Modeling with Discretized Ordered Choice Covariates

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

November 19, 2018

* Curtin University
** Central European University

Time: 1:38pm

Work in progress please do not quote!

Abstract

This paper deals with econometric models where some (or all) explanatory variables (or covariates) are observed as discretized ordered choices. Such variables are in theory continuous, but in this form are not observed at all, their distribution is unknown, and instead only a set of discrete (menu type) choices are observed. We explore how such variables influence inference, more precisely, we show that this leads to a very special form of measurement error, and consequently to endogeneity bias. We then propose appropriate sub-sampling and instrumental variables (IV) estimation methods to deal with the problem.

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 menu choices, i.e., by picking one option from a pre-set menu/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. Such practice is also common for major cities like London, Sydney and Hong Kong. Usually the reported values are discretized version of variables, like average personal distance travelled, or use of public or private transportation (Santos et al., 2011). Also in transportation research, the use of Likert-scale type data on intentions or attitudes is quite common, such as data from question on the likelihood of utilising certain transportation mode (see Heath and Gifford (2002)). In happiness economics variables are also often measured with Likert-scale data (see Frey and Stutzer (2002) or Stutzer (2004)). Such examples are also common in many other areas, like credit ratings in financial economics, corruption measures or institutional development in political economy. These are such discrete variables which have the characteristics of menu choices (see

Mauro (1995) and Méndez and Sepúlveda (2006) or 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 to download this application...?) and questions related to time (e.g., how much time did you spend last week commuting...?). The main question we try to investigate in this paper is how this may affect inference in an econometric model, when such variables are used as explanatory variables or covariates.

Consider $x_{it} \sim D_i(0,100)$ where $D_i(a,b)$ denotes a distribution with support in $[a,b]$ with mean μ_i for $i = 1,\ldots,N$. It is also assumed that it is stationary so the distribution may change over individual i but not over time, t. Also, quite importantly, the distribution $D_i(\cdot)$ is unknown (and can be continuous or discrete). Furthermore, define

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

Variable z_m, $m = 1,\ldots,M$ can be a measure of centrality of the given menu, or can be a completely arbitrarily assigned value (say, for example, if we consider preferences, etc.). The threshold c_m can be known, unknown or in some cases it can also be stochastic. For simplicity we may also refer to each menu point as class. The main difficulty is that instead of x_{it} we only observe x^*_{it}. Continuous variable x is in fact observed through the discrete ordered window of x^*_{it}.

2 Basic Setup

Let us assume that we have an econometric model of the form

$$y_{it} = g(w'_{it}\gamma + x^{*'}_{it}\beta) + \varepsilon_{it} \tag{2}$$

with the true Data Generating Process (DGP) being:

$$y_{it} = g(w'_{it}\gamma + x'_{it}\beta) + u_{it} \tag{3}$$

where $i = 1,\ldots,N$, $t = 1,\ldots,T$, w is a set of "usual" explanatory variables, x^* is a set of menu choice variable as defined in (1), γ and β are unknown parameters, $g(.)$ is a known linear or non-linear function, and u_{it} is an idiosyncratic disturbance term for model (3) with ε_{it} being its perceived counterpart in model (2). We also maintain the independence of observations across individuals assumption. The main question is therefore how estimating model (2) differs to estimating model (3).

Remark: If $D_i(.)$ is known, the expected value of each variable in x^* is therefore also know in each menu/class and have an unbiased/consistent estimate, then the LS estimator of model (2) is unbiased/consistent. This is in fact the Berkson model (see Berkson (1980) and Wansbeek and Meijer (2000) pp. 29-30).

2.1 An Example

Let us assume that we would like to model in a given city the factors explaining individual Transport Expenditures (TE), over a period of time with the simple model

$$TE_{it} = w_{it}'\gamma + \beta\, UPT_{it} + \text{Fixed Effects} + \varepsilon_{it}\,, \tag{4}$$

where TE_{it} is the transport expenditure for individual i in period t, UPT_{it} is the use of public transport in commuting measured in percentage points: 100% if only PT was used and 0% if PT was not used at all, for individual i $(i = 1, \ldots, N)$ on day t $(t = 1, \ldots, T)$, and w_{it} are "usual" controls. If UPT is not observed and instead we observe only the individual's choice from a pre-set menu menu list, UPT^* in the following form:

$$UPT^*_{it} = \begin{cases} 1, & \text{if} \quad 90\% \leq UPT_{it} = 100\% \\ 2, & \text{if} \quad 50\% \leq UPT_{it} < 90\% \\ 3, & \text{if} \quad 10\% \leq UPT_{it} < 50\% \\ 4, & \text{if} \quad 0\% \leq UPT_{it} < 10\%, \end{cases} \tag{5}$$

where the menu choices are:

$$\begin{array}{rl} 1 \rightarrow & \text{took almost only public transport} \\ 2 \rightarrow & \text{took mostly public transport} \\ 3 \rightarrow & \text{mostly did not take public transport} \\ 4 \rightarrow & \text{almost did not take public transport} \end{array} \tag{6}$$

Or alternatively, one could assign the mid value of each menu to UPT^*_{it} such that

$$\begin{array}{rl} 1 \rightarrow & \text{took almost only public transport} \\ 0.75 \rightarrow & \text{took mostly public transport} \\ 0.25 \rightarrow & \text{mostly did not take public transport} \\ 0 \rightarrow & \text{almost did not take public transport} \end{array} \tag{7}$$

and so

$$UPT^*_{it} = \begin{cases} 1, & \text{if} \quad 90\% \leq UPT_{it} \leq 100\% \\ 0.75, & \text{if} \quad 50\% \leq UPT_{it} < 90\% \\ 0.25, & \text{if} \quad 10\% \leq UPT_{it} < 50\% \\ 0, & \text{if} \quad 0\% \leq UPT_{it} < 10\%. \end{cases} \tag{8}$$

2.2 Related Work

To the best of our knowledge there has been no study investigating the estimation of categorized variable(s), when the categories/classes are not represented by the expected values of the underlying distribution(s). There has been though some work done on related issues. Taylor and Yu (2002) consider a regression model with three multivariate normal random variables. The first is linearly dependent on the second one. Then they dichotomize this second one and include into the model an another variable as well and derive the asymptotic bias for its parameter. However, they do not connect this to the bias in the parameter of the other variable(s). 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 class/menu choice values are properly representing each class/menu. 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/menu 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)).

