Modelling with Discretized Ordered Choice Covariates

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

[1]Curtin University
[2]Central European University

May 18, 2021

Email address: reguly_agoston@phd.ceu.edu.

Abstract

The 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 covariates. This situation is common when modelling with survey data where continuous random variables, such as income or expenditure, are being transformed into 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 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 non-linear models.

JEL: C01, C13, C21, C25, C83

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

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

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 (e.g., 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 e.g., 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 $[a_l, a_u]$, where $a_l, a_u \in \mathbb{R}, a_l < a_u$ and realizations $i = 1, \ldots, N$. Furthermore, define the discretized variable as

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

We refer to 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 an arbitrarily assigned value in C_m. x_i^* is considered as an interval data, as the class boundaries are known by the researchers. The main difficulty is that x_i is not directly observable, in fact only the response x_i^* is observed. 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

[1] A complete list of the notations used in the paper is given in Appendix D.

models using x_i^* instead of x_i as the latter is not available.[2] Manski and Tamer (2002) show 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 Least Squares (LS) estimator is inconsistent in general and can only be consistent in a few very specific and restrictive cases.

More importantly, this paper proposes a new data gathering technique, called *split sampling*,[3] which can map 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 LS 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 of the parameters in the linear regression model using discretized data. The finite sample performance of these techniques is analysed in Section 4. Section 5 discusses some possible extensions of the techniques and concluding remarks are made in Section 6. All technical proofs and additional Monte Carlo results can be found in the Appendix.

2 Motivation

Consider the following data generating process

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

[2]Let us note here, that an other common practice is to create $M-1$ dummy variable for each different choices and use these variables instead of x_i^*. This solution is feasible if the parameter of interest is not a measure such as the elasticity for this discretized variable. In this paper we focus on these cases, thus using dummy variables is not a solution.

[3]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.

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 the idiosyncratic disturbance term of model (2) and $\varepsilon_i = (x_i - x_i^*)'\beta + u_i$ denotes the disturbance term of model (3), while γ and β are unknown parameter vectors. 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 daily average use of public transport in commuting measured in minutes. UPT_i is not observed directly, but we observe only the individual's choice from a pre-set list UPT_i^* via a questionnaire. We ask the use of public transport in the following way

$$\begin{array}{ll} a) & \text{between 1 and 2 hours,} \\ b) & \text{between 30 minutes and 1 hour,} \\ c) & \text{between 15 and 30 minutes,} \\ d) & \text{between 5 and 15 minutes,} \\ e) & \text{less than 5 minutes,} \end{array} \tag{5}$$

For simplicity, let us neglect the possible answer of 'more than 2 hours' travelling time, but we will come back to this issue at Section 3. These choice options can be converted to given intervals and let us set the choice values (z_m) as the mid-value for each class. The discretized variable UPT_i^* has the following form

$$UPT_i^* = \begin{cases} 90\,, & \text{if} \quad 60 \le UPT_i \le 120, & \longleftarrow & \text{a) between 1 and 2 hours} \\ 45\,, & \text{if} \quad 30 \le UPT_i \le 60, & \longleftarrow & \text{b) between 30 minutes and 1 hour} \\ 22.5\,, & \text{if} \quad 15 \le UPT_i \le 30, & \longleftarrow & \text{c) between 15 and 30 minutes} \\ 10\,, & \text{if} \quad 5 \le UPT_i \le 15, & \longleftarrow & \text{d) between 5 and 15 minutes} \\ 2.5\,, & \text{if} \quad 0 \le UPT_i \le 5, & \longleftarrow & \text{e) less than 5 minutes} \end{cases} \tag{6}$$

Obviously, we can use many possible values for the z_m. 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), Terza (1987) 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 LS is inconsistent in general when assigning z_m using the mid-point values.[4] In a seminal paper, Manski and Tamer (2002) extend this 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 LS estimator on model (3) as it is inconsistent. In fact, it can be shown that with one regressor[5]

$$\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^{*}_{LS}-\beta\right)=\frac{\beta\sum_{m=1}^{M}z_m\left\{\mathbb{E}(x_i|x_i\in C_m)-z_m\right\}}{\sum_{m=1}^{M}z_m^2}. \tag{7}$$

Equation (7) is insightful for two reasons. First, the right-hand side is generally not zero which shows that LS is inconsistent in general. Second, the right-hand

[4] As a solution Hsiao (1983) offers an iterative maximum likelihood method to estimate β, using a strong distributional assumption for point-identification. Terza (1987) improves the method of Hsiao (1983), with two-stage maximum likelihood method, still requiring assumption on the distribution.

[5] Detailed derivations and in-depth analysis on the consistency of the LS estimator for Equation (2) and Model (3) can be found in Appendix B

side can be zero 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 LS 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 (7).

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 LS estimator, do not seem to be appropriate even when partial identification is possible.

In the next section, we introduce the split sampling approach that can provide a solution to these identification and estimation issues.

3 Split sampling

Since there is a limit on identification given the data, one 'natural' 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[6]. The set of respondents

[6]In the case of cross sectional data. For panel data one shall assign different questionnaires for

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.[7] Note that the number of observations across split samples can be the same or, more likely, different. Now a split sample looks like,

$$x_i^{(s)} = \begin{cases} z_1^{(s)} & \text{if } x_i \in C_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ & \qquad \text{1st choice for split sample } s, \\ z_2^{(s)} & \text{if } x_i \in C_2^{(s)} = [c_1^{(s)}, c_2^{(s)}), \\ \vdots & \vdots \\ z_m^{(s)} & \text{if } x_i \in C_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ z_M^{(s)} & \text{if } x_i \in C_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last choice for split sample } s. \end{cases} \tag{8}$$

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.

each individual through time.

[7] 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.

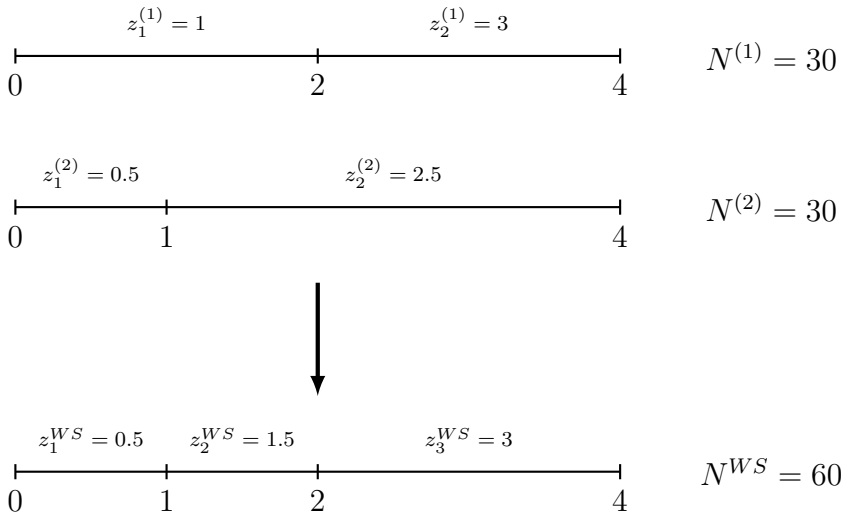

Figure 1: The basic idea of split sampling

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. The class boundaries in the working sample are constructed by 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'.

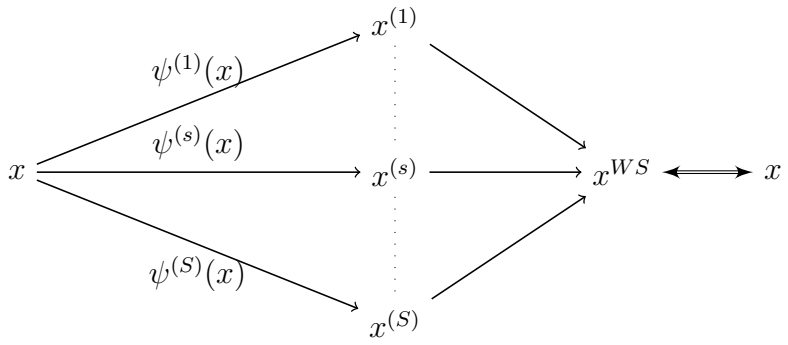

Figure 2: Creation of the working sample's random variable

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

$$\lim_{S \to \infty} \mathbb{E}_S \left[x^{WS} | y \right] = \mathbb{E} \left[x | y \right] , \tag{9}$$

which means that in the limit the conditional expectation of the working sample should be the same as for the true underlying variable.

3.2 Probabilities in the Working Sample

To show later on that Equation 9. holds for the introduced methods, we need to calculate the probability of an observation to fall into a working sample class. 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. Based on these, 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: $\Pr(x \in \mathcal{S}_s)$, where $\mathcal{S}_s$ denotes the s-th split sample. Uniformly assigning these individuals to split samples is the most straightforward procedure, ($\Pr(x \in \mathcal{S}_s) = 1/S$), 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,

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

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 any systematic bias during the merging process. [8]

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

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

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

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

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

of each of the 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 $M = 3, S = 4$.

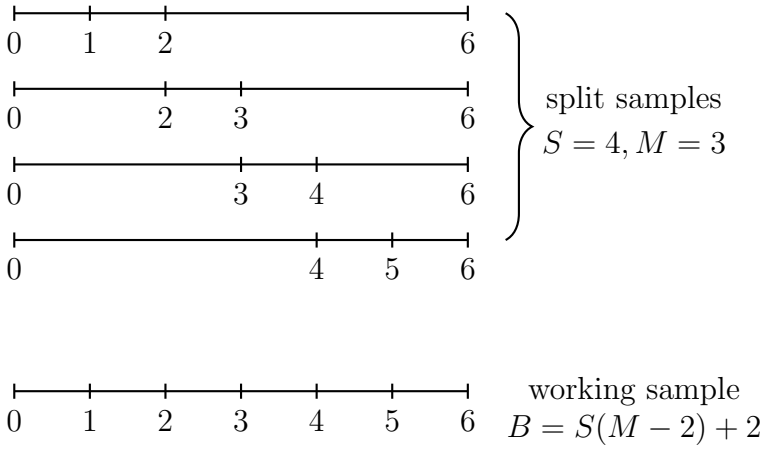

Figure 3: The magnifying method

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

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

Algorithm 1 Magnifying method – creation of the split samples ($\psi^{(s)}(\cdot)$)

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 goto Step 2.

The magnifying method works as it converges to the unknown distribution of x as 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.[9]

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

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

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

Note that $||C_m^{(s)}|| \leq ||C_1^{(s)}||$ and $||C_m^{(s)}|| \leq ||C_M^{(s)}||$. Formally, let us define $\zeta := \{C_m^{(s)} \mid 1 < m < M, 1 < s < S, C_m^{(1)} \mid 1 \leq m < M, C_m^{(S)} \mid 1 < m \leq M\}$ as the set of classes which have the class width $\frac{a_u - a_l}{S(M-2)+2}$. Then we can write $\Pr\left((x - x^{(s)})^2 | x \in \zeta \leq (x - x^{(s)})^2 \mid x \notin \zeta\right) = 1$, 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

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

and for the first and last split sample

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

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

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.

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

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}.[10]
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 describes how to construct the working sample, when using only the directly transferable observations.

