Modelling with Discretized Ordered Choice Covariates

Felix Chan*, Agoston Reguly** and Laszlo Matyas**

April 14, 2021

* Curtin University
** Central European University

Abstract

This paper proposes a new data gathering method, called split sampling which allows the identification and consistent estimation of parameters in a linear regression model with discretized observations. This situation is common when analysing survey data where continuous random variables, such as income or expenditure, for example, are being transformed into either a set of intervals. Such discretization prevents point-identification and least squares type estimators are inconsistent. Split sampling method resolves these problems by improving the design of the survey without creating additional disincentives for respondents and without additional complexity on the design of the survey questions. The proposed methods can consistently reconstruct the distribution of the underlying random variables, which leads to the consistent estimation of the parameters. Since the solution resides in the data collection stage, the proposed methods should also be applicable for the identification of parameters in a non-linear model.

JEL: C01, C13, C21, C25, C83

Keywords: Discretized variable, measurement error, sampling, survey methods.

Acknowledgement: Contribution by Balazs Kertesz to earlier versions of this paper is kindly acknowledged. Special thanks to the IAAE for the financial support of the 2019 conference presentation.

1 Introduction

There is an increasing number of survey-based large data sets where many (sometimes all) variables are observed through the window of individual choices, i.e., by picking one option from a pre-set class list, while the original variables themselves are in fact continuous. For example, in transportation modelling, the US Federal Transportation Office creates surveys to measure different transportation behaviours. This practice is also common for major cities like London, Sydney and Hong Kong. Usually the reported values are a discretized version of variables, like average personal distance travelled, or use of public or private transportation (Santos et al., 2011). Such examples emerges in many other areas, like credit ratings in financial economics, corruption measures or institutional development in political economy. These are discretized variables which have the characteristics of interval data (see Mauro (1995), Méndez and Sepúlveda (2006), Knack and Keefer (1995) and Acemoglu et al. (2002)). Typically, such variables are related to income, expenditure on something over a period of time, willingness to take some action (e.g., how much would you be willing to pay for ... ?)

or questions about likelihood(s) (e.g., how likely would you be to download this application ... ?) and questions related to time (e.g., how much time did you spend commuting last week ... ?).

To formalise the discussion, consider the random variable $x_i \sim f(a_l, a_u)$, where $f(a_l, a_u)$ denotes[1] an *unknown* distribution with support in $[a_l, a_u]$, $a_l, a_u \in \mathbb{R}$ with mean μ for $i = 1, \ldots, N$. Furthermore, define

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

We refer to variable z_m as the choice values for $m = 1, \ldots, M$. It can be a measure of centrality of the given choice class, or can be a completely arbitrarily assigned value. The class boundaries are known and set by the researchers. The main difficulty is that x_i is not directly observable and researchers only observe the response x_i^*. In other words, variable x is observed through the discrete ordered window of x_i^*.

It is not uncommon among empirical researchers to estimate linear regression models using x_i^* instead of x_i as the latter is not available. Manski and Tamer (2002) shows that the parameters in those cases are not point-identifiable, even though they may be partially identifiable. That is, it is possible to identify a region where the true parameters reside. This paper echoes the results in Manski and Tamer (2002) and shows that the Ordinary Least Squares (OLS) estimator is inconsistent in general but can be consistent in a few restricted cases.

More importantly, this paper proposes a new data gathering technique, that we call *split sampling*[2], which can consistently estimate the underlying distribution of the unobserved random variables, and thus, lead to consistent estimation of the parameters in (linear) regression models. The basic idea is to allow each survey to have different class boundaries. This induces additional information on the distribution of the random variables when considering all the responses as a whole. The proposed techniques do not induce any disincentive for respondents since the number of choices of each question remains the same. It also does not create additional complexity in the design of the questions, since the adjustments focus on the responses rather than the questions. Perhaps more importantly, the proposed solution focuses on the data collection stage and is invariant to the relation between the variables.

The organisation of the paper is as follows. Section 2 motivates the problem from both empirical and theoretical perspectives. It shows that OLS is inconsistent in general, except in a few restricted cases, and provides support to the results in Manski and Tamer (2002) on the limit of identification when using discretized data that share the same boundary points. Section 3 introduces the two split sampling techniques namely, the *magnifying* and *shifting* methods, that allow consistent estimation of the underlying distribution as well as consistent

[1] A complete list of the notations used in the paper is given in the on line Supplementary Materials here.

[2] 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.

estimation of the parameters in the linear regression model using discretized data. The finite sample performance of these techniques can be found in Section 4. Section 5 discusses some possible extensions of the techniques and some concluding remarks are in Section 6. All technical proofs, additional Monte Carlo results and algorithms for implementing the two sub-sampling methods can be found in the online Supplementary Materials here .

2 Motivation

Consider the following data generating process

$$y_i = w_i'\gamma + x_i'\beta + u_i, \tag{2}$$

and the following linear regression model

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

where $i = 1, \ldots, N$, w is a $K_1 \times 1$ vector of explanatory variables that can be directly observed, x is a $K \times 1$ vector of continuous random variables that cannot be directly observed and x^* is the corresponding $K \times 1$ vector of discretized choice variables as defined in (1). u_i is an idiosyncratic disturbance term for model (2) with $\varepsilon_i = (x_i - x_i^*)'\beta + u_i$ denotes the residuals for model (3). γ and β are unknown parameter vectors that conform to w_i and x_i, respectively. We also maintain the assumption of independence between individuals. The two main questions are the identification and consistent estimation of β based on model (3). Equation (2) and model (3) represent a common problem in empirical research.

Let us take an example from the transportation economics literature. Assume that in a given city we would like to model the factors explaining individual transport expenditures (TE) in a given period of time, using the simple model

$$TE_i = w_i'\gamma + \beta\, UPT_i + \varepsilon_i\,, \tag{4}$$

where TE_i is the transport expenditure for individual i, w_i are 'usual' controls and UPT_i is the use of public transport in commuting measured in percentage points. 100% if only public transport (PT) was used and 0% if PT was not used at all for individual i ($i = 1, \ldots, N$). Now UPT is not observed; instead we observe only the individual's choice from a pre-set list UPT^*. We ask the use of public transport in the following way

$$\begin{array}{ll} 1 \rightarrow & \text{took almost only public transport,} \\ 2 \rightarrow & \text{took mostly public transport,} \\ 3 \rightarrow & \text{mostly did not take public transport,} \\ 4 \rightarrow & \text{took almost no public transport,} \end{array} \tag{5}$$

which is referred to the following intervals

$$UPT_i^* = \begin{cases} 1, & \text{if} \quad 90\% \leq UPT_i \leq 100\%, \\ 2, & \text{if} \quad 50\% \leq UPT_i < 90\%, \\ 3, & \text{if} \quad 10\% \leq UPT_i < 50\%, \\ 4, & \text{if} \quad 0\% \leq UPT_i < 10\%. \end{cases} \tag{6}$$

We can set the choice value as the mid-value of each class to UPT_i^* so that

$$UPT_i^* = \begin{cases} 0.95 \rightarrow \text{took almost only public transport}, & \text{if} \quad 90\% \leq UPT_i \leq 100\%, \\ 0.70 \rightarrow \text{took mostly public transport}, & \text{if} \quad 50\% \leq UPT_i < 90\%, \\ 0.30 \rightarrow \text{mostly did not take public transport}, & \text{if} \quad 10\% \leq UPT_i < 50\%, \\ 0.05 \rightarrow \text{almost did not take public transport}, & \text{if} \quad 0\% \leq UPT_i < 10\%. \end{cases} \tag{7}$$

Obviously, we can use many possible representations for the responses. Using the mid-values seems to be reasonable when the only available information is that an observation is in a given class.

To the best of our knowledge, with the exception of Hsiao (1983) and Manski and Tamer (2002), there has been no study investigating the estimation of discretized continuous variable(s) when the categories/classes are not represented by the expected values of the underlying distribution(s). There is, however, some work on related issues. Taylor and Yu (2002) consider a regression model with three multivariate normal random variables. In their setting, the response variable is correlated with the first variable while the second variable does not affect the response variable conditional on the first. They show that if one dichotomizes the first variable, the least squares estimate of the coefficient for the second variable will be biased, However, they do not extend their results to the more general settings where the response variable may depend on more than two covariates. Lagakos (1988) analyses the correct cut values for the grouping of continuous explanatory variables. He derives a test on deviating from the expected group mean and the categorized value if the group mean is known. He refers to this solution as the optimization criterion for discretizing an explanatory variable, using the argument in Connor (1972).