3 Some Theory: Bias of the OLS Estimator

Let us assume for simplicity that $g(\cdot)$ in (2) is linear and that there is only one explanatory variable in the model, which is observed through menu choices. This case covers the issue, when the respondent knows the exact value of his or her characteristics, but because of the design of data gathering, the researcher makes the respondents choose a class, which is later assigned to some class value. It is also assumed, as said earlier, that it has a known support $[a, b]$ with known boundaries, and let z_m from equation (1) be the class/menu midpoint.[1] Taking specifically the classes and the class midpoints do not alter the results of the paper, we use them for illustration purposes.
Our classes/menus are the following with their respective class values:

$$\begin{aligned} C_1 &= \left[a, a + \frac{b-a}{M}\right) & z_1 &= a + \frac{b-a}{2M} \\ &\vdots \\ C_m &= \left[a + (m-1)\frac{(b-a)}{M}, a + m\frac{b-a}{M}\right) & z_m &= a + (2m-1)\frac{b-a}{2M} \\ &\vdots \\ C_M &= \left[a + (M-1)\frac{(b-a)}{M}, a + M\frac{b-a}{M}\right] & z_M &= a + (2M-1)\frac{b-a}{2M}. \end{aligned} \tag{9}$$

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 distribution pdf $f(\cdot)$ and cdf $F(\cdot)$,

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

[1] 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.

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

The standard OLS estimation is given by

$$\begin{aligned}
\hat{\beta}^*_{OLS} &= \left(x^{*'}x^*\right)^{-1}\left(x^{*'}y\right) \\
&= \frac{z_1\left(\sum_{i=1}^{N_1} y_i\right) + z_2\left(\sum_{i=N_1+1}^{N_1+N_2} y_i\right) + \cdots + z_M\left(\sum_{i=N-N_M+1}^{N_M} y_i\right)}{N_1 z_1^2 + N_2 z_2^2 + \cdots + N_M z_M^2} \\
&= \frac{z_1\left(\sum_{i=1}^{N_1} \beta x_i + u_i\right) + \cdots + z_M\left(\sum_{i=N-N_M+1}^{N_M} \beta x_i + u_i\right)}{N_1 z_1^2 + \cdots + N_M z_M^2} \\
&= \frac{z_1\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_1\}}(\beta x_i + u_i)\right] + \cdots + z_M\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_M\}}(\beta x_i + u_i)\right]}{N_1 z_1^2 + \cdots + N_M z_1^2} \\
&= \frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \qquad (10) \\
&= \frac{\sum_{m=1}^{M}\left[a + (2m-1)\frac{b-a}{2M}\right]\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m\left[a + (2m-1)\frac{b-a}{2M}\right]^2}.
\end{aligned}$$

Using equation (10), we can get the following general formula for the expected value of the OLS estimator

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) &= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}\left(\beta(x_i^* + \xi_i) + u_i\right)\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m\left[\beta\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^* + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i\right) + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^*}{\sum_{m=1}^{M} N_m z_m^2}\right\} + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&\qquad + \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m \xi_m}{\sum_{m=1}^{M} N_m z_m^2}\right\}, \qquad (11)
\end{aligned}$$

where the researcher makes the respondents cause an error ξ_i for each observation by setting the possible answer values at x_i^*, $x_i = x_i^* + \xi_i$. The last but one assertion in equation (11) is based on the disturbance term u_i being independent of the intended regressor x_i and $\mathbb{E}(u_i) = 0$ for all $i = 1, \ldots, N$. The last inference uses that the errors ξ_i have the same conditional distribution over the class C_m, $\xi_m \overset{\mathrm{d}}{=} \xi_i|C_m$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. Importantly, the second term in the expression (11) does not vanish in general, since $\xi_m|C_m$ is not independent of $N_m|C_m$, $\xi_m|C_m \not\perp\!\!\!\perp \mathbb{N}_m|C_m$ (see figure (1), right panel) nor $\mathbb{E}(\xi_i|C_m) = \mathbb{E}(\xi_m) = 0$

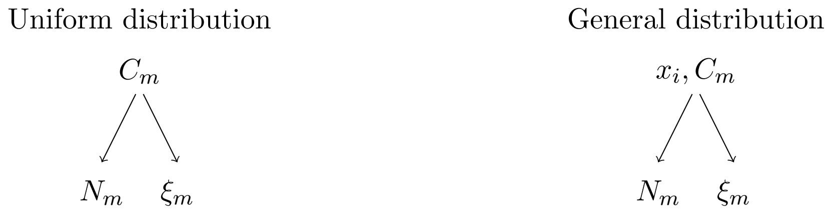


Figure 1: The dependence of the number of observations N_m and the error ξ_m on the distribution of the regressor variable x_i and the class C_m in uniform (left panel) and general distribution (right panel) cases

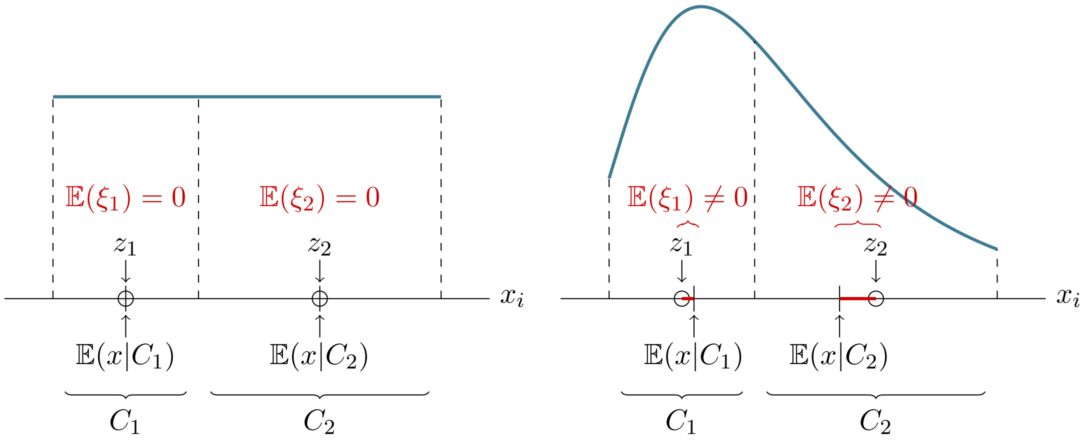


Figure 2: The difference between uniform (left panel) and general distributions (right panel)