Algorithm 2 Magnifying method - creation of the 'DTO' working sample ($\Psi_{DTO}(\cdot)$)

1: Set $m = 1, s = 1$ and $x_{i,DTO}^{WS}, y_{i,DTO}^{WS}, w_{i,DTO}^{WS} = \emptyset$.
2: If $C_m^{(s)} \in \zeta$, add observations from class $C_m^{(s)}$ to the working sample:

$$x_{i.DTO}^{WS} := \left\{ x_{i,DTO}^{WS}, \bigcup_{j=1}^{N} \left(x_j^{(s)} \in C_m^{(s)} \right) \right\},$$
$$y_{i.DTO}^{WS} := \left\{ y_{i,DTO}^{WS}, \bigcup_{j=1}^{N} y_j^{(s)} \mid \left(x_i^{(s)} \in C_m^{(s)} \right) \right\},$$
$$w_{i.DTO}^{WS} := \left\{ w_{i,DTO}^{WS}, \bigcup_{j=1}^{N} w_j^{(s)} \mid \left(x_i^{(s)} \in C_m^{(s)} \right) \right\},$$

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.

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

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

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

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

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

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

Assumption 1. *Let x be a continuous random variable with probability density function $f(x)$ with S, N and $C_m^{(s)}$ follow the definitions above then*

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

b. All split samples will have non-zero respondents. $(\Pr(x \in \mathcal{S}_s) > 0)$

c. $\int_a^b f(y)dy > 0$ *for any* $(a,b) \subset [a_l, a_u]$.

Assumption 1a. ensures that the number of respondents will always be higher than the number of split samples. Assumption 1b. ensures utilisation of all split samples, i.e. each split sample will have non-zero respondents. Assumption 1c. 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 1a - c,*

1.
$$\mathbb{E}(N_b^{WS}) > 0$$

2.
$$\Pr\left(\sum_{i=1}^{b} N_b^{WS} > 0\right) \to 1.$$

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

See the proof in Appendix C.1.

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 LS 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 c ≥ 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 $x^{WS}_{i,DTO}$). However, it seems that too many could fall into this category, resulting in a large efficiency loss in estimation.

Another approach is to use 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 $x^{WS}_{i,ALL}$. 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 $x^{(s)}_i \in C_\chi$, where

$$C_\chi := \bigcup_{s,m} C^{(s)}_m \bigcap_b C^{WS}_b = \zeta^{\complement}$$

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

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

Let us introduce $x^{WS}_{i,NDTO}$ as the variable which contains all the replaced values with $\hat{\pi}_\chi$, $\forall x^{(s)}_i \in C_\chi$. 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 $N, S \to \infty$ and under the same assumptions

as before. This also implies $x^{WS}_{i,NDTO} \to \mathbb{E}(x|x \in C_\chi)$, which means using working sample $x^{WS}_{i,ALL}$ satisfies Equation 9.

Algorithm 3 The magnifying method - creation of 'ALL' working sample ($\Psi_{ALL}(\cdot)$)

1: Let, $x^{WS}_{i,ALL} := \{x^{WS}_{i,DTO}\}$, $y^{WS}_{i,ALL} := \{y^{WS}_{i,DTO}\}$, $w^{WS}_{i,ALL} := \{w^{WS}_{i,DTO}\}$
2: Set, $m = 1, s = 1$
3: If $C^{(s)}_m \in C_\chi$, then calculate $\hat{\pi}_\chi$ and expand the working sample as,

$$x^{WS}_{i.ALL} := \left\{ x^{WS}_{i,ALL}, \bigcup_{j=1}^{N} \hat{\pi}_\chi \mid \left(x^{(s)}_j \in C^{(s)}_m \right) \right\},$$

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

$$w^{WS}_{i.ALL} := \left\{ x^{WS}_{i,ALL}, \bigcup_{j=1}^{N} w^{(s)}_j \mid \left(x^{(s)}_i \in C^{(s)}_m \right) \right\},$$

4: If $s < S$, then $s := s + 1$ and go to Step 3.
5: If $s = S$, then $s := 1$ and set $m = m + 1$ and go to Step 3.

We can obtain the asymptotic standard errors of this estimator, as 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 LS 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.[11] 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 LS 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 LS assumptions, we can write

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

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

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

The variance of the LS estimator is

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

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

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 $a_l = -\infty$ or/and C_B^{WS} if $a_u = \infty$. This solution means we artificially truncate $y \to y^{tr}, x \to x^{tr}$ and $w \to w^{tr}$. For the truncated distribution, we can use all the derivations presented above, and we end up with convergence in distribution. That is, $f(x^{WS}_{DTO} \in \zeta^{tr}) \xrightarrow{d} f(x^{tr})$.[12] Furthermore, the parameter estimates $\beta^{tr} = \beta$ (under some reasonable assumptions), which implies that the LS 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.

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.

[12] 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.

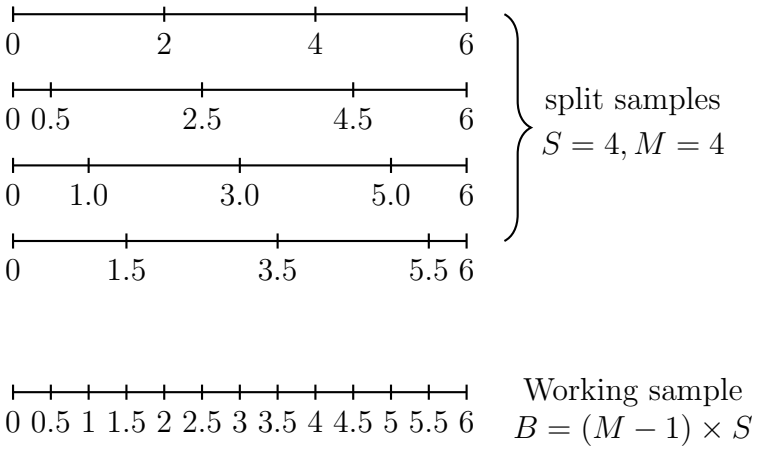

Figure 4: The shifting method

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 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 goto Step 2.

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

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

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 $C_m^{(s)}$,

$$Z(s,m) = \begin{cases} \{\emptyset\}, & \text{if, } s=1 \text{ and } m=1, \\ \bigcup_{b=1}^{s-1}\{z_b^{WS}\}, & \text{if, } s\neq 1 \text{ and } m=1, \\ \bigcup_{b=s-1+(m-1)(M-1)}^{s-1+(m-1)(M-1)}\{z_b^{WS}\}, & \text{if, } 1<m<M, \\ \bigcup_{b=B-S+s-1}^{B}\{z_b^{WS}\}, & \text{if } m=M. \end{cases} \tag{14}$$

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 $x_i^{\dagger}$.
The assignment mechanism can be written as

$$x_i^{\dagger}|x_i^{(s)} \in C_m^{(s)} = z \in Z(s,m), \text{ with } \begin{cases} \Pr(1), & \text{if } s=1 \text{ and } m=1, \\ \Pr(1/(s-1)), & \text{if } s\neq 1 \text{ and } m=1, \\ \Pr(1/S), & \text{if } 1<m<M, \text{ or} \\ \Pr(1/(S-s+1)), & \text{if } m=M. \end{cases} \tag{15}$$

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 $\Pr(x \in \mathcal{S}_s) = 1/S$,

$$\Pr\left(x_i^{\dagger} \in C_b^{WS}\right) = \begin{cases} 0, & \text{if } s=1 \text{ and } m=1, \\ \frac{1}{S}\sum_{s=2}^{S}\frac{1}{s-1}\int_{C_1^{(s)}|C_b^{WS}\in C_1^{(s)}} f(x)\mathrm{d}x, & \text{if } s\neq 1 \text{ and } m=1, \\ \frac{1}{S^2}\sum_{s=1}^{S}\int_{C_m^{(s)}|C_b^{WS}\in C_m^{(s)}} f(x)\mathrm{d}x, & \text{if } 1<m<M, \\ \frac{1}{S}\sum_{s=1}^{S}\frac{1}{S-s+1}\int_{C_M^{(s)}|C_b^{WS}\in C_M^{(s)}} f(x)\mathrm{d}x, & \text{if } m=M\,. \end{cases} \tag{16}$$

Algorithm 5 describes how to create an artificial variable which approximates the underlying distribution of x.

Algorithm 5 The shifting method – creation of artificial variable $(x_i^\dagger)$

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

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

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

where $\mathcal{A}_m^{(s)}$ has $N_m^{(s)}$ number of observations.

3: Create $Z(s,m)$, the set of possible working sample choice values,

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

4: Draw $\mathcal{Z}_j \in Z(s,m), j = 1, \ldots, N_m^{(s)}$, with uniform probabilities given by

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

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

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

$$x_i^\dagger := \left\{ x_i^\dagger, \bigcup_{j=1}^{N_m^{(s)}} \mathcal{Z}_j \right\}$$

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.

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 1a, 1c and* $\Pr(x \in \mathcal{S}_s) = 1/S$,

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

or

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

where

$$F_S(a) = \Pr(x^\dagger < a) \text{ and } F(a) = \Pr(x < a)$$

See the proof in the Appendix C.2.

In addition, 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[13], the method converges slower, with $\frac{\log S}{S}$, while for the rest it converges with $1/S$. See the derivations in Appendix C.3.

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

$$\hat{\pi}_\tau := \left(\sum_{i=1}^{N} \mathbf{1}'_{\{x_i^{(s)} \in C_m^{(s)}\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}'_{\{x_i^{(s)} \in C_m^{(s)}\}} x_i^\dagger \,. \tag{17}$$

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

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

[13]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: $c_M^{(1)} + (c_{M-1}^{(1)} - c_M^{(1)})$.

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 LS 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 LS 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 LS 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 Th shifting method – creation of working sample ($\Psi_{Shifting}(\cdot)$)

1: Set $s := 1, m := 1, \{x_i^{WS}, y_i^{WS}, w_i^{WS}\} = \emptyset$.

2: Calculate the sample conditional mean $\hat{\pi}_{\tau}$, for the given $C_m^{(s)}$ class, using

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

3: Add the conditional mean $\hat{\pi}_{\tau}$ and the observed values $y_j^{(s)}, w_j^{(s)}$ to the working sample,

$$x_i^{WS} := \left\{ x_i^{WS}, \bigcup_{j=1}^{N} \hat{\pi}_{\tau} \mid \left(x_j \in C_m^{(s)} \right) \right\}$$
$$y_i^{WS} := \left\{ y_i^{WS}, \bigcup_{j=1}^{N} y_j^{(s)} \mid \left(x_j \in C_m^{(s)} \right) \right\}$$
$$w_i^{WS} := \left\{ w_i^{WS}, \bigcup_{j=1}^{N} w_j^{(s)} \mid \left(x_j \in C_m^{(s)} \right) \right\}.$$

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.

4 Monte Carlo Experiments

In this section, we examine the finite sample performance of our split sampling methods through some Monte Carlo simulations. We use the following data generating process (DGP)

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

The explanatory variable, x_i, is generated as different distributions. Appendix A contains detailed results with uniform, normal, exponential and weibull distributions with different parameter settings. Here we present only some demonstrative results with specifications shown in Table 1.

$f(\cdot; a_l, a_u)$	$\mathbb{E}[x \mid x \in C_m]$ 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.