There are many papers considering the discretization of a continuous variable, but all assume that the choice values properly represent each class. In these papers, the main question is the effect of discretization in terms of efficiency loss (see, for example, Cox (1957), Cohen (1983), Johnson and Creech (1983)).

The measurement error literature has not considered the problem in details either, as it has been assumed that the class choice values are taking the expected values of the known underlying distribution (Wansbeek and Meijer, 2001), or the measurement error is on top of a categorized variable (Buonaccorsi, 2010).

Hsiao (1983) shows that OLS is inconsistent in general when assigning z_m using the mid-point values.[3] In a seminal paper, Manski and Tamer (2002) extend the result and show that β in model (3) is not point-identifiable without any further assumption and can only be partially identifiable. That is, it is possible to identify the region in which β resides. However, this region cannot be estimated using the OLS estimator on model (3) as it is inconsistent. In fact, it can be shown that when $K = 1$[4]

$$\operatorname*{plim}_{N\to\infty} \left(\hat{\beta}_{OLS}^* - \beta\right) = \frac{\beta \sum_{m=1}^{M} z_m \left\{\mathbb{E}(x_i | x_i \in C_m) - z_m\right\}}{\sum_{m=1}^{M} z_m^2}. \tag{8}$$

Equation (8) is insightful for two reasons. First, the right-hand side is generally not zero which shows that OLS is inconsistent in general. Second, the right-hand side can be zero

[3]Hsiao (1983) offers a maximum likelihood method to overcome this problem, using a strong distributional assumption for point-identification

[4]Detailed derivations and in-depth analysis on the consistency of the OLS estimator for Equation (2) and Model (3) can be found here.

when $\mathbb{E}(x_i|x_i \in C_m) = z_m$. That is, when the choice value equals the expectation of the explanatory variable given its value lies in the corresponding class.

The result here also justifies the Berkson model (see Berkson (1980) and Wansbeek and Meijer (2000) pp. 29-30). That is, if $f(\cdot)$, the probability density of x_i, is known with known boundaries, the expected value of each variable in x^* can be consistently estimated. As such, the OLS estimator of model (3) is consistent.

Another implication is that assigning mid-point values of each class to the choice values would make sense if one could safely assume that the explanatory variable follows a uniform distribution. In that case, the mid-point value equals the conditional expectation in equation (8).

Together with the results from Manski and Tamer (2002), there are two immediate conclusions: (i) There is a limit on the identification of parameters. That is, β cannot be point-identified under equation (2) and model (3) and the procedure for partial identification is complicated. (ii) It follows from (i) and the analysis above, that simple techniques, such as the OLS estimator, do not seem to be appropriate even when partial identification is possible.

In the next section, we introduce split sampling method that can resolve these identification and estimation issues.

3 Split sampling

Since there is a limit on identification given the data, one lateral solution is to improve the information content of the data at the data collection stage. This improvement must satisfy two criteria. First, it cannot induce additional disincentive for respondents. That is, the design of the survey cannot create an additional hurdle for respondents to answer the questions truthfully. Second, it cannot create additional complications in the design of the survey questions.

The main approach of the proposed methods is to create a number of split samples (S), while fixing the number (M) of choices in each split sample, in order to reduce the estimation bias. The reason for fixing M is the restricted human cognitive capacity as noted above. Nevertheless, we can achieve an increase in M through changing the class boundaries in each split sample, which in practice means different survey questionnaires for each split sample.

The intuition behind the method is that this leads to a better mapping of the unknown distribution of x and thus reduces the estimation bias. By merging the different split samples into one data set, which we call the '*working sample*'. With 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 (individual i) is given one questionnaire only[5]. The set of respondents who fill in the questionnaire with the same class boundaries defines a split sample. Each split sample has $N^{(s)}$, number of observations ($s = 1, \ldots, S$, $\sum_s N^{(s)} = N$).

In this setup, a split sample looks exactly as the problem introduced above in (1), with the only difference across the split sample that the class boundaries are different.[6] Note that the number of observations across split samples can be the same or, more likely, different.

[5] In the case of cross sectional data. For panel data one shall assign different questionnaires for each individual through time.

[6] In order to simplify the notation, we use instead of $x^{*(s)}$ simply $x^{(s)}$. For each split sample the discretization of x result in a new random variable.

Now a split sample looks like,

$$x_i^{(s)} = \begin{cases} z_1^{(s)} & \text{if } x_i \in C_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ & \qquad \text{1st choice for split sample } s, \\ z_2^{(s)} & \text{if } x_i \in C_2^{(s)} = [c_1^{(s)}, c_2^{(s)}), \\ \vdots & \vdots \\ z_m^{(s)} & \text{if } x_i \in C_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ z_M^{(s)} & \text{if } x_i \in C_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last choice for split sample } s. \end{cases} \tag{9}$$

Let us take a very simple illustrative example. Assume that $M = 2$, $S = 2$, $N = 60$, $N^{(1)} = 30$ and $N^{(2)} = 30$. Let x be a continuously distributed variable in $[0, 4]$ and define the class boundaries in the first split sample as $[0, 2)$ and $[2, 4]$, while in the second split sample $[0, 1)$ and $[1, 4]$, with $10, 20, 5$, and 25 observations respectively in each class. Next, we merge the information obtained in the two split samples in one working sample in such a way that we are not introducing any selection bias. This working sample now has $B = 3$ classes (or bins): $[0, 1)$, $[1, 2)$ and $[2, 4]$ and number of observations N^{WS} with the working sample's artificially created variable x_i^{WS}. Using the information from the 2nd split sample, we know that of 30 observations 5 are in the 1st bin. Similarly, we can deduce that in the 2nd bin there are 5 observations as well, while in the last 3rd bin 20 (see Figure 1 below). Piecing this information together, we can create x_i^{WS}. Clearly, this way the working sample maps the unknown distribution of x better than any one of the two split samples.

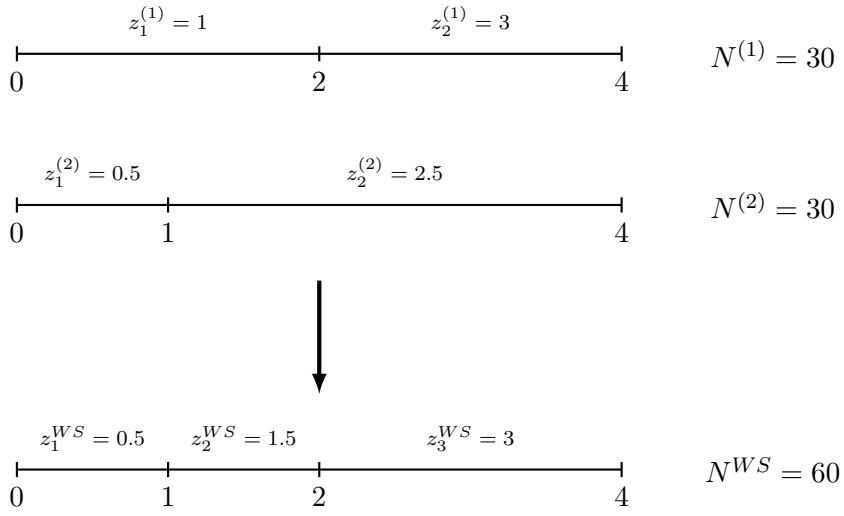

Figure 1: The basic idea of split sampling

3.1 Construction of the Working Sample

The construction of questionnaires for each split sample and the merger into the working sample can be done in many ways, depending on the assignments of boundary points ($c_m^{(s)}$) and on the choice values ($z_m^{(s)}$) for each split samples. We assume that the number of observations

(N), their allocation among split samples $(N^{(s)})$ and the number of split samples (S) are given, and also that the number of choices (M) is fixed across the split samples. We also assume now that the class boundaries in the working sample are the union of the split samples' class boundaries, that is

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

This translates in our example to the following: $c_0^{WS} = c_0^{(1)} = c_0^{(2)} = 0$; $c_1^{WS} = c_1^{(2)} = 1$; $c_2^{WS} = c_1^{(2)} = 2$; $c_3^{WS} = c_2^{(1)} = c_2^{(2)} = 4$.
Also, we restrict the domain of the underlying distribution for each split sample. We construct the split sample questionnaires' and the working sample's boundary points so that: $a_l = c_0^{(s)} = c_0^{WS},\ a_u = c_M^{(s)} = c_B^{WS},\ \forall s$. It is possible to accommodate distribution with infinite support $(a_l = -\infty, a_u = \infty)$. In this case all split samples will have infinite boundary points at the boundaries.