(see figure (2), right panel). These are the sufficient assumptions for the OLS to be unbiased estimator. 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 the error are independent of each other, $N_m | x_i, C_m \perp\!\!\!\perp \xi_m | x_i, C_m$, but knowing the underlying distribution makes the problem trivial. Nonetheless, because of both issues, the naive OLS estimator is biased.

However, the uniform distribution turns out to be a special case. Let us assume that $x_i \sim U(a, b)$ for all $i = 1, \ldots, N$, both the issues disappear (see the left panels on figure (1) and figure (2)). The former issue is resolved, because in case of the uniform distribution, both the number of observations N_m in each class C_m and the error term ξ_m are independent of the regressor's x_i distribution, while the latter issue does not appear trivially, since in this case, the class midpoints are proper estimates of the regressor's x_i expected value on the class C_m. From equation (11) we obtain that

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

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

Now turning back to equation (10), but instead to taking the expectation let us see what happens in the probability limit, when the sample size or the number of classes go to infinity. 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 menu/class selection is independent of the disturbance terms and also that with sample size N the number of menus/classes M is fixed.

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\hat{\beta}^*_{OLS} &= \operatorname*{plim}_{N\to\infty}\frac{\sum_{m=1}^{M} z_m\left[\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(\beta x_i+u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2}\\
&= \frac{\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(\beta x_i+u_i)\right]}{\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m}\\
&= \frac{\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\beta\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i\right]}{\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m}\\
&= \frac{\beta\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i\right]}{\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m},
\end{aligned}\tag{12}$$

where 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 $m=1,\dots,M$, $x^m=\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i$, therefore its asymptotic distribution can be calculated by applying the Lindeberg-Levy Central Limit Theorem,

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

$\hat{\beta}^*_{OLS}$ is consistent if and only if the probability limit in equation (12) equals β. To give a condition for consistency, first we rewrite the previous equation (12) in terms of the error terms ξ_i,

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\hat{\beta}^*_{OLS}-\beta &= \frac{\beta\left(\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}x_i\right]-\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m\right)}{\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m}\\
&= \frac{\beta\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}(x_i-x_i^*)\right]}{\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m}\\
&= \frac{\beta\sum_{m=1}^{M} z_m\left[\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}\mathbf{1}_{\{x_i\in C_m\}}\xi_i\right]}{\sum_{m=1}^{M} z_m^2\operatorname{plim}_{N\to\infty}N_m},
\end{aligned}\tag{13}$$

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

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

After substituting back to the expression (13), we get

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} \hat{\beta}^*_{OLS} - \beta &= \frac{\operatorname{plim}_{N\to\infty} \beta \sum_{m=1}^{M} z_m \xi^m}{\operatorname{plim}_{N\to\infty} \sum_{m=1}^{M} z_m^2 N_m} \\ &= \frac{\operatorname{plim}_{N\to\infty} O(N) \beta \sum_{m=1}^{M} z_m \xi^m / N_m}{\operatorname{plim}_{N\to\infty} O(N) \sum_{m=1}^{M} z_m^2} \\ &= \frac{\beta \sum_{m=1}^{M} z_m \operatorname{plim}_{N\to\infty} \xi^m / N_m}{\sum_{m=1}^{M} z_m^2} \\ &= \frac{\beta \sum_{m=1}^{M} z_m \left\{\mathbb{E}(x_m) - z_m\right\}}{\sum_{m=1}^{M} z_m^2}, \end{aligned} \tag{14}$$

where 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

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

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 to its expected value. Therefore, to obtain the (in)consistency of the OLS estimator $\hat{\beta}^*_{OLS}$ 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 (14) 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 (14) zero, hence the OLS estimator is consistent.

Notice that, the consistency of the OLS estimator is not guaranteed even in case of symmetric distributions and symmetric class boundaries. After appropriate transformations (e.g., demeaning), it can be see 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 3).

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 (14). As a result, the differences between the expectation and the class midpoints in general are not zero for all m, hence the formula cannot be made arbitrarily small. Therefore, the OLS estimator becomes inconsistent in N (see the simulation results in the Appendix).

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

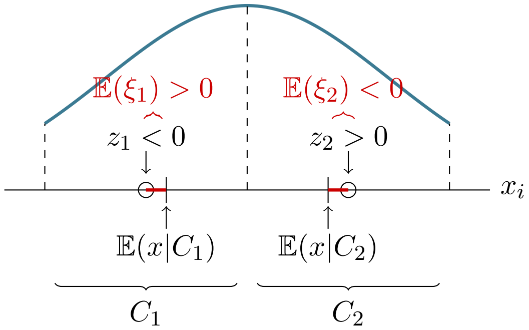


Figure 3: The estimator is inconsistent even in case of symmetric distributions, see equation (14)

of classes, the size of the classes itself are likely to shrink and become so narrow that only one observation can fall into each of the classes. In the limit we are going to hit the observations with the class midpoints (boundaries). To see that, we derive the consistency formula in the number of classes M assuming that $\text{plim}_{M\to\infty} \sum_{\{m:C_m\neq\emptyset,m=1,\dots,M\}} z_m u_{i_m} = 0$, 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 $\text{plim}_{N\to\infty} \sum_{i=1}^N x_i u_i = 0$.