Overall, the results are consistent with the theoretical findings. Tables 2, 3, 4 and 5 shows the Monte Carlo average of the biases $(\hat{\beta} - \beta)$, the 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 bias is 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.

		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: simple mid-values are used. For complete specification see Table 9 in Appendix A..

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: simple mid-values are used. For complete specification see Table 10 in Appendix A..

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: simple mid-values are used. For complete specification see Table 9 in Appendix A.

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: simple mid-values are used. For complete specification see Table 10 in Appendix A.

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

Let us summarize the results and conclusions from the Monte Carlo exercise.

- **Magnifying method – Truncated case**
  - $Exp(0.5; 0, 1)$: The bias decreases in S and N. The increase of M has no significant effect, because the conditional expected values and choice val-

ues are close to each other. The standard errors are decreasing in N, but slightly increasing in S. This is due to the fact that the share of directly transferable observations is decreasing in S. This implies more NDTOs, which increases the standard errors of the estimated coefficient. The absolute bias therefore first decreases, then starts to increase as the effect of standard errors starts to dominate. *Overall, with flat curvature and complete mapping of the probability mass, S/N should be above 0.01%, and M can be small.*

- $\mathcal{N}(0, 0.2; -1, 1)$: The bias decreases in S and N. There is a significant decrease in the bias if we increase M, because the conditional expected values and choice values are not close to each other. All other results are the same as in the exponential case above. *Overall, with steep curvature and complete mapping of probability mass, S/N should be above 0.01%, and increasing M can significantly reduce the bias.*

- **Magnifying method – Censored case**

  - $Exp(0.5; 0, \infty)$ and $\mathcal{N}(0, 0.2; -\infty, \infty)$: The bias first decreases, but then it starts to increase again. This is due to the fact there are only a few observations to calculate the replacement estimator values for non-directly transferable observations. This lack of precision introduces bias during the estimation of β. The number of observations is radically decreasing as S increases and the standard errors are increasing in S. The absolute bias is mainly driven by the standard errors. *Overall, without complete mapping of the probability mass, the main driver of the bias is the number of observations in the working sample. With fewer split samples, we can decrease the absolute bias, but using too many split samples is counterproductive. $S/N > 0.01\%$ is a good rule of thumb here as well.*

- **Shifting method – Truncated case**

  - $Exp(0.5; 0, 1)$: The bias decreases in S and N. Using larger S will not help reduce the bias on the same scale as in the magnifying method due to the boundary classes' slow convergence. On the other hand, using more choices (M) will reduce the bias. It is interesting to note that the standard errors remain unchanged as S increases. The absolute bias decreases and gets smaller than in the benchmark case (with no split sampling) if we have a large amount of observations. *Overall, with complete mapping of the probability mass and flat curvature distribution, increasing M helps to reduce the bias, and increasing S also decreases it, but at a much slower rate. We need a large amount of observations in order to reduce the standard errors as well. As a rule of thumb we may use a smaller number of split samples.*

- $\mathcal{N}(0, 0.2; -1, 1)$: The bias decreases in S and N. Using larger S helps to significantly reduce the bias similarly to using larger M. This makes the approximation much better at the boundaries. Standard errors are the same as in the benchmark case, and does not change as S or M increases. The absolute bias is decreasing in N and S. *Overall, with complete mapping of the probability mass and steep curvature distribution, increasing M and S helps to reduce the bias more effectively. The absolute bias is also decreasing in N, M and S.*

- **Shifting method – Censored case**

  - $Exp(0.5; 0, \infty)$: The bias is decreasing in N and S, but it decreases more slowly in S, because the main drivers of the bias are the boundary classes. Increasing M will help to significantly reduce the bias. The standard errors and the absolute bias behave similarly as in the truncated case. Note that the number of observations used for the estimation is much larger than in the magnifying case! *Overall, without complete mapping of the probability mass, with flat curvature distribution, using few split samples will eliminate the main bias, and increasing M can help to reduce it even more.*
  - $\mathcal{N}(0, 0.2; -\infty, \infty)$: The bias is decreasing in N and S. Now, the boundary classes only take up a small fraction of the probability mass of the distribution, so these classes have a much smaller role in driving the bias, resulting in a much faster bias reduction. Furthermore, increasing the number of choices decreases the bias further. The standard errors, however, are slightly larger than in the benchmark case. The absolute bias is decreasing in N, M and S as well. *Overall, without complete mapping of the probability mass, with steep curvature distribution, increasing both S and M will significantly reduce the bias.*

- **Comparison of the Magnifying and Shifting methods**

  - $Exp(0.5; \cdot)$: In the truncated case the performances are very similar. In the censored case, the *bias* is smaller for the magnifying method when $S/N < 0.01\%$. In all other cases, the shifting method outperforms the magnifying one. This is due to the fact that the magnifying method drops many more observations by construction.
  - $\mathcal{N}(0, 0.2; \cdot)$: In the truncated case, the magnifying method decreases the bias much more efficiently than the shifting method. For the censored case, the results are very similar to the exponential distribution if M is small. However, the shifting method becomes better if we use larger M.

- **Survey design implications**
  - When some features of the underlying distribution are known or some assumptions about them can be made (about the curvature and the probability mass's distribution), then the most suitable method, split sample size, etc. can be picked for a given application:
    * With steep curvature you should use larger M.
    * When only a small fraction of probability mass is covered by the surveys, you must choose your main aim. If you intend to minimize the absolute bias, use shifting; if you prefer a small bias but are not worried about a more noisy estimator, then use the magnifying method.
  - In the case of shifting and/or censoring, extra choices on the boundaries can help to improve the performance of the methods:
    * In the case of shifting, you may add an extra small class in the boundaries, which will result in a faster bias reduction.
    * In the case of censoring, there is a clear cut from where to drop the observations, which enables us to control the censoring and thus reduce the number of dropped observations.

5 Extensions

5.1 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)). Let us call this the *perception effect*. The presence of this effect is independent of the implementation of the two split sampling methods. However, with split 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

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

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

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

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

$$y_i = \beta \tilde{x}_i^{**} + u_i \tag{20}$$

is equivalent to

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

Rewrite the above in matrix form using standard definitions gives

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

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 $\mathbf{B}$. 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

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

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

$$\hat{\pi}_{\chi,s} = \left(\sum_{i=1}^{N} \mathbf{1}_{\{\tilde{x}_i^{**} \in C_\chi, \tilde{x}_i^{**} \in \mathcal{S}_s\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}_{\{\tilde{x}_i^{**} \in C_\chi, \tilde{x}_i^{**} \in \mathcal{S}_s\}} \tilde{x}_i^{**} \tag{24}$$

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 (23) 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.

5.2 Non-linear Models

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

$$y_i = g(x_i; \beta) + u_i \tag{25}$$

where $g(\cdot)$ 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 as covariates 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.

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

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.

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.

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.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons; Republished by Dover Publications in 1968; reprinted in 1978.

Kullback, S. (1987). Letter to the Editor: The Kullback-Liebler distance. *The American Statistician*, 41:340–341.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

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.

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

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.

Terza, J. V. (1987). Estimating linear models with ordinal qualitative regressors. *Journal of Econometrics*, 34(3):275–291.

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.

Appendix

Structure of the Appendix:
Section A contains detailed research of the Monte Carlo experiments. Section B provides theoretical exposition of the simple Least Squares (LS) estimator on model with discretized data. The discussion covers cross section and panel data. Section C contains technical proofs of all the Propositions in the paper. Section D provides a list of notations used in the paper.

A Monte Carlo Simulation Results on the Bias

This section contains detailed results from all the Monte Carlo experiments. Recall the basic setup of the Monte Carlo experiment is,

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

The explanatory variable, x, is generated as Uniform, Normal, Exponential, and Weibull distributions with several different parameter setups. One thousand Monte Carlo experiments ($mc = 1, \ldots, 1000$) were run for each setup, for sample sizes ($N =$) 10,000; 100,000 and 500,000 and different σ_ε^2 variances. When generating x^*, observation outside the support, whenever relevant, would be discarded (truncated approach), or assigned to the limit of the class (censored approach). We report the *average bias* (bias: $\sum_{mc}(\hat{\beta}_{mc} - \beta)/1000$), the *average absolute bias* (absbias: $\sum_{mc}|\hat{\beta}_{mc} - \beta|/1000$), and the *standard deviation* of the $\hat{\beta}$ estimated parameter (SD: $\sqrt{\sum_{mc}(\hat{\beta}_{mc} - \bar{\beta}_{mc})^2/999}$). The Kullback–Leibler proximity/discrepancy index (Kullback and Leibler (1951), Kullback (1959), Kullback (1987)) has also been calculated to appreciate how different a given distribution is from the uniform:

$$KL = \int p(x) \log \frac{p(x)}{f(x)} dx,$$

where $p(x)$ is the uniform distribution and $f(x)$ is the relevant truncated or censored normal distribution.

A.1 Uniform Distribution

		Uniform[-1,1]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
$\hat{\beta}-\beta$	**N=10,000**	-0.0005	-0.0005	-0.0005	-0.0005	-0.0006
	N=100,000	-0.0008	-0.0010	-0.0008	-0.0008	-0.0008
	N=500,000	-0.0008	-0.0010	-0.0010	-0.0010	-0.0010
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0322	0.0307	0.0303	0.0302	0.0300
	N=100,000	0.0103	0.0100	0.0098	0.0097	0.0097
	N=500,000	0.0049	0.0049	0.0049	0.0048	0.0048
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0406	0.0390	0.0384	0.0382	0.0380
	N=100,000	0.0129	0.0124	0.0123	0.0122	0.0122
	N=500,000	0.0060	0.0059	0.0058	0.0058	0.0058
		Uniform[0,1]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
$\hat{\beta}-\beta$	**N=10,000**	-0.0008	-0.0008	-0.0008	-0.0008	-0.0008
	N=100,000	-0.0006	-0.0007	-0.0006	-0.0006	-0.0006
	N=500,000	-0.0010	-0.0012	-0.0012	-0.0011	-0.0012
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0298	0.0295	0.0293	0.0292	0.0292
	N=100,000	0.0100	0.0098	0.0098	0.0098	0.0098
	N=500,000	0.0044	0.0044	0.0044	0.0044	0.0044
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0375	0.0372	0.0369	0.0369	0.0369
	N=100,000	0.0126	0.0123	0.0123	0.0123	0.0123
	N=500,000	0.0054	0.0054	0.0054	0.0054	0.0054
		Uniform[0,10]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
$\hat{\beta}-\beta$	**N=10,000**	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
	N=100,000	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
	N=500,000	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
$\lvert\hat{\beta}-\beta\rvert$	**N=10,000**	0.0031	0.0030	0.0029	0.0029	0.0029
	N=100,000	0.0010	0.0010	0.0010	0.0010	0.0010
	N=500,000	0.0005	0.0004	0.0004	0.0004	0.0004
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0038	0.0037	0.0037	0.0037	0.0037
	N=100,000	0.0013	0.0012	0.0012	0.0012	0.0012
	N=500,000	0.0006	0.0005	0.0005	0.0005	0.0005

Table 6: Uniform distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 5$

From Table 6 the unbiasedness and consistency (in sample size) of the LS estimator can clearly be seen in the case of the uniform distribution, similarly to the, somewhat slower, convergence in M. We have also done simulations with different σ_ε^2 and β, where the same results hold. For smaller σ_ε^2, the bias is smaller, for different β the results are almost exactly the same.