With the creation of S split samples, we introduce

$$x^{(1)}, \ldots, x^{(s)}, \ldots, x^{(S)}$$

new random variables $(x^{(s)} := \psi^{(s)}(x))$, where $\psi^{(s)}(\cdot)$ is the function that discretizes the continuous x into the choices of the split sample s. These then define a new random variable, $x^{WS} = \Psi(x^{(1)}, \ldots, x^{(s)}, \ldots, x^{(S)})$ representing the working sample, where $\Psi(\cdot)$ is the 'merging function'.

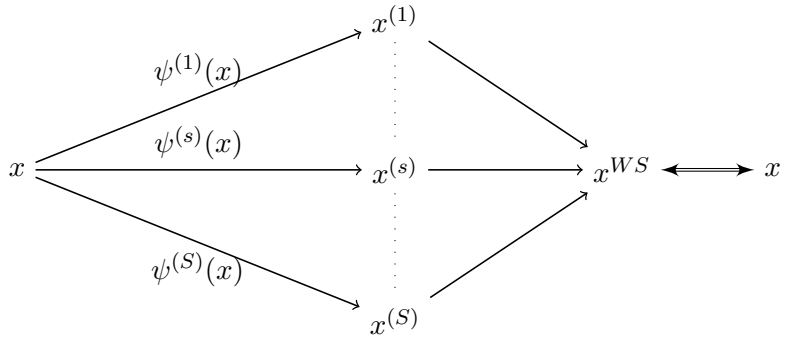

Figure 2: Creation of the working sample's random variable

Each method to be discussed below specifies the functions $\psi^{(s)}$, the merging function $\Psi(\cdot)$ and defines the random variable of the working sample x^{WS}. These functions are different across the methods, but all of them reflect the unknown random variable x. To do so, we need the following property to hold

$$\lim_{s \to \infty} \mathbb{E}_S \left[x^{WS} | y \right] = \mathbb{E} \left[x | y \right] ,$$

which means that no selection bias is introduced through the mapping.

3.2 Probabilities in the Working Sample

In addition to avoiding 'selection bias' through the merger of the split samples, we also need to calculate the probability of a given observation in a given choice class in the working sample.

To derive this, we have to derive the probability of an observation falling into a given split sample's choice class and introduce an assigning mechanism taking an observation in a split sample to a working sample class. Using this result, we can get the unconditional probability for an observation to be in a given class in the working sample.

All individuals are initially allocated into a split sample. This, of course, defines the number of observations in each split sample ($N^{(s)}$), which in turn translates into the probability of a given observation x being in split sample s: $\Pr(x \in s)$. Uniformly assigning these individuals to split samples is the most straightforward procedure, thus $\Pr(x \in s) = 1/S$, however for the general case we are going to use the probabilistic notations.
Now, we can define the probability for an observation to be in a given class in a given split sample,

$$\Pr\left(x \in C_m^{(s)}\right) = \Pr(x \in s) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,.$$

As we observe a response in a given split sample, we would like to derive the probability of this observation falling between given boundary points in the working sample. We then assign these uniformly into the working sample's classes to avoid introducing selection bias in the merging process. [7]

$$\Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) = \begin{cases} \frac{c_b^{WS} - c_{b-1}^{WS}}{c_m^{(s)} - c_{m-1}^{(s)}}, & \text{if} \quad c_b^{WS} \le c_m^{(s)} \text{ and } c_{b-1}^{WS} \ge c_{m-1}^{(s)}\,, \\ 0, & \text{otherwise}\,. \end{cases}$$

Using the above two equations, we need to assign each individual from all split samples into the working sample without any additional information. Thus, the unconditional probability of an individual falling in the working sample between given boundary points is

$$\Pr\left(x \in C_b^{WS}\right) = \sum_{s=1}^{S} \Pr(x \in s) \sum_{m=1}^{M} \Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,. \tag{10}$$

To simplify, we can assume uniform assignment of the observations to each split sample, and write

$$\Pr\left(x \in C_b^{WS}\right) = \frac{1}{S} \sum_{s=1}^{S} \sum_{\substack{m \\ \text{if } C_b^{WS} \in C_m^{(s)}}} \frac{c_b^{WS} - c_{b-1}^{WS}}{c_m^{(s)} - c_{m-1}^{(s)}} \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,.$$

In some cases x may have infinite support, which implies classes not bounded from below and/or above. Usually, this is related to survey questions like *"... or less"* or *"... or more"*. Here we face censoring. As a consequence, the difference between the class's choice value (e.g., $z_1^{(s)}$ in Equation (1)) and the class's conditional mean for the underlying distribution can be potentially infinite, resulting in very large estimation biases. We will return to this issue later in the paper.

3.3 Magnifying Method

In the magnifying method, we magnify the domain of the answers within the original domain of the unknown distribution of x by one equally sized choice class. The size of each of the

[7] Here we use the assumption that the boundary points in the working sample are the union of the split samples' boundary points.

classes depends on the number of split samples (S) and the number of choice values (M). As the number of split samples increases class sizes decrease, which is the main mechanism to uncover the unknown distribution. Figure 3 shows the main idea of the magnifying method: the last line shows the working sample, while above, we can see the individual questionnaires for the case of $M = 3, S = 4$.

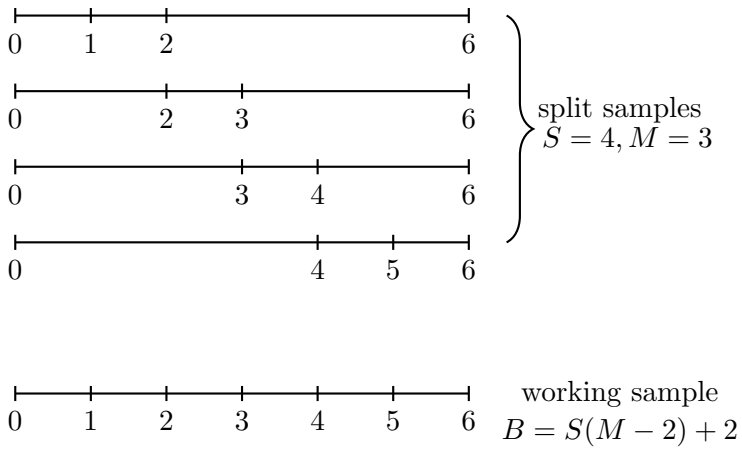

Figure 3: The magnifying method

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. To further explore the properties of the magnifying method, let us establish the connection between the number of magnified classes in the working sample (B), and the number of split samples (S) and choices (M)

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

As mentioned above, we have 2 split samples, which lie in the boundary of the domain and capture $M-1$ classes of equal size; and there are $S-2$ split samples in between which capture $M-2$ classes. After some manipulations, we get the number of the classes in the working sample.

Given the fact that there are B classes in the working sample, we get the widths of these classes, let us call it h such

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

By fixing the upper and lower bounds on the support for the split samples ($a_l = c_0^{WS} = c_0^{(s)};\ a_u = c_B^{WS} = c_M^{(s)},\ \forall s$), one can reduce the class size $h \to 0$ as $S \to \infty$. This can also be seen through the working sample's boundary points, which have the following simple form

$$c_b^{WS} = a_l + bh = a_l + b\frac{a_u - a_l}{S(M-2)+2}\,.$$

To show how the number of split samples affects the bias, we need the boundary points for

each split sample, which can be derived as

$$c_m^{(s)} = \begin{cases} a_l \text{ or } -\infty & \text{if } m = 0, \\ a_l + mh & \text{if } 0 < m < M \text{ and } s = 1, \\ a_l + h\left[(s-2)(M-2) + M + m - 2\right] & \text{if } 0 < m < M \text{ and } s > 1, \\ a_u \text{ or } \infty & \text{if } m = M. \end{cases} \tag{11}$$

The intuition behind this is that on the boundaries of the support, the split samples take the values of the lower and upper bounds. For the first split sample, one needs to shift the boundary points m times. However, for the other split samples, one needs to push by $h(M-1)$ times to shift through the first questionnaire and then $h(M-2)$ to shift through each split sample in between $s = 2$ and $s = S - 1$, $s - 2$ times. Deriving this process algebraically will result in the above expression.[8] Algorithm 1 in the online Supplementary Materials can be used to construct split samples using the shifting method.

