception effects and

$$\tilde{\beta} = \left(\mathbf{x}'\mathbf{x}\right)^{-1} \mathbf{x}'\tilde{y}^{**} \tag{25}$$

is consistent only in the absence of the perception effects or if the effects are uncorrelated with $\mathbf{x}$, then under the usual regularity conditions, the test statistic is

$$\left(\hat{\beta} - \tilde{\beta}\right)' \left[\text{Var}\left(\hat{\beta} - \tilde{\beta}\right)\right]^{-1} \left(\hat{\beta} - \tilde{\beta}\right) \overset{a}{\sim} \chi^2(K). \tag{26}$$

The exact regularity conditions and the construction of the test statistic would depend on the nature of the perception effect. For example, the case where $\mathbf{B}$ is fixed would be different to the case where $\mathbf{B}$ is a random vector. It would also appear that some assumptions on $\mathbf{B}$ are required in order to compute the test statistics. This is another interesting avenue for future research.

4.2 Non-linear Models

Another possible extension is to consider the application of the proposed methods in the context of non-linear models. Given the presented methods focus on data collection, they could also be applied to non-linear models. To see this, consider

$$y_i = g(x_i; \beta) + u_i \tag{27}$$

where $g(\cdot)$ denotes a continuous function. Let $\mathbf{y}$, $\mathbf{y}^{WS}$ and $\mathbf{x}$ be the data matrix of y_i (if we do observe it), y_i^{WS} (observations from the working sample, and x_i, respectively). Let $\hat{\beta}(\mathbf{y}, \mathbf{x})$ denotes a consistent estimator of β with $\rho(\mathbf{x}) = \sqrt{N}\left[\hat{\beta}(\mathbf{y}, \mathbf{x}) - \beta\right]$ such that $\rho(\mathbf{y}, \mathbf{x}) \xrightarrow{d} f(0, \Omega)$. Under the assumptions made earlier, $y_i^{WS} \xrightarrow{d} y_i$, and therefore $\rho(\mathbf{y}^{WS}, \mathbf{x}) \xrightarrow{d} \rho(\mathbf{y}, \mathbf{x})$ by the continuous mapping theorem under appropriate regularity conditions. The technical details of these conditions, however, could be an interesting subject of future research.

5 Conclusion

This paper deals with econometric models where the dependent variable is continuous but observed through a discretization process that results in interval data. When such a variable is modelled in a (linear) regression framework, the regression parameter(s) cannot be point-identified.

Ordered choice models – which are among the most commonly used to treat such outcome variables – rely on distributional assumptions for point-identification. Alternatively, Manski and Tamer (2002) offer set identifying conditions, which results in large ranges of estimated parameter intervals.

Our proposed split sampling approach does not rely on any distributional assumption and does not restrict the validity to set-identification. Instead, we propose changes in the data gathering process (the way data is collected). We show that parameters can be point-identified and estimated consistently in a regression model. The split sampling approach put forward ensures that the least squares estimator is unbiased and consistent, and it also works well in moderate sample sizes.

The two split sampling method put forward – magnifying and shifting methods – may guide survey designers and researchers who deal with questionnaires, as well as data providers. With the shifting method, data providers can take into account data privacy considerations as well: individuals are not identifiable from the data, however, the data can still be used to simply estimate the parameters of interest.

References

Alwin, D. F. (1992). Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement. *Sociological Methodology*, 22:83–118.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814.

Beresteanu, A., Molinari, F., and Darcy, M. (2010). *Asymptotics for partially identified models in Stata*.

Bhat, C. R. (1994). Imputing a continuous income variable from grouped and missing income observations. *Economics Letters*, 46(4):311–319.

Cameron, C. and Trivedi, P. (2010). *Microeconometrics Using Stata. College Stations*. United States: STATA Press.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284.

Chesher, A. and Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica*, 85(3):959–989.

Diamond, P. A. and Hausman, J. A. (1994). Contingent valuation: Is some number better than no number? *American Economic Review*, 8(4):45–64.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. In Lane, J. I., Doyle, P., Zayatz, L., and Theeuwes, J., editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, chapter 2. North-Holland, Washington DC.

Fox, C. R. and Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, 14(3):195–200.

Givon, M. M. and Shapira, Z. (1984). Response to rating scales: a theoretical model and its application to the number of categories problem. *Journal of Marketing Research*, 21(4):410–419.

Greene, W. H. and Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.

Haisley, E., Mostafa, R., and Loewenstein, G. (2008). Subjective relative income and lottery ticket purchases. *Journal of Behavioral Decision Making*, 21:283–295.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Kaido, H., Molinari, F., and Stoye, J. (2019). Confidence intervals for projections of partially identified parameters. *Econometrica*, 87(4):1397–1432.

Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903.

Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human resources*, pages 343–360.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546.

Micklewright, J. and Schnepf, S. V. (2010). How reliable are income data collected with a single question? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):409–429.

Schomburg, G., Behlau, H., Dielmann, R., Weeke, F., and Husmann, H. (1977). Sampling techniques in capillary gas chromatography. *Journal of Chromatography A*, 142:87 – 102.

Schomburg, G., Husmann, H., and Rittmann, R. (1981). "direct"(on-column) sampling into glass capillary columns: comparative investigations on split, splitless and on-column sampling. *Journal of Chromatography A*, 204:85–96.

Srinivasan, V. and Basu, A. K. (1989). The metric quality of ordered categorical data. *Marketing Science*, 8(3):205–230.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195.