Next, let us turn our attention to some other distributions.

A.2 Normal Distribution

From Table 7 it is clear that the LS estimator is biased and inconsistent, with a negative bias, as predicted by the theory, both in the case of truncation and censoring. Although the theory suggests that intercept picks up some of the bias, in practice the difference between with and without intercept – in this case – is small, approximately 3-5%. It also interesting to note that the Kullback-Liebler index gives a good indication of the bias (see Table 8). The bias tends to be smaller where this index is small, and vice versa.

	$\hat{\beta} - \beta$					
	Truncated			Censored		
	N=10,000	**N=100,000**	**N=500,000**	**N=10,000**	**N=100,000**	**N=500,000**
$\sigma_x^2 = 0.1$	-0.0593	-0.0603	-0.0607	-0.0582	-0.0567	-0.0575
$\sigma_x^2 = 0.2$	-0.0320	-0.0323	-0.0329	-0.0110	-0.0101	-0.0103
$\sigma_x^2 = 0.3$	-0.0224	-0.0223	-0.0226	0.0272	0.0283	0.0280
$\sigma_x^2 = 0.4$	-0.0176	-0.0171	-0.0173	0.0619	0.0630	0.0628
$\sigma_x^2 = 0.5$	-0.0142	-0.0139	-0.0141	0.0938	0.0950	0.0948
$\sigma_x^2 = 0.6$	-0.0118	-0.0118	-0.0120	0.1239	0.1248	0.1245
$\sigma_x^2 = 0.7$	-0.0102	-0.0103	-0.0105	0.1517	0.1527	0.1524
$\sigma_x^2 = 0.8$	-0.0092	-0.0091	-0.0093	0.1783	0.1791	0.1788
$\sigma_x^2 = 0.9$	-0.0082	-0.0082	-0.0084	0.2032	0.2042	0.2039
$\sigma_x^2 = 1$	-0.0074	-0.0075	-0.0077	0.2271	0.2280	0.2278
	$\lvert\hat{\beta} - \beta\rvert$					
	Truncated			Censored		
	N=10,000	**N=100,000**	**N=500,000**	**N=10,000**	**N=100,000**	**N=500,000**
$\sigma_x^2 = 0.1$	0.0730	0.0603	0.0607	0.0710	0.0568	0.0575
$\sigma_x^2 = 0.2$	0.0485	0.0326	0.0329	0.0417	0.0151	0.0106
$\sigma_x^2 = 0.3$	0.0416	0.0233	0.0226	0.0435	0.0285	0.0280
$\sigma_x^2 = 0.4$	0.0382	0.0188	0.0173	0.0651	0.0630	0.0628
$\sigma_x^2 = 0.5$	0.0363	0.0162	0.0141	0.0941	0.0950	0.0948
$\sigma_x^2 = 0.6$	0.0350	0.0147	0.0121	0.1239	0.1248	0.1245
$\sigma_x^2 = 0.7$	0.0339	0.0136	0.0107	0.1517	0.1527	0.1524
$\sigma_x^2 = 0.8$	0.0335	0.0129	0.0097	0.1783	0.1791	0.1788
$\sigma_x^2 = 0.9$	0.0331	0.0125	0.0089	0.2032	0.2042	0.2039
$\sigma_x^2 = 1$	0.0326	0.0121	0.0084	0.2271	0.2280	0.2278
	$SD\left[\hat{\beta}\right]$					
	Truncated			Censored		
	N=10,000	**N=100,000**	**N=500,000**	**N=10,000**	**N=100,000**	**N=500,000**
$\sigma_x^2 = 0.1$	0.0661	0.0212	0.0098	0.0662	0.0210	0.0088
$\sigma_x^2 = 0.2$	0.0520	0.0165	0.0079	0.0518	0.0156	0.0068
$\sigma_x^2 = 0.3$	0.0473	0.0150	0.0072	0.0457	0.0137	0.0059
$\sigma_x^2 = 0.4$	0.0451	0.0144	0.0068	0.0421	0.0128	0.0055
$\sigma_x^2 = 0.5$	0.0436	0.0139	0.0067	0.0403	0.0124	0.0053
$\sigma_x^2 = 0.6$	0.0428	0.0136	0.0065	0.0387	0.0120	0.0051
$\sigma_x^2 = 0.7$	0.0419	0.0134	0.0064	0.0379	0.0117	0.0050
$\sigma_x^2 = 0.8$	0.0415	0.0132	0.0064	0.0368	0.0115	0.0049
$\sigma_x^2 = 0.9$	0.0412	0.0132	0.0063	0.0360	0.0114	0.0047
$\sigma_x^2 = 1$	0.0408	0.0131	0.0063	0.0356	0.0113	0.0047

Table 7: Truncated and Censored Normal Distributions, estimated without intercept, $M = 5, \beta = 0.5, \sigma_\varepsilon^2 = 5, Supp = [-1, 1]$

	Truncated	**Censored**
$\sigma_x^2 = 0.1$	0.7396	0.7407
$\sigma_x^2 = 0.2$	0.2287	0.2536
$\sigma_x^2 = 0.3$	0.1091	0.1783
$\sigma_x^2 = 0.4$	0.0634	0.1829
$\sigma_x^2 = 0.5$	0.0414	0.2109
$\sigma_x^2 = 0.6$	0.0291	0.2463
$\sigma_x^2 = 0.7$	0.0216	0.2835
$\sigma_x^2 = 0.8$	0.0167	0.3203
$\sigma_x^2 = 0.9$	0.0132	0.3558
$\sigma_x^2 = 1$	0.0197	0.3899

Table 8: Kullback-Leibler ratio: Uniform vs. Truncated/Censored Normal with different σ_x^2 values, $a = -1, b = 1$

A.3 Exponential Distribution and Weibull Distributions

We carried out a large number of simulations with different parametrisations for both distributions. In Table 9 we report the bias from the exponential distribution, which highlights the effect of censoring. Although we do no observe large bias with truncation, when the choices are censored the bias increases dramatically. Table 10 shows results on normal distribution, while Table 11 uses a weibull distribution.
From Table 9-11, the main takeaway is that, as expected, there is no convergence in the sample size, while the convergence speed in M is 'slow' and depends heavily on the shape of the distribution. Also, the results about the Kullback-Liebler index (not reported here) are very similar to those obtained for the normal distribution, i.e., a larger index implies systematically a larger bias.
We have also tried several different distributions and parameterisation, and the main take away is very similar.

		$Exp[\lambda]$, $Supp = [0,1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
$\hat{\beta} - \beta$	**N=10,000**	-0.0182	-0.0074	-0.0027	-0.0015	-0.0011	0.1341	0.1304	0.1235	0.1190	0.1160
	N=100,000	-0.0185	-0.0072	-0.0025	-0.0014	-0.0011	0.1342	0.1307	0.1239	0.1193	0.1163
	N=500,000	-0.0190	-0.0078	-0.0032	-0.0020	-0.0017	0.1339	0.1303	0.1235	0.1190	0.1160
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0415	0.0394	0.0388	0.0388	0.0388	0.1342	0.1305	0.1237	0.1191	0.1162
	N=100,000	0.0208	0.0145	0.0133	0.0131	0.0131	0.1342	0.1307	0.1239	0.1193	0.1163
	N=500,000	0.0191	0.0090	0.0064	0.0060	0.0059	0.1339	0.1303	0.1235	0.1190	0.1160
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0489	0.0489	0.0489	0.0490	0.0490	0.0445	0.0437	0.0427	0.0422	0.0419
	N=100,000	0.0163	0.0165	0.0164	0.0164	0.0164	0.0137	0.0135	0.0131	0.0130	0.0129
	N=500,000	0.0073	0.0073	0.0073	0.0073	0.0073	0.0061	0.0059	0.0058	0.0057	0.0057

Table 9: Exponential distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 5, \lambda = 0.5$

		$\mathcal{N}\left(\mu_x, \sigma_x^2\right), Supp = [-1, 1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
$\hat{\beta} - \beta$	**N=10,000**	-0.0798	-0.0311	-0.0078	-0.0017	0.0000	-0.0552	-0.0097	0.0088	0.0120	0.0120
	N=100,000	-0.0800	-0.0313	-0.0079	-0.0017	0.0000	-0.0552	-0.0099	0.0084	0.0115	0.0114
	N=500,000	-0.0803	-0.0315	-0.0081	-0.0020	-0.0003	-0.0554	-0.0100	0.0082	0.0113	0.0112
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0798	0.0328	0.0198	0.0188	0.0187	0.0553	0.0195	0.0198	0.0209	0.0209
	N=100,000	0.0800	0.0313	0.0092	0.0066	0.0064	0.0552	0.0106	0.0092	0.0117	0.0117
	N=500,000	0.0803	0.0315	0.0081	0.0032	0.0028	0.0554	0.0100	0.0082	0.0113	0.0112
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0224	0.0226	0.0234	0.0234	0.0234	0.0220	0.0228	0.0230	0.0229	0.0228
	N=100,000	0.0074	0.0078	0.0080	0.0080	0.0080	0.0074	0.0074	0.0074	0.0074	0.0074
	N=500,000	0.0033	0.0033	0.0034	0.0034	0.0034	0.0031	0.0033	0.0033	0.0033	0.0032

Table 10: Normal distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 1, \mu_x = 0, \sigma_x^2 = 0.2$

		$Weibull\,[b, c]\,, Supp = [0, 1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
$\hat{\beta} - \beta$	**N=10,000**	-0.0369	-0.0128	-0.0031	-0.0010	-0.0004	1.8197	1.7475	1.6828	1.6486	1.6278
	N=100,000	-0.0369	-0.0130	-0.0033	-0.0011	-0.0005	1.8209	1.7487	1.6840	1.6498	1.6289
	N=500,000	-0.0371	-0.0131	-0.0035	-0.0013	-0.0007	1.8197	1.7475	1.6828	1.6486	1.6278
$\lvert\hat{\beta} - \beta\rvert$	**N=10,000**	0.0371	0.0178	0.0144	0.0142	0.0141	1.8197	1.7475	1.6828	1.6486	1.6278
	N=100,000	0.0369	0.0131	0.0056	0.0049	0.0048	1.8209	1.7487	1.6840	1.6498	1.6289
	N=500,000	0.0371	0.0131	0.0038	0.0024	0.0022	1.8197	1.7475	1.6828	1.6486	1.6278
$SD\left[\hat{\beta}\right]$	**N=10,000**	0.0174	0.0179	0.0179	0.0179	0.0179	0.0492	0.0474	0.0458	0.0450	0.0445
	N=100,000	0.0058	0.0060	0.0060	0.0060	0.0060	0.0154	0.0148	0.0144	0.0141	0.0140
	N=500,000	0.0026	0.0027	0.0027	0.0027	0.0027	0.0071	0.0069	0.0066	0.0065	0.0064

Table 11: Weibull distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 0.5, b = 1, c = 0.5$

B Properties of LS using Discretized Data

Recall the data generating process is assumed to be

$$y_i = w_i'\gamma + x_i'\beta + u_i \tag{26}$$

with the linear regression model using the discretized version of x_i namely,

$$y_i = w_i'\gamma + x_i^{*'}\beta + u_i \tag{27}$$

Let us assume for the sake of simplicity that there is only one explanatory variable in the model which is observed through discretized choices. It is also assumed, as said earlier, that it has a known support $[a_l, a_u]$ with known boundaries (C_m), and let z_m be the class midpoint.[14]

[14]In the special case of the uniform distribution, the midpoints coincide with the conditional expectation of the uniformly distributed explanatory variable x in that class.