From Equation (11), it is clear that the class widths differ from each other within a split sample. Let $||C_m^{(s)}|| = c_m^{(s)} - c_{m-1}^{(s)}$ be the m-th class width, then for the split samples which are in-between the boundaries ($1 < s < S$) and substituting for h, we can write

$$||C_m^{(s)}|| = \begin{cases} (a_u - a_l)\left(\frac{s(M-2)+2}{S(M-2)+2} + \frac{1-M}{S(M-2)+2}\right) & \text{if } m = 1,\ 1 < s < S, \\ \frac{a_u - a_l}{S(M-2)+2} & \text{if } 1 < m < M,\ 1 < s < S, \\ (a_u - a_l)\left(1 - \frac{s(M-2)+1}{S(M-2)+2}\right) & \text{if } m = M,\ 1 < s < S\,. \end{cases}$$

We can also define the class widths for the first and last split samples as

$$||C_m^{(1)}|| = \begin{cases} \frac{a_u - a_l}{S(M-2)+2} & \text{if } 1 \le m < M, \\ (a_u - a_l)\left(1 - \frac{M-1}{S(M-2)+2}\right) & \text{if } m = M, \end{cases}$$

$$||C_m^{(S)}|| = \begin{cases} (a_u - a_l)\left(1 - \frac{M-1}{S(M-2)+2}\right) & \text{if } m = 1, \\ \frac{a_u - a_l}{S(M-2)+2} & \text{if } 1 < m \le M. \end{cases}$$

Note that $||C_m^{(s)}|| \le ||C_1^{(s)}||$ and $||C_m^{(s)}|| \le ||C_M^{(s)}||$. Formally, let us define $\zeta := \{C_m^{(s)} \mid 1 < m < M, 1 < s < S, C_m^{(1)} \mid 1 \le m < M, C_m^{(S)} \mid 1 < m \le M\}$ as the set of classes which have the class width $\frac{a_u - a_l}{S(M-2)+2}$. Then we can write $\Pr\left((x - x^{(s)})^2 | x \in \zeta \le (x - x^{(s)})^2 \mid x \notin \zeta\right) = 1$, which is true if and only if, $\mathbb{E}[x] = \mathbb{E}\left[x^{(s)}\right], \forall x$. One example is when x is uniformly distributed.

Now, let us check the limit in the number of split samples. We end up with the following limiting cases

$$\lim_{S\to\infty}\left(||C_m^{(s)}||\right) = \begin{cases} 0 & \text{if } 1 \le m < M,\ 1 < s < S, \\ a_u - a_l & \text{if } m = M,\ 1 < s < S; \end{cases}$$

and for the first and last split sample

$$\lim_{S\to\infty}\left(||C_m^{(1)}||\right) = \begin{cases} 0 & \text{if } 1 \le m < M, \\ a_u - a_l & \text{if } m = M, \end{cases}$$

$$\lim_{S\to\infty}\left(||C_m^{(S)}||\right) = \begin{cases} a_u - a_l & \text{if } m = M, \\ 0 & \text{if } 1 < m \le M. \end{cases}$$

[8] There is an alternative way to formalize the boundary points, when one starts from a_u. The formalism will result in the same conclusions.

This formulation takes a_l as the starting point and expresses the boundary points given a_l. However, we can use a_u as the starting point as well to shift the boundary point. This implies that the convergences on the bounds $(||C_1^{(s)}||, ||C_M^{(s)}||)$ will change, resulting in those parts not converging to 0 in general.

Based on the different magnitudes of measurement errors and depending on class widths, it is clear that there are two types of observations: The first type is $x_i^{(s)} \in \zeta$. Here, the error is the smallest and has the feature of $\lim_{S\to\infty} ||C_m^{(s)}|| = 0$. Moreover, these observations have the same class width as the working sample's classes and each of them can be directly linked to a certain working sample class by design. Formally, $\exists\, C_m^{(s)} \cong C_b^{WS}$ such that $c_m^{(s)} = c_b^{WS}$, $c_{m-1}^{(s)} = c_{b-1}^{WS}$. We call these values '*directly transferable observations*', as we can directly transfer and use them in the working sample. These observations are denoted by $x_{i,DTO}^{WS} := x_i^{(s)} \in \zeta,\ \forall s$, and the related random variable by x_{DTO}^{WS}.[9]

The second type of observations are all others for which none of the above is true. We call them '*non–directly transferable observations*'. Algorithm 2 in the online Supplementary Materials describes how to construct the working sample in practice, using the directly transferable observations.

Before proving the consistency of $\hat{\beta}$, using only $x_{i,DTO}^{WS}$ — the *directly transferable observations* in the working-sample — we need to make some assumptions on these observations.

The probability that a *directly transferable observation* lies in a given class of the working sample can be written based on Equation (10) as follows

$$\Pr\left(x \in C_b^{WS}\right) = \Pr(x \in s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x\,.$$

Here we used the fact that individual i being assigned to a split sample s is independent of i choosing the class with choice value $z_m^{(s)}$.

We want to ensure that in each class in the working sample, there are directly transferable observations. This will ensure that the estimation is feasible. Thus, for each split sample the expected number of directly transferable observations is

$$\begin{aligned}\mathbb{E}(N_b^{WS}) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_b^{WS}\}}\right) \\ &= N \Pr(x \in s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x.\end{aligned} \tag{12}$$

Following from Equation (12), consider the following assumptions,

Assumptions

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

[9]Notation: for the estimation we use the superscript 'WS' and define the construction method in the subscript – here 'DTO'.

Assumption 1 ensures that the number of respondents will always be higher than the number of split samples. Assumption 2 ensures utilisation of all split samples, i.e. each split sample will have non-zero respondents. Assumption 3 imposes a mild assumption on the underlying distribution. That is, the support of the random variable is not disjoint. This implies $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x > 0$. These assumptions allow us to establish the following.

Proposition 1. *Under Assumptions 1 - 3,*

1.

$$\mathbb{E}(N_b^{WS}) > 0 \tag{13}$$

2.

$$\Pr\left(\sum_{i=1}^{b} N_b^{WS} > 0\right) \to 1. \tag{14}$$

3. $\Pr\left(x_{DTO}^{WS} < c\right) = \Pr\left(x < c\right)$ *for any* $c \in [a_l, a_u]$.

The proposition established convergence in distribution which allows consistent estimation of the underlying continuous distribution. This implies that the classical econometric results stand and the OLS estimator is consistent for β.

Note that 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 $\mathrm{c} \geq 1$, 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.

Next, let us consider the placement of the *non-directly transferable observations*. We have seen that these observations do not reduce the measurement error in a systematic way. One way to proceed is to remove them completely so that they do not appear in the working sample (thus only using $x_{i,DTO}^{WS}$). However, it seems that too many could fall into this category, resulting in a large efficiency loss in estimation.

Another approach is to utilise the information available for these observations namely, the known boundary points for these values. Then we could use all the *directly transferable observations* from the working sample to calculate the conditional averages for all *non-directly transferable observations* and replace them with those values. This way one could construct a variable, which has the same number of observations as the number of respondents. Let us denote this new variable $x_{i,ALL}^{WS}$. This represents all the directly transferable observations and the replaced values for non-directly transferable observations.
Let us formalize the non-directly transferable observations as $x_i^{(s)} \in C_\chi$, where

$$C_\chi := \bigcup_{s,m} C_m^{(s)} \bigcap_b C_b^{WS} = \zeta^{\complement}$$

is the set for non-directly transferable observations from all split samples, with $\chi = 1, \ldots, 2(S-1)$. We can then replace $x_i^{(s)} \in C_\chi$ with $\hat{\pi}_\chi$, which denotes the sample conditional averages

$$\hat{\pi}_\chi = \left(\sum_{i=1}^{N} \mathbf{1}_{\{x_{i,DTO}^{WS} \in C_\chi\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}_{\{x_{i,DTO}^{WS} \in C_\chi\}} x_{i,DTO}^{WS}.$$

Let us introduce $x^{WS}_{i,NDTO}$ as the variable which contains all the replaced values with $\hat{\pi}_\chi$, $\forall x_i^{(s)} \in C_\chi$. This way we can create a new working sample as $x^{WS}_{i,ALL} := \{x^{WS}_{i,DTO}, x^{WS}_{i,NDTO}\}$, which contains information from both types of observations.

Let us call $\hat{\pi}_\chi$ the 'replacement estimator' of the conditional expectation of the given class. Under the WLLN, it is straightforward to show that the 'replacement estimator' for the sample conditional averages converges to the conditional expectations, thus $\hat{\pi}_\chi \to \mathbb{E}(x|x \in C_\chi)$ as $N, S \to \infty$ under the same assumptions as before. This also implies $x^{WS}_{i,NDTO} \to \mathbb{E}(x|x \in C_\chi)$, which means we are not introducing any errors during the estimation when working sample $x^{WS}_{i,ALL}$. Algorithm 3 in the online Supplementary Materials describes how to create in practice the working sample using all observations.