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

where the index $i_m \in \{1,\dots,N\}$ denotes observation i in class m (at the beginning there might be several observation i that belong to the same class m), but index $m_i \in \{1,\dots,M\}$ denotes the class m that contains observation i (at the and of the derivation one class m includes only one observation i). Notice that the derivation do 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.

Remark: The above results hold for much simpler cases as well. If instead of model (2) we just take the simple sample average of x, $\bar{x} = \sum_i x/N$ (when $t = 1$), then $\bar{x}^* = \sum_i x_i^*/N$ is going to be a biased and inconsistent estimator of $\bar{x}$.

The measurement error due to menu 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 (2).
Consider a simple example where there is an unobserved variable x_i with an observed menu choice version:

$$x_i^* = \begin{cases} z_1 & \text{if } c_0 \leq x_i < c_1 \\ z_2 & \text{if } c_1 \leq x_i < c_2 \end{cases} \tag{15}$$

and

$$y_i = x_i\beta + \varepsilon_i. \tag{16}$$

Using the menu choice variable means:

$$y_i = x_i^*\beta + (x_i - x_i^*)\beta + \varepsilon_i \tag{17}$$

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 does not include an intercept. This result generalizes naturally to variable with multiple menu choice values.
These results about the behaviour of the OLS estimator are summarized in the following table:

$x \sim$ truncated	Biasedness	Consistency in N	Consistency in M
$U(a,b)$	✓	✓	✓
$N(\mu, \sigma^2)$	✗	✗	✓
$Exp(\lambda)$	✗	✗	✓
$Weibull(\lambda, k)$	✗	✗	✓

Table 1: Biasedness and consistency of the OLS estimator $\hat{\beta}^*_{OLS}$

4 Estimation Reconsidered

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{18}$$

where $\mathbf{X}$ and $\mathbf{W}$ are $N \times K$ and $N \times K_1$ data matrices of the explanatory variables, respectively. $\mathbf{y}$ is a $N \times 1$ vector containing the data of the dependent variable and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of disturbance terms. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $K \times 1$ and $K_1 \times 1$ parameter vectors, respectively. $\mathbf{X}$ is not observed, only its menu choice version, $\mathbf{X}^*$ is observed. Define the $MK \times K$ matrix

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

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

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

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 $\mathbf{X}^* = \mathbf{E}'\mathbf{Z}$.

4.1 The OLS Estimator

Following from equation (18), consider the regression based on the observed data:

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

then the OLS 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 usual derivation shows

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\boldsymbol{\varepsilon}. \tag{19}$$

This implies OLS is unbiased if and only if $(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z})^{-1}\,\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{I}$. This allows us to address the following:

1. Investigate the bias analytically by examining the elements in $\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}$ and $\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}$.

2. Derive the conditions for IV by constructing an orthogonal matrix to $\mathbf{X} - \mathbf{E}'\mathbf{Z}$.

4.2 Bias of the OLS

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

$$\mathbf{M_W X} = \mathbf{X} \tag{20}$$

$$\mathbf{M_W X}^* = \mathbf{X}^*. \tag{21}$$

In other words, we assume independence between $\mathbf{W}$ and $\mathbf{X}$ as well as its menu choice version. This may appear to be a strong assumption but it does allow us to see what is going on a little better. We can relax this at a latter stage.
The OLS estimator in this case becomes:

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

The OLS is unbiased if $\left(\mathbf{Z'EE'Z}\right)^{-1}\mathbf{Z'EX} = \mathbf{I}$. Let's consider a typical element in $\mathbf{Z'EE'Z}$ first. Since $\mathbf{Z}$ is non-stochastic as it contains only all the pre-defined menu choice values, it is sufficient to examine $\mathbf{EE'}$.

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

Note that each entry in $\mathbf{E}$ is a vector, so $\mathbf{EE'}$ will result in a partition matrix which 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 $\mathbf{b}_{kl}$ be a typical block element in $\mathbf{EE'}$ then

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

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

$$\mathbb{E}\left(\mathbf{1}^{ki}_{m}\mathbf{1}^{li}_{n}\right) = \int_{\Omega} f(x_k, x_l)dx_k dx_l\,, \tag{22}$$

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 (22) under the usual WLLN.
Following similar method, let a_{kl} be the (k, l) element in $\mathbf{Z'EX}$ then

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

Now,

$$\begin{aligned}\mathbb{E}\left[\sum_{m=1}^{M} z_{km}\mathbf{1}_m^{ki}x_{li}\right] &= \sum_{m=1}^{M} z_{km}\mathbb{E}\left[\mathbf{1}_m^{ki}x_{li}\right] \\ &= \sum_{m=1}^{M} z_{km}\int_{\Omega_1} x_l f(x_k, x_l)dx_k dx_l \,,\end{aligned} \tag{23}$$

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 (23) under the usual WLLN.

In the case when equations (20) and (21) do not hold, the analysis becomes more tedious algebraically but it does not affect the result that OLS is biased. Recall equation (19), and let ω_{ij} be the (i, j) element in $\mathbf{M_W}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, K_1$, then following the same argument as above $\mathbf{EM_W E'}$ 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

$$g_{kl} = \sum_{j=1}^{N}\sum_{i=1}^{N} \omega_{ij}\mathbf{1}_m^{ki}\mathbf{1}_n^{li} \tag{24}$$

with its expected value being

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{\Omega} \omega_{ij} f\left(x_k, x_l, \mathbf{w}\right) dx_k dx_k d\mathbf{w} \tag{25}$$

where $\mathbf{w} = (w_1, \ldots, w_{K_1})$, $d\mathbf{w} = \prod_{i=1}^{K_1} dw_i$ 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 (25) is complicated. However, under the assumption that the integral in equation (25) exits, then $N^{-1}g_{kl}$ should converge to equation (25) 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 (25) to equation (22).