The classes are now the following with their respective class values:

$$
\begin{aligned}
C_1 &= \left[a_l, a_l + \frac{a_u - a_l}{M}\right) && z_1 = a_l + \frac{a_u - a_l}{2M}, \\
&\vdots \\
C_m &= \left[a_l + (m-1)\frac{(a_u - a_l)}{M}, a_l + m\frac{a_u - a_l}{M}\right) && z_m = a_l + (2m-1)\frac{a_u - a_l}{2M}, \\
&\vdots \\
C_M &= \left[a_l + (M-1)\frac{(a_u - a_l)}{M}, a_l + M\frac{a_u - a_l}{M}\right] && z_M = a_l + (2M-1)\frac{a_u - a_l}{2M}.
\end{aligned} \tag{28}
$$

Let N_m be the number of observations in each class C_m, that is $N_m = \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}$, where $\mathbf{1}_{\{x \in C\}}$ denotes the indicator function defined as

$$
\mathbf{1}_{\{x \in C\}} := \begin{cases} 1, & \text{if } x \in C, \\ 0, & \text{if } x \notin C. \end{cases}
$$

When x has a cumulative distribution cdf $F(\cdot)$,

$$
\begin{aligned}
\mathbb{E}(N_m) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\right) \\
&= N \int_{C_m} f(x)\,\mathrm{d}x \\
&= N \Pr(c_{m-1} < x \leq c_m),
\end{aligned}
$$

using the independence assumption. When, for example, x has a uniform distribution, we have $\mathbb{E}(N_m) = N/M$ for all $m = 1, \ldots, M$.

$$
\begin{aligned}
\hat{\beta}^*_{LS} &= (x^{*'}x^*)^{-1}(x^{*'}y) \\
&= \frac{z_1\left(\sum_{i=1}^{N_1} y_i\right) + z_2\left(\sum_{i=N_1+1}^{N_1+N_2} y_i\right) + \cdots + z_M\left(\sum_{i=N-N_M+1}^{N_M} y_i\right)}{N_1 z_1^2 + N_2 z_2^2 + \cdots + N_M z_M^2} \\
&= \frac{z_1\left(\sum_{i=1}^{N_1} \beta x_i + u_i\right) + \cdots + z_M\left(\sum_{i=N-N_M+1}^{N_M} \beta x_i + u_i\right)}{N_1 z_1^2 + \cdots + N_M z_M^2} \\
&= \frac{z_1\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_1\}}(\beta x_i + u_i)\right] + \cdots + z_M\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_M\}}(\beta x_i + u_i)\right]}{N_1 z_1^2 + \cdots + N_M z_1^2} \\
&= \frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \\
&= \frac{\sum_{m=1}^{M}\left[a_l + (2m-1)\frac{a_u - a_l}{2M}\right]\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m\left[a_l + (2m-1)\frac{a_u - a_l}{2M}\right]^2}.
\end{aligned}
$$

Using the result above, we can get the following general formula for the expected value of the LS estimator,

$$
\begin{aligned}
\mathbb{E}\left(\hat{\beta}^*_{LS}\right) &= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\left(\beta(x_i^* + \xi_i) + u_i\right)\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m\left[\beta\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^* + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\xi_i\right) + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^*}{\sum_{m=1}^{M} N_m z_m^2}\right\} + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&\qquad + \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m \upsilon^m}{\sum_{m=1}^{M} N_m z_m^2}\right\}. \qquad (29)
\end{aligned}
$$

where a respondent makes an error $\xi_i = x_i - x_i^*$ for each observation by setting the possible answer values at x_i^*. The derivation above is based on the disturbance term u_i being independent of regressor x_i and $\mathbb{E}(u_i) = 0$ for all $i = 1, \ldots, N$. The

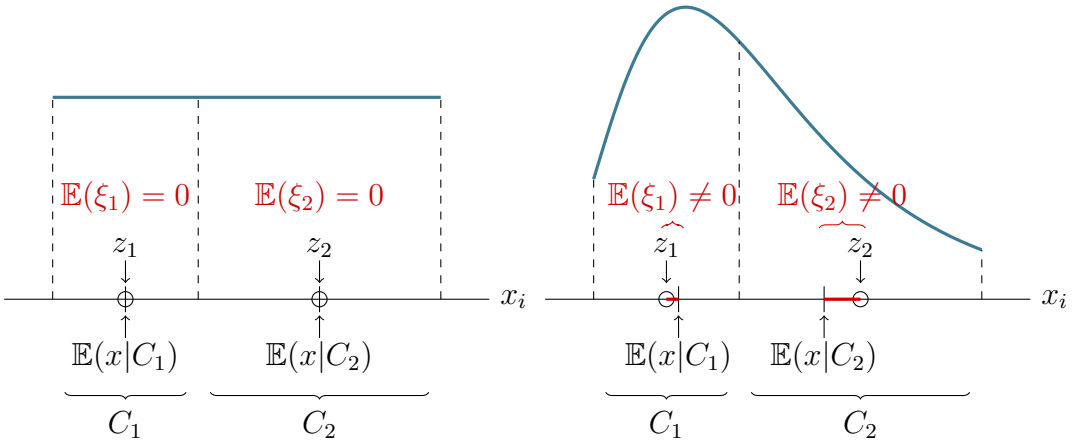

Figure 5: The difference between uniform (left panel) and general distributions (right panel)

last inference uses the fact that the errors ξ_i have the same conditional distribution over the class C_m, $\upsilon^m \overset{\text{d}}{=} \xi_i|C_m$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. Importantly, the second term in Equation (29) does not vanish in general, since $\upsilon^m|C_m$ is not independent of $N_m|C_m$, $\upsilon^m|C_m \not\perp\!\!\!\perp N_m|C_m$ nor $\mathbb{E}(\xi_i|C_m) = \mathbb{E}(\upsilon^m) = 0$ (see Figure 5, right panel). These would be sufficient assumptions for the LS to be unbiased. The former issue can be eliminated by conditioning on the underlying distribution of x_i. Conditional on the distribution x_i and the class C_m, the number of observations in the class and assuming that the errors are independent of each other, $N_m|x_i, C_m \perp\!\!\!\perp \upsilon^m|x_i, C_m$, but knowing the underlying distribution makes the problem trivial. Nonetheless, because of both issues, the 'naive' LS estimator is biased.

The uniform distribution, however, turns out to be a special case. Let us assume that $x_i \sim U(a_l, a_u)$ for all $i = 1, \ldots, N$, then both of the above disappear (see the left panel in Figure 5) if we are using the class mid points. The first problem is resolved, because in the case of the uniform distribution, both the number of observations N_m in each class C_m and the error term υ^m are independent of the regressor's x_i distribution, while the second problem does not appear trivially, since now the class midpoints are proper estimates of the regressor's x_i expected value in the class C_m. From Equation (29), we obtain that

$$\mathbb{E}\left(\hat{\beta}^*_{LS}\right) = \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m \upsilon^m}{\sum_{m=1}^{M} N_m z_m^2}\right\} = \beta,$$

where υ^m is a uniformly distributed random variable with zero expected value, $\mathbb{E}(\upsilon^m) = 0$ for all $m = 1, \ldots, M$. Hence, in the case of uniform distribution, unlike for other distributions, the LS is unbiased.

B.1 N (in)consistency

This subsection considers the large sample properties of the estimator. First, assume that $\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}u_i = 0$, in other words that the choice set selection is independent of the disturbance terms, and also that with sample size N the number of classes M is fixed. Then

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\hat{\beta}^{*}_{LS} &= \operatorname*{plim}_{N\to\infty}\frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \\
&= \frac{\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\beta\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\beta\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m}.
\end{aligned} \tag{30}$$

Define $x^m = \sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i$, then x^m sums the truncated version of the original random variables x_i on the class C_m, $x_m \overset{\mathrm{d}}{=} x_i|C_m$, for all $m = 1,\dots,M$, therefore its asymptotic distribution can be calculated by applying the Lindeberg-Levy Central Limit Theorem,

$$x^m/N_m \overset{\mathrm{a}}{\sim} N\big(\mathbb{E}(x_m), \mathrm{V}(x_m)/N_m\big).$$

The $\hat{\beta}^{*}_{LS}$ estimator is consistent if and only if the probability limit in Equation (30) equals β. To give a condition for consistency, first we rewrite the previous Equation (30) in terms of the error terms ξ_i,

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^{*}_{LS} - \beta\right) &= \frac{\beta\left(\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i\right] - \sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m\right)}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\beta\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(x_i - x_i^{*})\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\beta\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}\xi_i\right]}{\sum_{m=1}^{M} z_m^2 \operatorname{plim}_{N\to\infty} N_m},
\end{aligned}$$

where the asymptotic distribution of the sum of errors in class C_m, $\xi^m = \sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}\xi_i$, $m = 1,\dots,M$, can be given by

$$\xi^m/N_m \overset{\mathrm{d}}{=} x^m/N_m - z_m \overset{\mathrm{a}}{\sim} N\big(\mathbb{E}(x^m) - z_m, \mathrm{V}(x^m)/N_m\big).$$

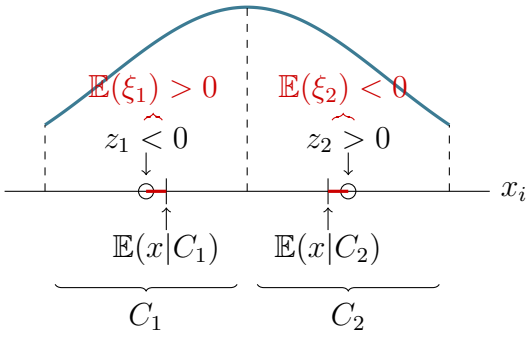

Figure 6: The estimator is inconsistent even in case of symmetric distributions (see Equation (31)).

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^*_{LS}-\beta\right) &= \frac{\operatorname{plim}_{N\to\infty}\beta\sum_{m=1}^{M} z_m\xi^m}{\operatorname{plim}_{N\to\infty}\sum_{m=1}^{M} z_m^2 N_m} \\
&= \frac{\operatorname{plim}_{N\to\infty} O(N)\beta\sum_{m=1}^{M} z_m\xi^m/N_m}{\operatorname{plim}_{N\to\infty} O(N)\sum_{m=1}^{M} z_m^2} \\
&= \frac{\beta\sum_{m=1}^{M} z_m \operatorname{plim}_{N\to\infty}\xi^m/N_m}{\sum_{m=1}^{M} z_m^2} O(N) \\
&= \frac{\beta\sum_{m=1}^{M} z_m\left\{\mathbb{E}(x_m)-z_m\right\}}{\sum_{m=1}^{M} z_m^2} O(N). \qquad (31)
\end{aligned}$$

The last step in the above derivation can simply be obtained from the definition of the plim operator, i.e., for any $\varepsilon > 0$ given Therefore, to obtain the (in)consistency of the LS estimator $\hat{\beta}^*_{LS}$ in the number of observations N, we only need to calculate the expected value of the truncated random variable x_m, $m = 1, \ldots, M$ and check whether the expression (31) equals 0 to satisfy a sufficient condition.