We also need to obtain the asymptotic standard errors of this estimator, because if these are large, the replacement might not be favorable, as it may induce more uncertainty relative to the potential loss of efficiency by not including all the observations. To obtain the standard errors, one can think of $\hat{\pi}_\chi$ as an OLS estimator, regressing $\mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}$ on $x^{WS}_{i,DTO}$. Here $\mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}$ is a vector of indicator variables, created by $2(S-1)$ indicator functions: It takes the value of one for the directly transferable observations, which are within C_χ.[10] We can now write the following:

$$x^{WS}_{i,DTO} = \boldsymbol{\pi}_\chi \mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}} + \eta_i,$$

where $\boldsymbol{\pi}_\chi$ stands for the vector of $\pi_\chi, \forall \chi$. The OLS estimator of $\boldsymbol{\pi}_\chi$ is

$$\hat{\boldsymbol{\pi}}_\chi = \left(\mathbf{1}'_{\{x^{WS}_{i,DTO} \in C_\chi\}} \mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}\right)^{-1} \mathbf{1}'_{\{x^{WS}_{i,DTO} \in C_\chi\}} x^{WS}_{i,DTO}\,,$$

and under the standard OLS assumptions, we can write

$$\sqrt{N^{WS}_{DTO}}\,(\hat{\boldsymbol{\pi}}_\chi - \boldsymbol{\pi}_\chi) \overset{\text{a}}{\sim} \mathcal{N}\,(\mathbf{0}, \boldsymbol{\Omega}_\chi)\,,$$

where $\boldsymbol{\pi}_\chi = \mathbb{E}(x|x \in C_\chi), \forall \chi$.

The variance of the OLS estimator is

$$\boldsymbol{\Omega}_\chi = \mathrm{V}\,(\eta_i)\left(\mathbf{1}'_{\{x^{WS}_{i,DTO} \in C_\chi\}} \mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}\right)^{-1}.$$

Using this result, we may decide whether to replace or not to replace.

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 not having 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, $F_{n,s}(x^{WS} \in \zeta^{tr}) \xrightarrow{d} F(x^{tr})$.[11] Furthermore, the parameter estimates $\beta^{tr} = \beta$ (under some reasonable assumptions), which implies that the OLS estimator is consistent for the truncated observations. 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.

[10] The indicator variables are not independent of each other, while the non-transferable observation classes (C_χ) are overlapping each other.

[11] $\zeta^{tr} := \zeta \cap \{C_1^{WS}, C_B^{WS}\}$.

3.4 Shifting Method

The shifting method approaches the problem in a different way. It takes the original questionnaire as given, with fixed class widths, and shifts the boundaries of each choice with a given fixed value. This results in fixed class widths for the different split samples, except in the boundary classes where the widths are changing. Increasing the split sample size does not affect the boundary widths in between the support, only the size of the shift. We can approach this method in two ways. Logically we could consider the original questionnaire, and take the number of choices as fixed here, then as we shift the boundaries, add an extra class for each split sample at the boundary where, due to the shift, the class width has increased. For the mathematical derivations, however, it is more convenient to look at each split sample separately and take the number of classes in each split sample as given, with the exception of the first split sample, which we will regard as the starting benchmark. The first split sample in this case has one fewer class. The discussion below focuses on this approach and Figure 4 shows the split samples following this logic with $S = 4$ and with $M = 4$ classes.

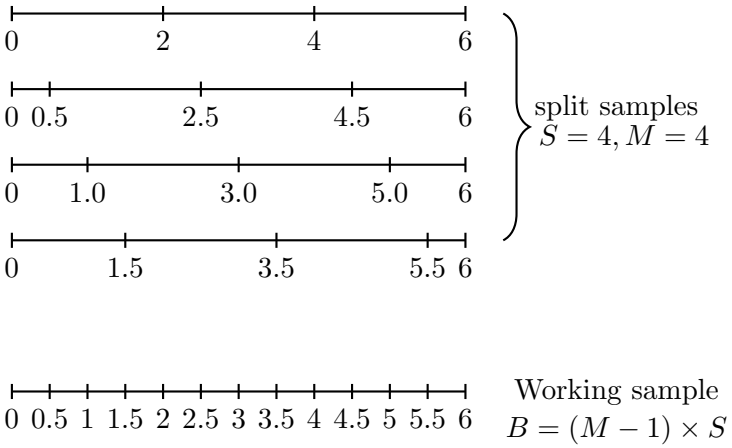

Figure 4: The shifting method

As Figure 4 shows there is one split sample (the benchmark $s = 1$) where there is one class fewer $(M - 1)$. Or, if we prefer, we can look at the benchmark as where we shifted the boundaries with zero. To get the properties of the working sample, let us define the class widths for the first split sample as $\frac{a_u - a_l}{M-1}$. We want to split this into S part in order to be able to shift the boundaries S times in order to have S split samples. Thus, the size of the shift is $\frac{a_u - a_l}{S(M-1)}$. This way we can define the number of classes in the working sample as

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

Now, the boundary points for each split sample are

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

For the working sample, we get $c_b^{WS} = a_l + b\frac{a_u - a_l}{S(M-1)}$. The class widths are

$$||C_m^{(s)}|| = \begin{cases} 0, & \text{if } s = 1 \text{ and } m = 1, \\ \frac{a_u - a_l}{M-1}, & \text{if } 1 < m < M, \\ (s-1)\frac{a_u - a_l}{S(M-1)}, & \text{otherwise.} \end{cases}$$

and for the class size in the working sample, $||C_b^{WS}|| = \frac{a_u - a_l}{S(M-1)}$.

Some additional remarks on the boundary points:

- $C_1^{(1)}$ has size 0 and does not exist in practice. Theoretically, it is induced by the formalism.
- There are only two classes in the split samples which are congruent (with the same boundary points) with the classes in the working sample: $C_1^{(2)} \cong C_1^{WS}, C_M^{(S)} \cong C_B^{WS}$. This means that directly transferable observations will not help us here.
- One cannot decrease the class widths between $C_2^{(s)}$ and $C_{M-1}^{(s)}$ in the split samples by increasing the number of split samples.
- However, the class widths in the working sample can be decreased by increasing the number of split samples.

Algorithm 4 in the online Supplementary Materials describes how to create in practice split samples using the shifting method. The general idea is to reconstruct the underlying distribution $f(x)$, with creating a new random variable, which incorporates the information content of the boundary points.

The observations from a particular class in the split sample s can end up in several classes in the working sample so the union of these classes gives one of the classes from the split samples

$$C_m^{(s)} = \begin{cases} \emptyset, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1} C_b^{WS}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-2)(M-1)}^{s-1+(m-1)(M-1)} C_b^{WS}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B} C_b^{WS}, & \text{if } m = M. \end{cases} \tag{15}$$

Now, define $Z(s, m)$, which creates sets for the scalar values of the working sample's choice values (z_b^{WS}) for each split sample class $C_m^{(s)}$,

$$Z(s,m) = \begin{cases} \{\emptyset\}, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1} \{z_b^{WS}\}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-2)(M-1)}^{s-1+(m-1)(M-1)} \{z_b^{WS}\}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B} \{z_b^{WS}\}, & \text{if } m = M. \end{cases} \tag{16}$$

The number of elements in $Z(s, m)$ depends on the split sample and its class. We use these sets to create a new artificial variable $x_i^\dagger$.

In order to demonstrate this, let us start with an example. *Let an observation $x_i^{(s)} \in C_m^{(s)}$. From Equation (15) we know which working sample classes are included in $C_m^{(s)}$. Furthermore, we also have a set of possible working sample choice values $Z(s, m)$. Now $x_i^\dagger$ will be a randomly*

chosen element of $Z(s,m)$, using uniform probabilities.
The assignment mechanism can be written as

$$x_i^\dagger | x_i^{(s)} \in C_m^{(s)} = z \in Z(s,m), \text{ with } \begin{cases} \Pr(1), & \text{if } s=1 \text{ and } m=1, \\ \Pr(1/(s-1)), & \text{if } s \neq 1 \text{ and } m=1, \\ \Pr(1/S), & \text{if } 1 < m < M, \text{ or} \\ \Pr(1/(S-s+1)), & \text{if } m = M. \end{cases} \tag{17}$$

While by the definition, there is a direct mapping between $z \in Z(s,m)$ and C_b^{WS}, we can write the probability of $x_i^\dagger \in C_b^{WS}$, using Equation (10) and assuming $\Pr(x \in s) = 1/S$,

$$\Pr\left(x_i^\dagger \in C_b^{WS}\right) = \begin{cases} 0, & \text{if } s=1 \text{ and } m=1, \\ \frac{1}{S}\sum_{s=2}^{S} \frac{1}{s-1} \int_{C_1^{(s)} | C_b^{WS} \in C_1^{(s)}} f(x)\mathrm{d}x, & \text{if } s \neq 1 \text{ and } m=1, \\ \frac{1}{S^2}\sum_{s=1}^{S} \int_{C_m^{(s)} | C_b^{WS} \in C_m^{(s)}} f(x)\mathrm{d}x, & \text{if } 1 < m < M, \\ \frac{1}{S}\sum_{s=1}^{S} \frac{1}{S-s+1} \int_{C_M^{(s)} | C_b^{WS} \in C_M^{(s)}} f(x)\mathrm{d}x, & \text{if } m = M\,. \end{cases} \tag{18}$$

Algorithm 5 in the online Supplementary Materials describes how to create an artificial variable which approximates the underlying distribution of x.
It is possible to show that the distribution of this new variable converges to the distribution of the true underlying random variable (x) as we increase the number of split samples.