Again, following the same derivation as above, a typical element in $\mathbf{Z'EM_W X}$ is

$$h_{kl} = \sum_{m=1}^{M}\sum_{i=1}^{N} z_{km}\mathbf{1}_m^{ki}u_{li} \tag{26}$$

where $u_{li} = \sum_{\tau=1}^{N} \omega_{i\tau}X_{l\tau}$. 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

$$\sum_{m=1}^{M} z_{km}\int_{\Omega_m} u_l f(x_k, x_l, \mathbf{w})dx_k dx_l d\mathbf{w} \tag{27}$$

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 (27) to equation (23).

4.3 Extension to Panel Data

So far we have assumed that $t = 1$, that we are dealing with cross-sectional data. Next, let us see what changes if $t > 1$, i.e., when we have panel data at hand. Now the most important problem is identification. If the menu choice of an individual does not changes over the time periods covered, the individual effects in the panel and the parameter associated with the menu choice variable cannot be identified separately. The Within transformation would wipe it out the menu choice variable as well. When the menu choice does change overt time, but little, 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 solution is to have different menu choice classes (boundaries), for the different time periods as, for example, explained in the next section. After the appropriate Within transformation, the OLS can be applied with properties outlined in the previous sections and in the next.

The bias of the panel data Within estimator can be shown easily. Let us re-write equation (4.1) in a panel data context

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

from which 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

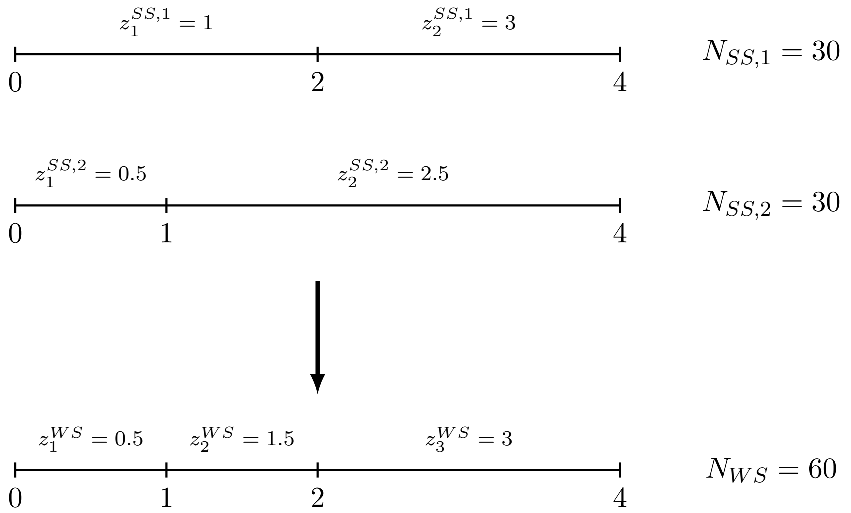
$$\mathbf{M}_{\mathbf{D}_N}\mathbf{y} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^*\boldsymbol{\beta} + \mathbf{M}_{\mathbf{D}_N}[(\mathbf{X} - \mathbf{X}^*)\boldsymbol{\beta} + \boldsymbol{\varepsilon}].$$

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

5 Consistent Estimation: Sub-sampling and Instrumental Variables

In this section, we propose two possible approaches to ensure consistent estimation, which later is extended to instrumental variables techniques as well. Most importantly, we depart from the classical econometric approach to estimation: we do not start assuming that the sample is given, but the main aim is to design an environment and sampling that delivers estimation with good enough precision. In other words we investigate what is the best method to gather the data, (what is the best survey design) to reduce the estimation bias presented earlier.

The main approach of the proposed methods is to create a number of sub-samples (B) using the same number (M) of menu choice classes in each sub-sample, but where the class boundaries are different. In fact, this approach utilizes the M consistency result previously discussed in section **??** and transforms it into N consistency, through using several sub-samples. The intuition behind the methods is that this leads to a better mapping of the unknown distribution of x and so reduce the estimation bias. By merging the different sub-samples into one data set (let us call this the working sample), we get M_{WS} overall number of menu choice classes (or bins) across the merged sub-samples, where M_{WS} is much larger than M. In a given sub-sample each respondent (individual i) is given one questionnaire (in the case of cross sectional data). The set of respondents who fill the questionnaire with the same class boundaries define a sub-sample. Each sub-sample has $N_{SS,j}$, $j = 1, \dots, B$ number of observations ($\sum_j N_{SS,j} = N$). In this setup a sub-sample looks exactly as the problem introduced above in the paper, and the only difference across sub-sample is that the menu choice (class) boundaries vary. Note that the number of observations across sub-samples can be the same or, more likely, different. Let us see a very simple illustrative example about this. Assume that $M = 2$, $B = 2$ as well, $N = 60$, $N_{SS,1} = 30$ and similarly $N_{SS,2} = 30$. Let x be a continuously distributed variable in the $[0, 4]$ domain and let the class boundaries in the first sub-sample be $[0, 2)$ and $[2, 4]$, while in the second sub-sample $[0, 1)$ and $[1, 4]$, with 10, 20, 5, and 25 observations respectively in each class. Next, let us merge the information obtained in the two sub-samples in one working sample. This will have 3 classes (or bins): $[0, 1)$, $[1, 2)$ and $[2, 4]$. From the 2nd sub-sample we know that with 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 the picture below). Clearly the working sample maps (slightly) better the unknown distribution of x than any of the two sub-samples.