Let us apply these results to the uniform distribution. In this case, there is no consistency issue because the class midpoints coincide with the expected value of the truncated uniform random variable in each class, making the expression (31) zero, hence the LS estimator is consistent.

Note that the consistency of the LS estimator is not guaranteed even in the case of symmetric distributions and symmetric class boundaries. After appropriate transformations (e.g., demeaning), it can be seen that the sign of the differences between the expectation of the truncated random variables x_m and the class midpoints is opposite to the sign of the class midpoints on either side of the distribution, which implies negative overall asymptotic bias in N (see Figure 6).

In the case of a (truncated) normal variable, for example, we need to substitute the expected value of the truncated normal random variable x_m for each

$m = 1, \ldots, M$ in the consistency formula (31). As a result, the difference between the expectation and the class midpoints in general is not zero for all m, hence the formula cannot be made arbitrarily small. Therefore, the LS estimator becomes inconsistent in N (see Table 7. on the size of the bias).

So far we have focused on the estimation of β in Equation (27). But how about γ? It can be shown that the bias and inconsistency presented above is contagious. Estimation of all parameters of a model is going to be biased and inconsistent unless the measurement error and x are orthogonal (independent), which is quite unlikely in practice. This is important to emphasize: a single choice type variable in a model is going to infect the estimation of all variables of the model.

B.2 M Consistency

Let us see next the case when N is fixed but $M \to \infty$. Now, we may have some classes that do not contain any observations, while others still do. Omitting, however, empty classes does not cause any bias because of our iid assumption. Furthermore, while we increase the number of classes, the size of the classes itself is likely to shrink and become so narrow that only one observation can fall into each. In the limit we are going to hit the observations with the class boundaries. To see that, we derive the consistency formula in the number of classes M assuming that $\text{plim}_{M\to\infty} \sum_{\{m:C_m\neq\emptyset, m=1,\ldots,M\}} z_m u_{i_m} = 0$, or with re-indexation $\text{plim}_{M\to\infty} \sum_{i=1}^{N} z_{m_i} u_i = \sum_{i=1}^{N} x_i u_i = 0$, which should hold in the sample and is a

stronger assumption than the usual $\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N} x_i u_i = 0$:

$$\begin{aligned}
\operatorname*{plim}_{M\to\infty}\left(\hat{\beta}^{*}_{LS}-\beta\right) &= \operatorname*{plim}_{M\to\infty}\frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(\beta x_i+u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2}-\beta \\
&= \operatorname*{plim}_{M\to\infty}\frac{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m\left[\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(\beta x_i+u_i)\right]}{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} N_m z_m^2}-\beta \\
&= \operatorname*{plim}_{M\to\infty}\frac{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m(\beta x_{i_m}+u_{i_m})}{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m^2}-\beta \\
&= \operatorname*{plim}_{M\to\infty}\beta\left\{\frac{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m x_{i_m}}{\sum_{\{m:C_m\neq\emptyset,m=1,\ldots,M\}} z_m^2}-1\right\} \\
&= \operatorname*{plim}_{M\to\infty}\beta\left\{\frac{\sum_{i=1}^{N} z_{m_i} x_i}{\sum_{i=1}^{N} z_{m_i}^2}-1\right\} \\
&= \beta\left\{\frac{\sum_{i=1}^{N}\operatorname{plim}_{M\to\infty} z_{m_i} x_i}{\sum_{i=1}^{N}\operatorname{plim}_{M\to\infty} z_{m_i}^2}-1\right\} \\
&= \beta\left\{\frac{\sum_{i=1}^{N} x_i x_i}{\sum_{i=1}^{N} {x_i}^2}-1\right\} \\
&= 0,
\end{aligned}$$

where the index $i_m \in \{1,\ldots,N\}$ denotes observation i in class m (at the beginning there might be several observations that belong to the same class m), and index $m_i \in \{1,\ldots,M\}$ denotes the class m that contains observation i (at the and of the derivation one class m includes only one observation i). Note that the derivation does not depend on the distribution of the explanatory variable x, so consistency in the number of classes M holds in general. Let us also note, however, that this convergence in M is slow. Also, as $M \to \infty$, the class sizes go to zero, and the smaller the class sizes the smaller the bias. Of course, in practice, the number of classes M cannot be too large due to the limits of our cognitive capacities. Typically, the optimal number of choices for a survey is relatively small, $M = 3, 5, 7$ or at most $M = 10$.[15]

B.3 Some Remarks

The above results hold for much simpler cases as well. If instead of model (27) we just take the simple sample average of x, $\bar{x} = \sum_i x_i/N$, then $\bar{x}^* = \sum_i x_i^*/N$ is going to be a biased and inconsistent estimator of $\bar{x}$.

[15]There is an abundant 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).

The measurement error due to discretized choice variables, however, not only induces correlation between the error terms and the observed variables, but it also induces a non-zero expected value for the disturbance terms of the regression in (27). Consider a simple example where there is an unobserved variable x_i with an observed discretized choice version:

$$x_i^* = \begin{cases} z_1 & \text{if } c_0 \leq x_i < c_1, \\ z_2 & \text{if } c_1 \leq x_i < c_2, \end{cases} \tag{32}$$

and

$$y_i = x_i\beta + \varepsilon_i. \tag{33}$$

Using the discretized choice variable means:

$$y_i = x_i^*\beta + (x_i - x_i^*)\beta + u_i \tag{34}$$

and $\mathbb{E}\left[x_i - x_i^*\right]$ is

$$\begin{aligned} \mathbb{E}\left[x_i - x_i^*\right] =& \mathbb{E}(x_i) - \mathbb{E}(x_i^*) \\ =& \mathbb{E}(x_i) - \mathbb{E}\left[z_1\mathbf{1}(c_0 \leq x_i < c_1) + z_2\mathbf{1}(c_1 \leq x_i < c_2)\right] \\ =& \mathbb{E}(x_i) - z_1 \Pr(c_0 \leq x_i < c_1) - z_2 \Pr(c_1 \leq x_i < c_2). \end{aligned}$$

The last line above is not zero in general. Thus, it would induce a bias in the estimator if the regression did not include an intercept. This result generalizes naturally to variables with multiple choice values.

B.4 Estimation Reconsidered

Let us generalise the problem and re-write it in matrix form. Consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{35}$$

where $\mathbf{X}$ and $\mathbf{W}$ are $N \times K$ and $N \times J$ data matrices of the explanatory variables, $\mathbf{y}$ is a $N \times 1$ vector containing the data of the dependent variable, $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of disturbance terms, and finally $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $K \times 1$ and $J \times 1$ parameter vectors.

$\mathbf{X}$ is not observed, only its discretized ordered choice version $\mathbf{X}^*$ is. Define the $MK \times K$ matrix as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{0} & \dots & \dots \\ \mathbf{0} & \mathbf{z}_2 & \mathbf{0} & \mathbf{0} \\ \vdots & \dots & \ddots & \vdots \\ \dots & \dots & \mathbf{0} & \mathbf{z}_K \end{bmatrix},$$

where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iM})'$ contains the choice values for variable i. Let $\mathbf{E} = \{\mathbf{e}_{ki}\}$, where $k = 1, \ldots K$ and $i = 1, \ldots, N$ such that

$$\mathbf{e}_{ki} = \begin{bmatrix} \mathbf{1}(c_{k0} \leq x_{ki} < c_{k1}) \\ \mathbf{1}(c_{k1} \leq x_{ki} < c_{k2}) \\ \vdots \\ \mathbf{1}(c_{kM-1} \leq x_{ki} < c_{kM}) \end{bmatrix},$$

where x_{ki} denotes the value of the i^{th} observation from the explanatory variable x_k.

This implies $\mathbf{E}$ is a $MK \times N$ matrix since each entry $\mathbf{e}_{ki}$ is a $M \times 1$ vector. Following the definition of x_i^* in the paper, we can rewrite $\mathbf{X}^* = \mathbf{E}'\mathbf{Z}$.

B.4.1 The LS Estimator

From Equation (35), consider the regression based on the observed data:

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + (\mathbf{X} - \mathbf{X}^*)\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{36}$$

then the LS estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{W}}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{W}}\mathbf{y}\,,$$

where $\mathbf{M}_{\mathbf{W}} = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$ defines the usual residual maker. The standard derivation shows that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\boldsymbol{\varepsilon}. \tag{37}$$

This implies LS is unbiased if and only if $\left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{I}$. This allows us to investigate the bias analytically by examining the elements in $\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}$ and $\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}$.

To simplify the analysis, we assume for the time being the following:

$$\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{X} \tag{38}$$

$$\mathbf{M}_{\mathbf{W}}\mathbf{X}^* = \mathbf{X}^*. \tag{39}$$

In other words, we assume independence between $\mathbf{W}$ and $\mathbf{X}$, as well as its discretized choice version. This may appear to be a strong assumption but it does allow us to see what is happening somewhat better. We relax this at a latter stage.

The LS estimator in this case becomes:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\boldsymbol{\varepsilon}.$$

The LS is unbiased if $\left(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{X} = \mathbf{I}$. Note that $\mathbf{Z}'$ and $\mathbf{E}$ are of size $K \times MK$ and $MK \times N$, respectively. This means $\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}$ are invertible as long as $N > K$,

which is a standard assumption in classical regression analysis. Let us consider a typical element in $\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}$ first. Since $\mathbf{Z}$ is non-stochastic as it contains only all the pre-defined choice values, it is sufficient to examine $\mathbf{E}\mathbf{E}'$:

$$\mathbf{E}\mathbf{E}' = \begin{bmatrix} \mathbf{e}_{11} & \dots & \mathbf{e}_{1i} & \dots & \mathbf{e}_{1N} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}_{k1} & \dots & \mathbf{e}_{ki} & \dots & \mathbf{e}_{kN} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}_{K1} & \dots & \mathbf{e}_{Ki} & \dots & \mathbf{e}_{KN} \end{bmatrix} \begin{bmatrix} \mathbf{e}'_{11} & \dots & \mathbf{e}'_{k1} & \dots & \mathbf{e}'_{K1} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}'_{1i} & \dots & \mathbf{e}'_{ki} & \dots & \mathbf{e}'_{ki} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \mathbf{e}'_{1N} & \dots & \mathbf{e}'_{kN} & \dots & \mathbf{e}'_{KN} \end{bmatrix}.$$

Note that each entry in $\mathbf{E}$ is a vector, so $\mathbf{E}\mathbf{E}'$ will result in a partition matrix whose elements are the sums of the outer products of $\mathbf{e}_{ki}$ and $\mathbf{e}_{lj}$ for $k, l = 1, \dots, K$ and $i, j = 1, \dots, N$. Specifically, let $\mathbf{q}_{kl}$ be a typical block element in $\mathbf{E}\mathbf{E}'$, then

$$\mathbf{q}_{kl} = \sum_{i=1}^{N} \mathbf{e}_{ki}\mathbf{e}'_{li}.$$

Let $\mathbf{1}_m^{ki} = \mathbf{1}\left(c_{km-1} \leq x_{ki} < c_{km}\right)$, then the (m, n) element in $\mathbf{q}_{kl}$, q_{mn} is $\sum_{i=1}^{N} \mathbf{1}_m^{ki}\mathbf{1}_n^{li}$ for $m, n = 1, \dots, M$. Thus, $\mathbb{E}\left(\mathbf{E}\mathbf{E}'\right)$ exists if $\mathbb{E}\left(\mathbf{1}_m^{ki}\mathbf{1}_n^{li}\right)$ exists,

$$\mathbb{E}\left(\mathbf{1}_m^{ki}\mathbf{1}_n^{li}\right) = \int_{\Omega} f(x_k, x_l) dx_k dx_l \,, \tag{40}$$

where $f(x_k, x_l)$ denotes the joint distribution of x_k and x_l and $\Omega = [c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}]$ defines the region for integration. Thus, $N^{-1}b_{mn}$ should converge into Equation (40) under the usual WLLN.