Proposition 2. *Under Assumptions 1, 3 and* $\Pr(x \in s) = 1/S$,

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

or

$$\lim_{S\to\infty} F_S(c) = F(c) \qquad \forall c \in (a_l, a_u),$$

where $F_S(c) = \Pr(x^\dagger < c)$ *and* $F(c) = \Pr(x < c)$.

In addition, we are also able to investigate the speed of convergence, as we increase the number of split samples (S). Details are given in the online Supplementary Materials here.

Note that we cannot directly use $x_i^\dagger$ 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

$$\mathbb{E}\left(x_i^\dagger \in C_m^{(s)}\right) = \mathbb{E}\left(x_i^{(s)} \in C_m^{(s)}\right) \neq \mathbb{E}\left(x_i \in C_m^{(s)}\right).$$

However, while $F_S(x^\dagger)$ approximates the underlying distribution, we can use these values to calculate the sample conditional means for a given split sample class. Thus, the idea is to use this artificial distribution to calculate the conditional means and replace the class observations with these values.
Let $\hat{\pi}_\tau$ be the replacement estimator for the shifting method, where $\tau = 1, \ldots, S \times M$. Let us define

$$\hat{\pi}_\tau := \left(\sum_{i=1}^{N} \mathbf{1}'_{\{x_i^{(s)} \in C_m^{(s)}\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}'_{\{x_i^{(s)} \in C_m^{(s)}\}} x_i^\dagger\,. \tag{19}$$

Using the WLLN, it can be shown that the $\hat{\pi}_\tau$ for the sample conditional averages are in fact converging to the true underlying distribution's conditional expectations, thus

$$\hat{\pi}_\tau \to \mathbb{E}(x|x \in C_m^{(s)})$$

as $N, S \to \infty$ under the same assumptions as before.
Using this fact, we can replace $x_i^{(s)} \in C_m^{(s)}$ with $\hat{\pi}_\tau$ for each value, thus the working sample becomes the set of replacement estimators for each observation

$$x_{i,Shifting}^{WS} := \{\hat{\pi}_\tau\}.$$

We can also check the standard errors of the replacement estimator to have an idea how precise our results are

$$x_i^\dagger = \boldsymbol{\pi}_\tau \mathbf{1}_{\{x_i^\dagger \in C_m^{(s)}\}} + \eta_i,$$

where $\boldsymbol{\pi}_\tau$ denotes the vector of $\pi_\tau, \forall\tau$. Using the standard OLS technique we can derive

$$\hat{\boldsymbol{\pi}}_\tau = \left(\mathbf{1}'_{\{x_i^\dagger \in C_m^{(s)}\}} \mathbf{1}_{\{x_i^\dagger \in C_m^{(s)}\}}\right)^{-1} \mathbf{1}'_{\{x_i^\dagger \in C_m^{(s)}\}} x_i^\dagger.$$

Under the standard OLS assumption, we can write

$$\sqrt{N^{WS}} \left(\hat{\boldsymbol{\pi}}_\tau - \mathbb{E}\left[\boldsymbol{\pi}_\tau\right]\right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_\tau\right),$$

where $\mathbb{E}\left(\boldsymbol{\pi}_\tau\right) = \mathbb{E}(x|x \in C_m^{(s)}), \forall\tau$. Furthermore, the variance of the OLS estimator is given by

$$\boldsymbol{\Omega}_\tau = V\left(\eta_i\right) \left(\mathbf{1}'_{\{x_i^\dagger \in C_m^{(s)}\}} \mathbf{1}_{\{x_i^\dagger \in C_m^{(s)}\}}\right)^{-1},$$

where $\hat{\pi}_\tau$ represents the first moments of the underlying random variable, thus using $x_{i,Shifting}^{WS}$ for estimation will result in a consistent estimator for β. Algorithm 6 describes how to create in practice a working sample with the shifting method.

4 Monte Carlo Experiments

In this section, we examine the finite sample performance of our split sampling methods through some Monte Carlo simulations. The Data Generating Process (DGP) is

$$y_i = 0.5x_i + \varepsilon_i \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The explanatory variable, x_i, is generated as different distributions. The online Supplementary Materials contains detailed results with uniform, normal, exponential and weibull distributions with different parameter settings. Here we present only some demonstrative results, shown in Table 1.
Overall, the results are consistent with the theoretical findings. Tables 2, 3, 4 and 5 shows the Monte Carlo average bias $(\hat{\beta} - \beta)$, Monte-Carlo average of the absolute biases $(|\hat{\beta} - \beta|)$, the Monte Carlo standard deviation $\left(SD\left[\hat{\beta}\right]\right)$ and the average of the number of effective observations $\left(N^{eff}\right)$. The estimation biases are in general decreasing as the number of observations and the number of split samples increase. The relative performance of the methods depends

on two characteristics of the underlying distribution namely, curvature (or the classes' conditional expectations relative to the choice values, $\mathbb{E}[x \mid x \in C_m]$ and z_m), and the fraction of the probability mass covered by the surveys (or what is the probability that a certain part of the distribution is neglected by the surveys: $\Pr(x < a_l)$ or $\Pr(x > a_u)$).

The Monte Carlo setup allows us to disentangle these two effects as shown in Table 1. The exponential distribution with parameter $\lambda = 0.5$ provides a distribution with flat curvature. Hence, $\mathbb{E}[x \mid x \in C_m]$ and z_m are close to each other and the performance of the two methods are similar to each other. The normal distribution with $\mu_x = 0, \sigma_x^2 = 0.2$ has steeper curvature. Thus, $\mathbb{E}[x \mid x \in C_m]$ and z_m are far from each other. The magnifying method appears to be better than the shifting method in this case. In general, a large M appears to be critical if the distribution has a steep curvature. Furthermore, we have checked the truncated case, where the probability mass is completely covered by the surveys and the censored case, where there is a non-negligible part of the probability mass which cannot be utilized for the estimation. The detailed analysis of each of these cases is in the online Supplementary Materials.

$f(\cdot; a_l, a_u)$	$\mathbb{E}[x \mid x \in C_m]$ and z_m	$\int_{a_l}^{a_u} f(\cdot)$
$Exp(0.5; 0, 1)$	close to each other	complete mapping (100%)
$Exp(0.5; 0, \infty)$		weak mapping (39%)
$\mathcal{N}(0, 0.2; -1, 1)$	far from each other	complete mapping (100%)
$\mathcal{N}(0, 0.2; -\infty, \infty)$		good mapping (99%)

Table 1: Distributions used for the underlying random variable x.