Construction of the working sample can be done in many different ways. We propose that the class boundaries in the working sample is the union of the sub-samples' class boundaries:

$$\bigcup_{i=0}^{B} c_i^{WS} = \bigcup_{j=1}^{B} \bigcup_{i=0}^{M} c_i^{SS,j}$$

This will result that in our example $c_0^{WS} = c_0^{SS,1}, c_1^{WS} = c_1^{SS,2}, c_2^{WS} = c_1^{SS,1}, c_3^{WS} = c_2^{SS,1}$. The probability of an observation is in a sub-sample between given boundary points is,

$$\Pr\left(c_{m-1}^{SS,j} < x \leq c_m^{SS,j}\right) = \Pr(x \in j) \int_{c_{m-1}^{SS,j}}^{c_m^{SS,j}} f(x) \mathrm{d}x$$

The probability of an observation in the working sample between given boundary points, given the observed menu choice value in a given sub-sample is observed is:[2]

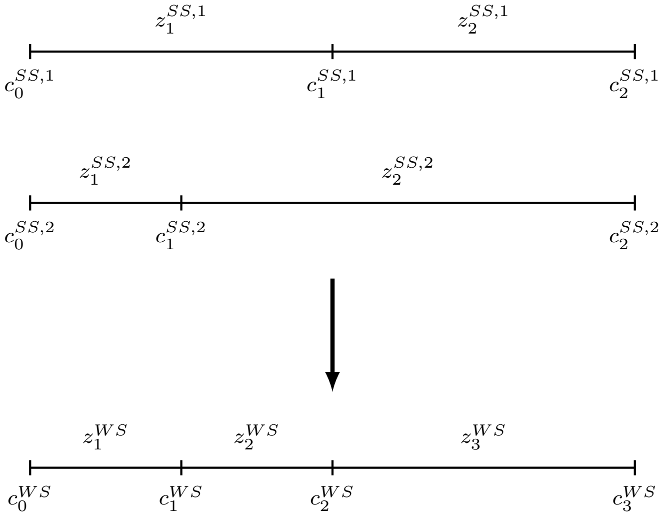
$$\Pr\left(c_{m-1}^{WS} < x \leq c_m^{WS} \mid c_{l-1}^{SS,j} < x \leq c_l^{SS,j}\right) = \begin{cases} \frac{c_m^{WS} - c_{m-1}^{WS}}{c_l^{SS,j} - c_{l-1}^{SS,j}}, & \text{if} \quad c_m^{WS} \leq c_l^{SS,j} \text{ and } c_{m-1}^{WS} \geq c_{l-1}^{SS,j} \\ 0, & \text{otherwise} \end{cases}$$

General case:

$$\Pr\left(c_{m-1}^{WS} < x \leq c_m^{WS} \mid c_{l-1}^{SS,j} < x \leq c_l^{SS,j}\right) = \begin{cases} \frac{c_m^{WS} - c_{l-1}^{SS}}{c_l^{SS,j} - c_{l-1}^{SS,j}}, & \text{if??} \\ \frac{c_m^{WS} - c_{m-1}^{WS}}{c_l^{SS,j} - c_{l-1}^{SS,j}}, & \text{if} \quad c_m^{WS} \leq c_l^{SS,j} \text{ and } c_{m-1}^{WS} \geq c_{l-1}^{SS,j} \\ \frac{c_m^{SS} - c_{m-1}^{WS}}{c_l^{SS,j} - c_{l-1}^{SS,j}}, & \text{if??} \\ 0, & \text{otherwise} \end{cases}$$

From these we can easily get the uncondtional probability of an observation in the working sample between given boundary points:

$$\Pr\left(c_{m-1}^{WS} < x \leq c_m^{WS}\right) = \begin{cases} \sum_{j=1}^{B} \Pr(x \in j) \sum_{m=1}^{M} \frac{c_m^{WS} - c_{m-1}^{WS}}{c_l^{SS,j} - c_{l-1}^{SS,j}} \int_{c_{m-1}^{SS,j}}^{c_m^{SS,j}} f(x)\mathrm{d}x, & \\ \qquad \text{if} \quad c_m^{WS} \leq c_l^{SS,j} \text{ and } c_{m-1}^{WS} \geq c_{l-1}^{SS,j} & \\ 0, \qquad \text{otherwise} & \end{cases}$$