Following a similar method, let a_{kl} be the (k, l) element in $\mathbf{Z}'\mathbf{E}\mathbf{X}$, then

$$a_{kl} = \sum_{i=1}^{N} \sum_{m=1}^{M} z_{km}\mathbf{1}_m^{ki}x_{li}.$$

Now,

$$\begin{aligned} \mathbb{E}\left[\sum_{m=1}^{M} z_{km}\mathbf{1}_m^{ki}x_{li}\right] &= \sum_{m=1}^{M} z_{km}\mathbb{E}\left[\mathbf{1}_m^{ki}x_{li}\right] \\ &= \sum_{m=1}^{M} z_{km} \int_{\Omega_1} x_l f(x_k, x_l) dx_k dx_l \,, \end{aligned} \tag{41}$$

where $\Omega_1 = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}}$ with $\Omega_{\mathbf{X}}$ denotes the sample space of x_k and x_l. Thus, $N^{-1}a_{kl}$ should converge into Equation (41) under the usual WLLN.

In the case when Equations (38) and (39) do not hold, the analysis becomes more tedious algebraically, but it does not affect the result that LS is biased. Recall Equation (37), and let ω_{ij} be the (i,j) element in $\mathbf{M_W}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, then following the same argument as above, $\mathbf{EM_WE'}$ can be expressed as a $M \times M$ block partition matrix with each entry a $K \times K$ matrix. The typical (m,n) element in the (k,l) block is

$$g_{kl} = \sum_{j=1}^{N} \sum_{i=1}^{N} \omega_{ij} \mathbf{1}_m^{ki} \mathbf{1}_n^{li} \tag{42}$$

with its expected value being

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \int_{\Omega} \omega_{ij} f\left(x_k, x_l, \mathbf{w}\right) dx_k dx_k d\mathbf{w}, \tag{43}$$

where $\mathbf{w} = (w_1, \ldots, w_J)$, $d\mathbf{w} = \prod_{i=1}^{J} dw_i$ and $\Omega = [c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}] \times \Omega_{\mathbf{w}}$ where $\Omega_{\mathbf{w}}$ denotes the sample space of $\mathbf{w}$. Note that ω_{ij} is a nonlinear function of $\mathbf{w}$, and so the condition of existence for Equation (43) is complicated. However, under the assumption that the integral in Equation (43) exits, then $N^{-1} g_{kl}$ should converge to Equation (43) under the usual WLLN. It is also worth noting that $\mathbb{E}\left[\mathbf{M_W X}\right] = \mathbb{E}\left[\mathbf{M_W}\right] \mathbb{E}\left[\mathbf{X}\right] = \mathbb{E}\left[\mathbf{X}\right]$ and $\mathbb{E}\left[\mathbf{M_W X^*}\right] = \mathbb{E}\left[\mathbf{M_W}\right] \mathbb{E}\left[\mathbf{X^*}\right] = \mathbb{E}\left[\mathbf{X^*}\right]$ under the assumption of independence, which reduces Equation (43) to Equation (40). Again, following the same derivation as above, a typical element in $\mathbf{Z'EM_WX}$ is

$$h_{kl} = \sum_{m=1}^{M} \sum_{i=1}^{N} z_{km} \mathbf{1}_m^{ki} u_{li}, \tag{44}$$

where $u_{li} = \sum_{v=1}^{N} \omega_{iv} X_{lv}$. Note that u_{li} is the i^{th} residual of the regression of X_l on $\mathbf{W}$. The expected value of h_{kl} can be expressed as

$$\sum_{m=1}^{M} z_{km} \int_{\Omega_m} u_l f(x_k, x_l, \mathbf{w}) dx_k dx_l d\mathbf{w}, \tag{45}$$

where u_l denotes the random variable corresponding to the i^{th} column of $\mathbf{M_W X}$ and $\Omega_m = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}} \times \Omega_{\mathbf{w}}$ with $\Omega_{\mathbf{w}}$ denotes the sample space of $\mathbf{W}$. Note that $u_l = x_l$ under the assumption of independence, which reduces Equation (45) to Equation (41).

B.5 Extension to Panel Data

So far, we have dealt with cross-sectional data. Next, let us see what changes if we have panel data at hand, which is closer to the reality of data gathering through surveys. We can extend our basic model using Equation (27) to

$$y_{it} = w_{it}'\gamma + x_{it}^{*'}\beta + \varepsilon_{it}, \tag{46}$$

and adjust the DGP, based on Equation (26)

$$y_{it} = w_{it}'\gamma + x_{it}'\beta + u_{it}, \tag{47}$$

where $x_{it} \sim f_i(a_l, a_u)$ denotes an individual distribution with mean μ_i for $i = 1, \ldots, N$. Here we need to assume that $f_i(\cdot)$ is stationary, so the distribution may change over individual i but not over time, t.

Now, the most important problem is identification. If the choice of an individual does not change over the time periods covered, the individual effects in the panel and the parameter associated with the choice variable cannot be identified separately. The within transformation would wipe out the choice variable as well. When the choice does change over time, but not much, then we are facing weak identification, i.e., in fact very little information is available for identification, so the parameter estimates are going to be highly unreliable. This is a likely scenario when M is small, for example $M = 3$ or $M = 5$.

The bias of the panel data within estimator can be easily shown. Let us re-write Equation (36) in a panel data context

$$\mathbf{y} = \mathbf{D}_N\boldsymbol{\alpha} + \mathbf{X}^*\boldsymbol{\beta} + [(\mathbf{X} - \mathbf{X}^*)\boldsymbol{\beta} + \boldsymbol{\varepsilon}],$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ and $\mathbf{D}_N$ is a $NT \times N$ zero-one matrix that appropriately selects the corresponding fixed effect elements form $\boldsymbol{\alpha}$. The Within estimator is

$$\hat{\boldsymbol{\beta}}_W^* = (\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{D}_N}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{M}_{\mathbf{D}_N}\mathbf{y},$$

or equivalently

$$\hat{\boldsymbol{\beta}}_W^* = (\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\mathbf{X}\boldsymbol{\beta} + (\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{D}_N}\boldsymbol{\varepsilon},$$

where

$$\mathbf{M}_{\mathbf{D}_N}\mathbf{y} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^*\boldsymbol{\beta} + \mathbf{M}_{\mathbf{D}_N}[(\mathbf{X} - \mathbf{X}^*)\boldsymbol{\beta} + \boldsymbol{\varepsilon}].$$

The Within estimator is biased as $\mathbb{E}(\hat{\boldsymbol{\beta}}_W^*) \neq \boldsymbol{\beta}$, because $\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^* \neq \mathbf{M}_{\mathbf{D}_N}\mathbf{X}$.

$$
\begin{aligned}
&\operatorname*{plim}_{N\to\infty} \xi^m = \mathbb{E}(X_m) - z_m \\
&\quad \iff \lim_{N\to\infty} \Pr\left(|\xi^m - \{\mathbb{E}(X_m) - z_m\}| > \varepsilon\right) \\
&\qquad = \lim_{N\to\infty} F_{\xi^m}\left(-\varepsilon + \mathbb{E}(X_m) - z_m\right)\left[1 - F_{\xi^m}\left(\varepsilon + \mathbb{E}(X_m) - z_m\right)\right] = 0.
\end{aligned}
$$

The convergence holds, because for any given $\delta > 0$, there is a threshold N_0 for which the term in the limit becomes less than δ. This can be seen from $F_{\xi^m}(\cdot)$ being close to a degenerate distribution above a threshold number of observations N_0, or intuitively, since the variance of the sequence of random variables ξ^m collapses in N, its probability limit equals its expected value.

C Technical Proofs

C.1 Proof of Proposition 1

Recall

$$
\begin{aligned}
\mathbb{E}(N_b^{WS}) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_b^{WS}\}}\right) \\
&= N \Pr(x \in \mathcal{S}_s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x.
\end{aligned}
\tag{48}
$$

We can reformulate Equation (48) by considering the number of observations up to a certain boundary point, rather than the number of observations in a particular class. That is checking for

$$
\Pr\left(\mathbb{E}\left[\sum_{i=1}^{b} N_i^{WS}\right] > 0\right) \to 1\,.
$$

This gives the possibility to replace $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x$ with $\int_{c_0^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x$. Since this is a CDF, and hence a non-decreasing function, which is effectively showing that each class has non-empty observations, we can write the following:

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{b} N_i^{WS}\right) =& \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i < c_b^{WS}\}}\right) \\
=& N \Pr(x \in \mathcal{S}_s) \int_{c_0^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x.
\end{aligned}
$$

Next, we need to show that this is an increasing function in C_b^{WS}. Now as $N \to \infty$, under the assumption that $\Pr(x \in \mathcal{S}_s) = 1/S$ and $S/N \to \mathrm{c}$ with $\mathrm{c} \in (0,1)$ (this is satisfied when $S = cN$)

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{b} N_i^{WS}\right) =& N \Pr(x_i < C_b^{WS}) \\ =& \frac{1}{\mathrm{c}} \int_{C_0^{WS}}^{C_b^{WS}} f(x)dx.\end{aligned}$$

Note that the derivative with respect to C_b^{WS} is $\frac{1}{\mathrm{c}} f\left(C_b^{WS}\right) > 0$, so the expected number of observations in each class is not 0. This completes our proof.

C.2 Proof of Proposition 2

Recall

$$\Pr\left(x \in C_b^{WS}\right) = \sum_{s=1}^{S} \Pr(x \in \mathcal{S}_s) \sum_{m=1}^{M} \Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,. \quad (49)$$

As $S \to \infty$, $\exists c_b^{WS} = c$ for any $c \in (a_l, a_u)$, by construction. Furthermore, for any c_b^{WS}, $\exists l \in [1, S]$, $m \in [1, M]$ such that $c_b^{WS} = c_m^{(l)}$. Also note that as $S \to \infty$, we need $N \to \infty$ as well. Now consider $\Pr(x^\dagger < c_b^{WS}) = \Pr(x^\dagger < c_m^{(l)})$, given $\Pr(x \in \mathcal{S}_s) = 1/S$ and using equation (49) gives

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

Note that the summation over the different classes in Equation (49) is being replaced as we are considering the cumulative probability and that no value greater than $c_m^{(l)}$ will be used as a candidate in the working sample for c_b^{WS}. Under the shifting method, $c_m^{(s)} \leq c_m^{(l)}$ for $s < l$ and using the definition of conditional probability gives

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

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

$$\begin{aligned}\Pr(x^\dagger < c_m^{(l)}) =& \frac{S-l-1}{S}\Pr(x < c_m^{(l)}) + \frac{1}{S}\sum_{s=1}^{l}\Pr(x < c_m^{(s)}) \\ \lim_{S\to\infty}\Pr(x^\dagger < c_m^{(l)}) =& \Pr(x < c_m^{(l)}).\end{aligned}$$

This completes the proof.