		Magnifying method - used as $x_{i,All}^{WS}$							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta} - \beta$	**N=10,000**	-0.0182	0.0032	0.0020	0.0015	0.1341	-0.0026	-0.1147	-0.2728
	N=100,000	-0.0185	0.0020	0.0012	0.0016	0.1342	-0.0029	-0.0151	-0.0473
	N=500,000	-0.0190	0.0004	0.0004	0.0008	0.1339	-0.0008	-0.0013	-0.0182
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0415	0.0807	0.0902	0.0929	0.1342	0.3105	0.4537	0.5312
	N=100,000	0.0208	0.0284	0.0312	0.0320	0.1339	0.0971	0.1676	0.2264
	N=500,000	0.0191	0.0121	0.0138	0.0140	0.1342	0.0438	0.0760	0.1049
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0489	0.1024	0.1147	0.0445	0.0785	0.3872	0.5653	0.6019
	N=100,000	0.0163	0.0355	0.0392	0.0401	0.0137	0.1218	0.2108	0.2794
	N=500,000	0.0073	0.0152	0.0172	0.0175	0.0061	0.0554	0.0961	0.1301
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	696	212	181
	N=100,000	100,000	100,000	100,000	100,000	100,000	6,874	1,693	941
	N=500,000	500,000	500,000	500,000	500,000	500,000	34,348	8,267	4,310
		Shifting method - used as x_i^{WS}							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta} - \beta$	**N=10,000**	-0.0182	0.0023	0.0025	0.0027	0.1341	0.0864	0.0843	0.0861
	N=100,000	-0.0185	0.0019	0.0021	0.0021	0.1342	0.0859	0.0809	0.0801
	N=500,000	-0.0190	0.0018	0.0017	0.0016	0.1339	0.0865	0.0815	0.0805
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0415	0.0703	0.0701	0.0701	0.1342	0.1811	0.1642	0.1630
	N=100,000	0.0208	0.0238	0.0236	0.0235	0.1339	0.0926	0.0873	0.0869
	N=500,000	0.0191	0.0103	0.0103	0.0103	0.1342	0.0866	0.0816	0.0806
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0489	0.0879	0.0878	0.0879	0.0785	0.2078	0.1891	0.1864
	N=100,000	0.0163	0.0297	0.0294	0.0293	0.0137	0.0683	0.0633	0.0632
	N=500,000	0.0073	0.0130	0.0130	0.0130	0.0061	0.0308	0.0283	0.0280
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	5,071	5,334	5,387
	N=100,000	100,000	100,000	100,000	100,000	100,000	50,704	53,162	53,491
	N=500,000	500,000	500,000	500,000	500,000	500,000	253,492	265,711	267,270

*BM = Benchmark. See Table 4. in the online Supplementary Materials.

Table 2: Monte Carlo statistics for $x_i \sim Exp(0.5)$, M=3

		Magnifying method - used as $x_{i,ALL}^{WS}$							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta}-\beta$	**N=10,000**	-0.0798	-0.0051	-0.0015	-0.0005	-0.0552	-0.0053	-0.1419	-0.3182
	N=100,000	-0.0800	-0.0055	-0.0002	-0.0003	-0.0552	-0.0053	-0.0188	-0.0751
	N=500,000	-0.0803	-0.0057	-0.0002	0.0000	-0.0554	-0.0054	-0.0037	-0.0160
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0798	0.0264	0.0318	0.0356	0.0553	0.0669	0.1699	0.3198
	N=100,000	0.0800	0.0100	0.0109	0.0120	0.0552	0.0226	0.0461	0.0863
	N=500,000	0.0803	0.0063	0.0050	0.0054	0.0554	0.0104	0.0194	0.0301
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0224	0.0329	0.0401	0.0447	0.0220	0.0842	0.1534	0.1485
	N=100,000	0.0074	0.0111	0.0136	0.0151	0.0074	0.0282	0.0540	0.0721
	N=500,000	0.0033	0.0051	0.0063	0.0068	0.0031	0.0117	0.0241	0.0349
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	946	241	195
	N=100,000	100,000	100,000	100,000	100,000	100,000	9,381	1,983	1,069
	N=500,000	500,000	500,000	500,000	500,000	500,000	46,891	9,730	4,953
		Shifting method							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta}-\beta$	**N=10,000**	-0.0811	-0.0244	-0.0240	-0.0242	-0.0552	0.0106	0.0067	0.0049
	N=100,000	-0.0810	-0.0246	-0.0241	-0.0241	-0.0552	0.0103	0.0069	0.0062
	N=500,000	-0.0811	-0.0246	-0.0242	-0.0242	-0.0554	0.0102	0.0071	0.0065
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0811	0.0288	0.0285	0.0286	0.0553	0.0346	0.0323	0.0316
	N=100,000	0.0810	0.0246	0.0241	0.0241	0.0552	0.0137	0.0115	0.0112
	N=500,000	0.0811	0.0246	0.0242	0.0242	0.0554	0.0104	0.0076	0.0072
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0224	0.0251	0.0253	0.0253	0.0220	0.0421	0.0401	0.0395
	N=100,000	0.0071	0.0083	0.0083	0.0082	0.0074	0.0134	0.0127	0.0126
	N=500,000	0.0033	0.0036	0.0037	0.0037	0.0031	0.0059	0.0056	0.0055
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	8,064	8,250	8,280
	N=100,000	100,000	100,000	100,000	100,000	100,000	80,631	82,428	82,654
	N=500,000	500,000	500,000	500,000	500,000	500,000	403,167	412,108	413,203

*BM = Benchmark. See Table 5. in the online Supplementary Materials.

Table 3: Monte Carlo statistics for $x_i \sim \mathcal{N}(0, 0.2)$, M=3

		Magnifying method - used as $x_{i,All}^{WS}$							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta}-\beta$	**N=10,000**	-0.0074	0.0037	0.0004	-0.0002	0.1304	-0.0063	-0.1343	-0.2709
	N=100,000	-0.0072	0.0014	0.0013	0.0012	0.1307	-0.0038	-0.0156	-0.0494
	N=500,000	-0.0078	0.0005	0.0006	0.0007	0.1303	-0.0011	-0.0033	-0.0099
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0394	0.0841	0.0908	0.0919	0.1305	0.2472	0.4654	0.5010
	N=100,000	0.0145	0.0291	0.0314	0.0325	0.1307	0.0809	0.1599	0.2147
	N=500,000	0.0090	0.0124	0.0138	0.0140	0.1303	0.0365	0.0746	0.1027
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0489	0.1068	0.1155	0.1164	0.0437	0.3081	0.5710	0.5767
	N=100,000	0.0165	0.0364	0.0395	0.0405	0.0135	0.1025	0.2048	0.2647
	N=500,000	0.0073	0.0155	0.0173	0.0177	0.0059	0.0458	0.0938	0.1278
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	804	218	183
	N=100,000	100,000	100,000	100,000	100,000	100,000	7,962	1,750	955
	N=500,000	500,000	500,000	500,000	500,000	500,000	39,775	8,547	4,383
		Shifting method - used as x_i^{WS}							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta}-\beta$	**N=10,000**	-0.0074	0.0020	0.0018	0.0016	0.1304	0.0269	0.0292	0.0311
	N=100,000	-0.0072	0.0016	0.0015	0.0015	0.1307	0.0218	0.0209	0.0212
	N=500,000	-0.0078	0.0007	0.0006	0.0006	0.1303	0.0221	0.0218	0.0218
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0489	0.0674	0.0678	0.0680	0.1305	0.1112	0.1057	0.1053
	N=100,000	0.0165	0.0230	0.0230	0.0230	0.1307	0.0394	0.0380	0.0381
	N=500,000	0.0073	0.0099	0.0099	0.0099	0.1303	0.0247	0.0243	0.0243
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0843	0.0837	0.0844	0.0846	0.0437	0.1359	0.1294	0.1280
	N=100,000	0.0279	0.0285	0.0286	0.0285	0.0135	0.0444	0.0425	0.0425
	N=500,000	0.0131	0.0124	0.0124	0.0124	0.0059	0.0200	0.0195	0.0193
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	6,379	6,552	6,589
	N=100,000	100,000	100,000	100,000	100,000	100,000	63,814	65,404	65,619
	N=500,000	500,000	500,000	500,000	500,000	500,000	319,061	326,956	327,963

*BM = Benchmark. See Table 4. in the online Supplementary Materials.