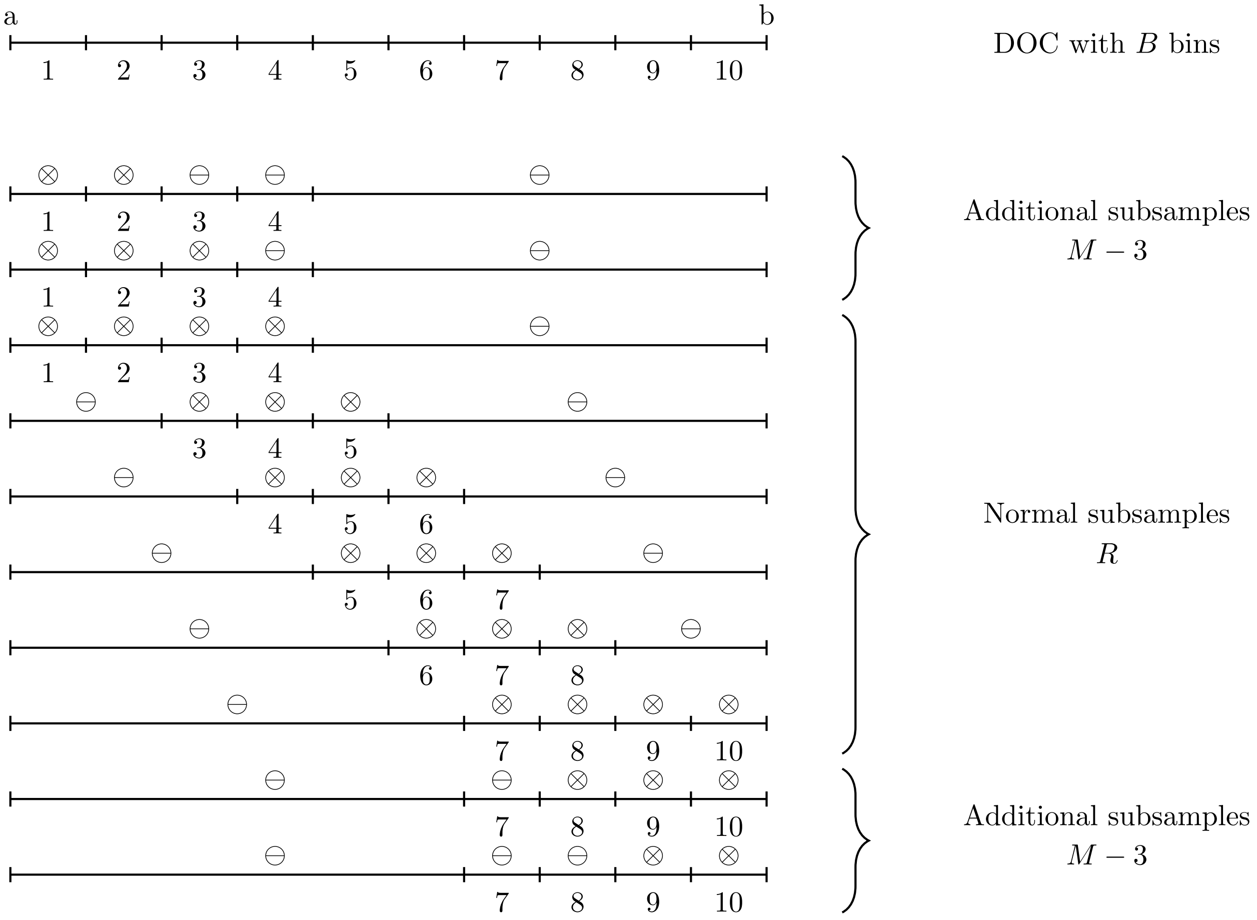


5.1 Magnifying Method

In case of the magnifying method, the researcher magnifies the domain of the answers within the original domain of the unknown distribution of x by one equally sized bin. The size of the bins are depending on the number of sub-surveys (R) used. As the number of sub-surveys increases the bin's size decrease, which will the main cause of exploring the underlying distribution. This method allows to magnify the domain of the underlying distribution without creating a survey where the number of discretized ordered choices overflow the survey. 4.

[2] This is not true for the general case when the working sample's boundary points can be anything.

figure shows the main idea of the magnifying method: the first line shows the union of sub-surveys, while below that one can see the individual sub-surveys.



⊗: Retained observations in each subsample

⊖: Dropped observations in each subsample

Figure 4: Magnifying method for create B descrete ordered choice values with multiple sub-samples

The number of magnified bins B in the US has a direct link to the number of survey types R, by design. One can easily see that, in this case, the number of overall magnified bins B the researcher uses in her study is given by

$$B = M - 1 + R,$$

intuitively, as M classes have $M-1$ common boundaries, they are magnified R times. We suggest that the number of choices falling into the same bin is the same across sub-samples (namely $M-2$). Therefore, in each sub-sample, some observations have to be dropped. The number of sub-samples S can also be easily given by

$$S = R + 2(M - 3),$$

intuitively, the initial R surveys should be augmented by $M-3$ additional surveys on either side in order to have $M-2$ classes of subsamples fall into the same bin.

First let us describe a sub-survey r, where $r = 1, \dots, R$. The discrete ordered choices given in a specific survey is $z_{r_1}, \dots, z_{r_M}$, where class boundaries are given by $c_{r_0}, \dots, c_{r_M}$. With magnifying method the sub-survey's binwidth is given by:

$$c_{r_i} - c_{r_{i-1}} = \frac{b-a}{B}$$

By design, the effective DOC's, which are retained during the modeling:

$$z_r^{eff} = \begin{cases} z_{r_1}, \dots, z_{r_{M-2}}, & \text{if} \quad r \leq M-3 \\ z_{r_2}, \dots, z_{r_{M-1}}, & \text{if} \quad M-3 < r < R+M-3 \\ z_{r_3}, \dots, z_{r_M}, & \text{if} \quad r \geq R+M-3 \end{cases}$$

and let call effective boundary points c_r^{eff}, the boundary points where the effective DOC's are observed, such as:

$$c_r^{eff} = \begin{cases} c_{r_0}, \dots, c_{r_{M-2}}, & \text{if} \quad r \leq M-3 \\ c_{r_1}, \dots, c_{r_{M-1}}, & \text{if} \quad M-3 < r < R+M-3 \\ c_{r_2}, \dots, c_{r_M}, & \text{if} \quad r \geq R+M-3 \end{cases}$$

As a second step we characterize the union of the sub-surveys and then show the links between individual and overall surveys. We have for US: $z_1^{US}, \dots, z_B^{US}$ DOC's, where $c_0^{US}, \dots, c_B^{US}$ are the boundaries.

Third step: number of effective observations (N_j^{eff}). As we have S subsamples and we distribute the number of observations equally among all surveys we have $n_s = N/S$ number of observations for each surveys. For each sub-samples the expected number of effective observations n_s^{eff} will depend on the actual underlying distribution, in the following way:

For each sub-survey the expected number of effective observations in a given bin:

$$\begin{aligned} \mathbb{E}(n_{r_j}^{eff}) &= \mathbb{E}\left(\sum_{i=1}^{n_s} \mathbf{1}_{x_i \in c_{r_j}^{eff}}\right) \\ &= n_s \int_{c_{r_j}^{eff}} f(x) \, \mathrm{d}x \\ &= n_s \Pr(c_{r_{j-1}}^{eff} < x \leq c_{r_j}^{eff}) \\ &= n_s \Pr(c_{r_{j-1}}, c_{r_j} \in c_r^{eff}) \Pr(c_{r_{j-1}} < x \leq c_{r_j}) \\ &= n_s \frac{M-2}{B} \Pr(c_{j-1}^{US} < x \leq c_j^{US}) \end{aligned}$$

For the sub-survey it is the sum of the effective bins:

$$\mathbb{E}(n_r^{eff}) = \mathbb{E}\left(\sum_{j=1}^{M-2} n_{r_j}^{eff}\right)$$

For the union of sub-surveys, it is the sum of individual sub-surveys, given the observations in each sub-samples are uniformly distributed:

$$\begin{aligned}\mathbb{E}(N_j^{eff}) &= \mathbb{E}\left(\sum_{r=1}^{R}\sum_{j=1}^{M-2} n_{r_j}^{eff}\right)\\ &= \frac{N(M-2)(M-3)}{SB}\Pr(c_{j-1}^{US} < x \leq c_j^{US})\end{aligned}$$