C.3 Speed of Convergence for the Shifting Method

Recall

$$\Pr\left(x_i^\dagger \in C_b^{WS}\right) = \begin{cases} 0, & \text{if } s = 1 \text{ and } m = 1, \\ \frac{1}{S}\sum_{s=2}^{S}\frac{1}{s-1}\int_{C_1^{(s)}|C_b^{WS}\in C_1^{(s)}} f(x)\mathrm{d}x, & \text{if } s \neq 1 \text{ and } m = 1, \\ \frac{1}{S^2}\sum_{s=1}^{S}\int_{C_m^{(s)}|C_b^{WS}\in C_m^{(s)}} f(x)\mathrm{d}x, & \text{if } 1 < m < M, \\ \frac{1}{S}\sum_{s=1}^{S}\frac{1}{S-s+1}\int_{C_M^{(s)}|C_b^{WS}\in C_M^{(s)}} f(x)\mathrm{d}x, & \text{if } m = M\,. \end{cases} \tag{50}$$

For each of the conditions in Equation (50), the corresponding expression is $o(1)$. To see this, note that $f(x)$ is a density, so the integral is less than 1. First, consider the case of $s \neq 1$ and $m = 1$,

$$\begin{aligned}\frac{1}{S}\sum_{s=2}^{S}\frac{1}{s-1}\int_{C_1^{(s)}|C_b^{WS}\in C_1^{(s)}} f(x)\mathrm{d}x, &\leq \frac{1}{S}\sum_{s=2}^{S}\frac{1}{s-1} \\ &= \frac{1}{S}\sum_{s=1}^{S}\frac{1}{s} \\ &= \frac{1}{S}\int_1^S \frac{1}{s}ds \\ &= \frac{\log S}{S}.\end{aligned}$$

As $S \to \infty$, the ratio in the last line goes to 0. This is expected if the widths of the classes in the working sample go to zero. This is straightforward, while the probability that an observation belongs to a point is 0. The same derivations applies

to the case when $m = M$. Now, consider the case of $1 < m < M$,

$$\frac{1}{S^2}\sum_{s=1}^{S}\int_{C_m^{(s)}|C_b^{WS}\in C_m^{(s)}} f(x)\mathrm{d}x \leq \frac{1}{S^2}\sum_{s=1}^{S} 1 = \frac{1}{S},$$

which also converges to 0 as $S \to \infty$, but at a faster rate than in the previous cases.

D Summary of the Notation Used in the Paper

Scalars:
N – number of individuals
T – number of time period (panel case)
a_l – lower boundary point for distribution's ($f(\cdot)$) support
a_u – upper boundary point for distribution's ($f(\cdot)$) support
μ or μ_i – first moment for distribution $f(\cdot)$ or $f_i(\cdot)$
M – number of possible choice values for a questionnaire
z_m – choice value of class m
c_m – m'th class's lower boundary point
β – parameter for DOC (x) variable
γ– parameter for control (w) variables
K – number of DOC (x) variables (matrix notations)
J – number of control variables (matrix notations)
B – number of working sample classes
S – number of split samples
$N^{(s)}$ – number of observations in split sample s
$z_m^{(s)}$ – choice value of class m in split sample s
$c_m^{(s)}$ – s'th split sample, m'th class's lower boundary point
c_b^{WS} – working sample b'th class's lower boundary point
h – working sample's class widths
Δ – size of shift for the shifting method

Running indexes
i – refers to individual $i = 1, \ldots, N$, and in some places it is a running index.
t – refers to time $t = 1, \ldots, T$
m – refers to class $m = 1, \ldots, M$
k – refers to a DOC variables in matrix formulation, $k = 1, \ldots, K$
j – refers to a control variables in matrix notation, $j = 1, \ldots, J$, and in some places

it is a running index.
b – working sample classes, $b = 1, \ldots, B$
s – split sample index
i_m – running index, where m is the indication in which class that observation is (M consistency)
m_i – i-th observation in the m-th class (M consistency)

Random variables
X or x – true, but unobserved variable with distribution $f(\cdot)$ (unknown)
X^* – discretized choice (DOC), with distribution $\psi(X)$ (observed)
$\hat{\beta}$ – parameter estimate for β with LS (estimate)
$\hat{\gamma}$ – parameter estimate for γ with LS (estimate)
$\bar{x}$ – sample average of the underlying variable x (not observed)
$\bar{x}^*$ – sample average of the observed discretized variable x^* (estimate)
x^{WS} – working sample (concept)
$\hat{\pi}_\chi$ – replacement estimator for non-directly transferable observations (estimate)
y^{tr}, x^{tr} – artificially truncated variables of the original r.v. (concept)
$\hat{\pi}_\tau$ – replacement estimator for shifting method (estimate)

Individual observations of random variables
x_i – true choice values for individual i (not observed)
x_i^* – discretized choice values (DOC) for individual i (observed)
y_i – outcome variable's values for individual i (observed)
w_i – control variable's values for individual i (observed)
ϵ_i – model disturbance term
u_i – idiosyncratic disturbance term for DGP (not observed)
N_m – number of observations in class m (observed)
ξ_i – error due to discretization $\xi_i = x_i - x_i^*$ (not observed)
υ^m – conditional distribution for errors of class m, formally: $\upsilon^m \stackrel{d}{=} \xi_i | C_m$ (not observed)
x_m – conditional distribution for x_i within class m, formally: $x_m \stackrel{d}{=} x_i | C_m$ (not observed)
x^m – sum of the true observed values in class m, formally: $x^m = \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\} x_i}$ (not observed)
ξ^m – sum of the errors in class m, formally: $\xi^m = \sum_{i=1}^{N} \mathbf{1}_{\{\xi_i \in C_m\} \xi_i}$ (not observed)
$x_i^{(s)}$ – discretized choice values (DOC) for individual i in split sample s (observed)
N^{WS} – number of observations in the working-sample (observed)
$N_m^{(s)}$ – number of observations in split sample s in class $C_m^{(s)}$ (observed)
x_i^{WS} – working-samples DOC observations (observed)
$x_{i,DTO}^{WS}$ – magnifying method's working sample, constructed by only the directly transferable observations (observed)

N_{DTO}^{WS} – number of observations in the magnifying method's 'DTO' working sample. (observed)
$x_{i,NDTO}^{WS}$ – magnifying method's working sample, constructed by only the directly transferable observations (observed)
η_i – error component from models to get $\hat{\pi}_\chi$ or $\hat{\pi}_\tau$ (observed)
$x_i^\dagger$ – artificial variable created during the shifting method (constructed)
$x_{i,Shifting}^{WS}$ – shifting method's working sample (constructed)

Functions
$f(\cdot)$ – probability distribution function
$\psi(\cdot)$ – discretization function $\psi(x_i) = x_i^*$
$\mathbf{1}_{\{\cdot\}}$ – indicator function, which takes 1 if the condition in the subscript is true, otherwise 0
$F(\cdot)$ – cdf of x
$U(\cdot)$ – Uniform distribution
$\psi^{(s)}(\cdot)$ – discretization function for split sample s
$\Psi(\cdot)$ – merging function
$||\cdot||$ – width of a class (or euclidean distance)
$Z(s, m)$ – set 'creator' function: given a split sample class, creates a set of choice values, which lies in the interval of the working-sample
$\mathcal{F}^\dagger$ – assign choice values from $Z(s, m)$ to each observation $x_i^{(s)} \in C_m^{(s)}$, with a given (uniform) probability
$\mathcal{F}^{WS}$ – assign estimated values $\hat{\pi}_\tau$ to each observation $x_i^{(s)} \in C_m^{(s)}$

Intervals
C_m – m'th class
$C_m^{(s)}$ – s split sample's, m'th class
C_b^{WS} – working sample's, b'th class

Sets
$\mathcal{S}_s$ – s'th split sample
ζ – set of classes, which contains the directly transferable observations
C_χ – set of classes, which contains the non-directly transferable observations
ζ^{tr} – ζ without the first and last class
$\mathcal{A}_m^{(s)}$ – set for observations $x_i^{(s)}$ which are in class $C_m^{(s)}$

Matrix notations
$\mathbf{y}$ – $y_i, N \times 1$
$\mathbf{X}$ – $(x_{1,i}, \ldots, x_{k,i}, \ldots, x_{K,i}), N \times K$
$\mathbf{W}$ – $(w_{1,i}, \ldots, w_{j,i}, \ldots, w_{K,i}), N \times J$
$\boldsymbol{\varepsilon}$ – $\epsilon_i, N \times 1$

$\boldsymbol{\beta}$ – $\beta_k, K \times 1$
$\boldsymbol{\gamma}$ – $\gamma_j, J \times 1$
$\mathbf{z}_k$ – $(z_{1,i}, \ldots, z_{m,i}, \ldots, z_{M,i}), 1 \times M$
$\mathbf{Z}$ – $\text{diag}(\mathbf{z}_{1,i}, \ldots, \mathbf{z}_{k,i}, \ldots, \mathbf{z}_{K,i}), MK \times K)$
$\mathbf{e}_{ki}$ – is the indicator vector for k'th DOC variable
$\mathbf{E}$ – matrices for the indicator vectors, $MK \times N$
$\mathbf{X}^* = \mathbf{E}'\mathbf{Z}$
$\mathbf{M}_W$ – residual maker
$\mathbf{q}_{kl}$ – typical block element in $\mathbf{E}\mathbf{E}'$
Ω – region for integration $[c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}]$
a_{kl} – auxiliary variable for $\mathbf{Z}'\mathbf{E}\mathbf{X}$
$\Omega_{\mathbf{X}}$ – sample space of x_k and x_l
ω_{ij} – (i, j) element in $\mathbf{M}_W$
g_{kl} – auxiliary variable for proof Eq. 26
h_{kl} – auxiliary variable for proof Eq. 28
u_{li} – auxiliary variable for proof Eq. 28
Panel
$\boldsymbol{\beta}_W$ – within estimator for panel
$\mathbf{D}_N$ – individual fixed effect
$\mathbf{M}_{D_N}$ – panel projection matrix
split sampling
$\hat{\boldsymbol{\pi}}_\chi$ – vector of replacement estimator for magnifying method
$\boldsymbol{\Omega}\chi$ – asymptotic standard errors for $\hat{\boldsymbol{\pi}}_\chi$
$\hat{\boldsymbol{\pi}}_\tau$ – vector of replacement estimator for shifting method
$\boldsymbol{\Omega}_\tau$ – asymptotic standard errors for $\hat{\boldsymbol{\pi}}_\tau$