Table 4: Monte Carlo statistics for $x_i \sim Exp(0.5)$, M=5

		Magnifying method - used as $x_{i,All}^{WS}$							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta} - \beta$	**N=10,000**	-0.0311	-0.0005	-0.0002	-0.0004	-0.0097	-0.0044	-0.1536	-0.3267
	N=100,000	-0.0313	-0.0004	0.0001	0.0000	-0.0099	-0.0010	-0.0178	-0.0794
	N=500,000	-0.0315	-0.0008	-0.0000	0.0000	-0.0100	-0.0010	-0.0039	-0.0164
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0328	0.0274	0.0324	0.0368	0.0195	0.0649	0.1780	0.3282
	N=100,000	0.0313	0.0093	0.0111	0.0121	0.0106	0.0204	0.0460	0.0901
	N=500,000	0.0315	0.0042	0.0050	0.0054	0.0100	0.0095	0.0200	0.0306
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0226	0.0343	0.0404	0.0461	0.0234	0.0815	0.1523	0.1495
	N=100,000	0.0078	0.0116	0.0139	0.0152	0.0074	0.0257	0.0544	0.0747
	N=500,000	0.0033	0.0052	0.0063	0.0068	0.0033	0.0118	0.0243	0.0346
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	973	243	196
	N=100,000	100,000	100,000	100,000	100,000	100,000	9,643	1,994	1,072
	N=500,000	500,000	500,000	500,000	500,000	500,000	48,204	9,771	4,965
		Shifting method - used as x_i^{WS}							
		Truncated				Censored			
		BM*	**S=10**	**S=50**	**S=100**	**BM***	**S=10**	**S=50**	**S=100**
$\hat{\beta} - \beta$	**N=10,000**	-0.0311	-0.0052	-0.0050	-0.0052	-0.0097	0.0049	0.0038	0.0034
	N=100,000	-0.0313	-0.0049	-0.0047	-0.0047	-0.0099	0.0054	0.0038	0.0035
	N=500,000	-0.0315	-0.0052	-0.0049	-0.0049	-0.0100	0.0053	0.0037	0.0035
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0328	0.0198	0.0198	0.0197	0.0195	0.0248	0.0241	0.0240
	N=100,000	0.0313	0.0077	0.0076	0.0076	0.0106	0.0089	0.0082	0.0081
	N=500,000	0.0315	0.0054	0.0052	0.0052	0.0100	0.0058	0.0046	0.0045
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0226	0.0242	0.0243	0.0243	0.0234	0.0307	0.0300	0.0299
	N=100,000	0.0078	0.0082	0.0083	0.0083	0.0074	0.0098	0.0095	0.0095
	N=500,000	0.0033	0.0036	0.0036	0.0036	0.0033	0.0044	0.0043	0.0043
N^{eff}	**N=10,000**	10,000	10,000	10,000	10,000	10,000	9,089	9,156	9,168
	N=100,000	100,000	100,000	100,000	100,000	100,000	90,884	91,525	91,606
	N=500,000	500,000	500,000	500,000	500,000	500,000	454,421	457,602	457,994

*BM = Benchmark. See Table 5. in the online Supplementary Materials.

Table 5: Monte Carlo statistics for $x_i \sim \mathcal{N}(0, 0.2)$, M=5

5 Extensions

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)). Let us call this the *perception effect*. The presence of this effect is independent of the implementation of the two sub-sampling methods. However, with sub-sampling, there is a way to tackle this issue, much akin to a familiar approach in the panel data literature.

More specifically, the definition of classes may affect participants' responses to the survey question. A way to formalize such effects is by redefining the discretization of x_i as follows

$$x_i^{**} = \begin{cases} z_1 & \text{if} \quad c_0 < x_i + B_s < c_1 \\ & \vdots \\ z_m & \text{if} \quad c_{m-1} < x_i + B_s < c_M, \end{cases} \tag{20}$$

where B_s denotes the perception effect for split sample s, $s = 1, \ldots, S$. Let $\tilde{x}_i^*$ and $\tilde{x}_i^{**}$ denote the observations in the working sample that derived from x_i^* and x_i^{**}, respectively. Following the derivation of the working sample from the methods above, all observations in the working samples can be expressed as

$$\tilde{x}_i^{**} = \tilde{x}_i^* + B_s \tag{21}$$

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

$$y_i = \beta \tilde{x}_i^{**} + u_i \tag{22}$$

is equivalent to

$$y_i = \beta \tilde{x}_i^* + B_s \beta + u_i. \tag{23}$$

Rewrite the above in matrix form using standard definitions gives

$$\mathbf{y} = \tilde{\mathbf{x}}^* \beta + \mathbf{DB}\beta + \mathbf{u}, \tag{24}$$

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 $\mathbf{B}$. Thus, 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

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

is a consistent estimator of β following the similar argument for the standard fixed effect estimator in the panel data literature.
We also need to slightly modify the replacement estimator in order for the above to hold. The main problem is to keep track of the perception effects. This means we need to keep track of which split sample each observation comes from when estimating the conditional averages. This means

$$\hat{\pi}_{\chi,s} = \left(\sum_{i=1}^{N}\mathbf{1}_{\{\tilde{x}_i^{**}\in C_\chi, \tilde{x}_i^{**}\in s\}}\right)^{-1}\sum_{i=1}^{N}\mathbf{1}_{\{\tilde{x}_i^{**}\in C_\chi, \tilde{x}_i^{**}\in s\}}\tilde{x}_i^{**} \tag{26}$$

and as $N \to \infty$

$$\hat{\pi}_{\chi,s} = \mathbb{E}(x_i | x_i \in C_\chi) + B_s + o_p(1).$$

This shows that equation (25) provides a valid replacement estimator in the presence of perception effects.
While the discussion above focuses on the case with one regressor, the generalisation to K regressors is straightforward. Perhaps a more interesting question is the presence of perception effects over different m. In principle, this can also be incorporated by replacing B_s with B_{sm} for $s = 1, \ldots, S$ and $m = 1, \ldots, M$. Therefore, this particular setup does not just allow for perception effects due to different split samples, but rather, it provides a framework to investigate different types of perception effects. This would be an interesting avenue of future research in this area.

Another possible extension is to consider the application of the proposed methods in the context of non-linear models. So far the discussion has focused on the linear model as defined in equation (2). Given the presented methods focus on data collection, they could also be applied for non-linear model. To see this, consider

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

where $h(\cdot)$ denotes a continuous function. Let $\mathbf{x}$ be the data matrix of x_i and $\hat{\beta}(\mathbf{x})$ denotes a consistent estimator of β with $\rho(\mathbf{x}) = \sqrt{N}\left[\hat{\beta}(\mathbf{x}) - \beta\right]$ such that $\rho(\mathbf{x}) \xrightarrow{d} D(0, \Omega)$. Under the assumptions made earlier, $x_i^* \xrightarrow{d} x_i$ and therefore $\rho(\mathbf{x}^*) \xrightarrow{d} \rho(\mathbf{x})$ by the continuous mapping theorem under appropriate regularity conditions. The technical details of these conditions, however, could be an interesting subject of future research.

6 Conclusion

This paper has investigated the effects of using interval data in a linear regression model when the underlying discretized continuous variable is not observed. This situation often arises in survey data when such variables – like income – are not captured directly, but rather, are replaced by a set of m choices. Unlike other studies in the literature, our approach has considered the more realistic case when the underlying distribution of the unobserved explanatory variables is unknown and the values of each choice can be arbitrarily assigned. With fixed

m, the results show that using the discretized ordered choices as explanatory variables in a linear regression will lead to biased and inconsistent parameter estimates. The well-known techniques to create consistent estimators require information from the distributions of the underlying explanatory variables, which are presumed to be unknown, and therefore cannot be applied here.

This paper proposes a novel data gathering method that we called split sampling. Using the fact that the discretized variables approach their unobserved continuous counterparts when m grows, the proposed approach essentially replaces the requirement of m being sufficiently large with the more standard scenario where the number of individuals, N is very large, utilizing different questionnaires for each split sample. Theoretical results show that these techniques will lead to a proper mapping of the true underlying distribution. Monte Carlo simulations show that the proposed methods work reasonably well, and may have significant implications for the future of survey design.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *The Quarterly Journal of Economics*, 117(4):1231–1294.

Berkson, J. (1980). Minimum chi-square, not maximum mikelihood! *The Annals of Statistics*, 8:457–487.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC Press.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253.

Connor, R. J. (1972). Grouping for testing trends in categorical data. *Journal of the American Statistical Association*, 67(339):601–604.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, 52(280):543–547.

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

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

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

Hsiao, C. (1983). Regression analysis with a categorized explanatory variable. In Karlin, S., Amemiya, T., and Goodman, A. L., editors, *Studies in Econometrics, Time Series, and Multivariate Statistics*, chapter 5, pages 93–129. Academic Press.

Johnson, D. R. and Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, pages 398–407.

Knack, S. and Keefer, P. (1995). Institutions and economic performance: cross-country tests using alternative institutional measures. *Economics & Politics*, 7(3):207–227.

Lagakos, S. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine*, 7(1-2):257–274.

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

Mauro, P. (1995). Corruption and growth. *The Quarterly Journal of Economics*, 110(3):681–712.

Méndez, F. and Sepúlveda, F. (2006). Corruption, growth and political regimes: Cross country evidence. *European Journal of Political Economy*, 22(1):82–98.

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., and Liss, S. (2011). Summary of travel trends: 2009 national household travel survey. Technical report, United States Department of Transportation.

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.

Taylor, J. M. and Yu, M. (2002). Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*, 83(1):248–263.

Wansbeek, T. and Meijer, E. (2000). *Measurement Error and Latent Variables in Econometrics*. North-Holland Elsevier.

Wansbeek, T. and Meijer, E. (2001). Measurement error and latent variables. In Baltagi, B. H., editor, *A Companion to Theoretical Econometrics*, chapter 8, pages 162–179. John Wiley & Sons.