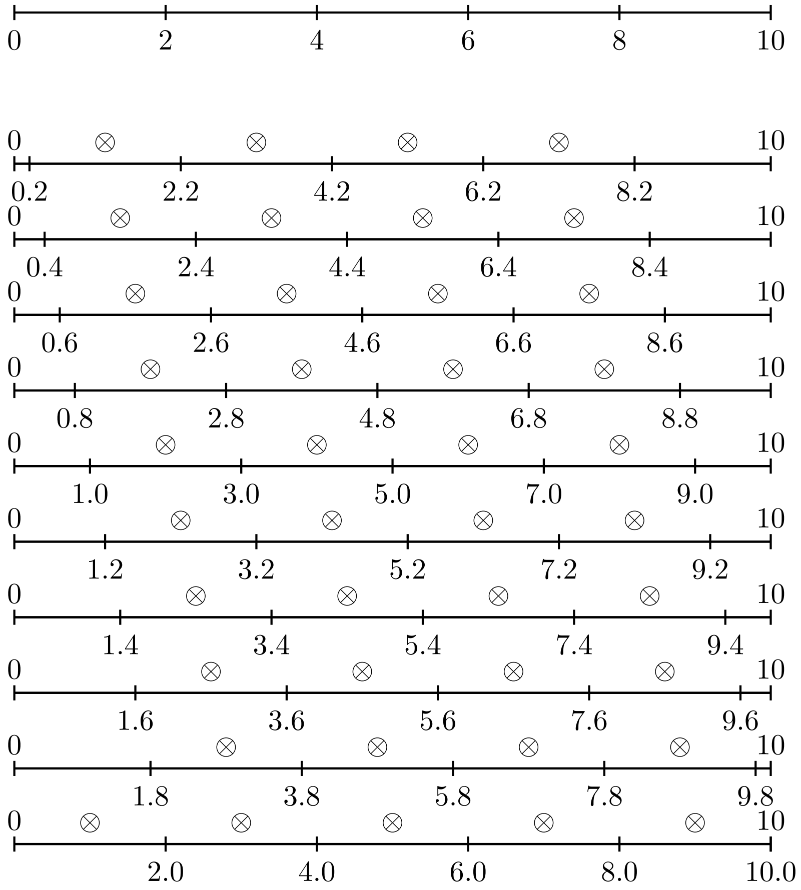
—— OLD PART NOT READY!!!! ——-
If the underlying distribution is censored we can retain or drop the boundary observations. This method can recover the underlying distribution as $S \to \infty$, without $M \to \infty$. However there are some issues with procedure to note:

- We may drop a large number of the observations: the surveys are censored in that sense they got menu choice values which has open thresholds, thus we need to drop those observations. Also we must drop some of the observations in the boundaries in order to preserve uniform frequencies along classes. This means in a simulation where we have $S = 100$, we may only retain $2 - 3\%$ of the original sample. This is not a big problem, when we are working with large dataset (eg., 100k or 500k), but causes problems with small number of observations.
- Formally, need N to increase faster than S in order to get observations in each survey: for each survey we have N/S observations, thus we need $S/N \to 0$ to have enough observation. (E.g.,: when we have $N = 10,000, S = 50, N/S = 200$)
- Observed menu choices also depend on the class width and the survey's observation: $S \to \infty$, thus $h = c_i^{(S)} - c_{i-1}^{(S)} \to 0$. As in the nonparametric literature, we need $N/S \to \infty$ as $h \to 0$ at a faster rate. This can cause some problems if we do not have large samples, e.g., in simulation with $N = 1000, S = 50$ we only observe 19 menu choice values, while for $S = 100$, this is only 9. ($\mathbb{P}\left[c_j^{\zeta_k} \leq x_{it} < c_{j+1}^{\zeta_k}\right]$ can be quite low, especially in the tails.)

Creating artificial observations:
While the number of dropped observation increases dramatically as we increase S, it may be a good option to fill/replace somehow the dropped values. One solution is to replace them with the expected value of the constructed distribution. In order to do this we need the expectation of the truncated distribution for each survey's dropped value, where the truncation depends on the menu choice value we dropped. E.g.,: if we are in the left boundary then we drop the (censored) largest menu choice value. We replace this menu choice value $x_{it}^{*,(\zeta_i)}(c_{M-1}^{(\zeta_i)} < x_{it})$ with the constructed distribution's conditional expectation: $\mathbb{E}\left[x_{it}^{*,(S)}|c_{M-1}^{(\zeta_i)} < x_{it}^{*,(S)}\right]$. The drawback of this method is $\mathbb{E}\left[x_{it}^{*,(S)}|c_{M-1}^{(\zeta_i)} < x_{it}^{*,(S)}\right] \neq \mathbb{E}\left[x_{it}|c_{M-1}^{(\zeta_i)} < x_{it}\right]$, thus we "pass on" the bias we make with S menu choice values to the replaced values. In practice we only recommend this method, if the assumed bias is expected to be small.

5.2 Shifting Method

The second method is to fix the class widths and push the thresholds along the support for different surveys. To do this one can define the effective number of surveys (EM), and the menu choice values in each surveys (M). Then one can define a benchmark sequence of menu choice values $z_1^{(bm)}, z_2^{(bm)}, \dots, z_M^{(bm)}$, and the thresholds $c_0^{(bm)}, c_1^{(bm)}, \dots, z_M^{(bm)}$. The class width is fixed at $h = c_{i+1}^{(bm)} - c_i^{(bm)}$. Let us define the push measure such: $\nu = \frac{h}{EM}$. Using this we can push the thresholds EM times, so the ζ_i survey will have one extra threshold such: $c_0^{(bm)}, c_1^{(bm)} + i\nu, \dots, c_{M-1}^{(bm)} + i\nu, z_M^{(bm)}$: so it has the benchmark's boundary thresholds and push them in-between. It results in an extra menu choice value, the menu choices between the boundaries are pushed, and in the boundaries they are changing in a specific way: $z_0^{(\zeta_i)}, z_1^{(bm)} + i\nu, z_2^{(bm)} + i\nu, \dots, z_{M-1}^{(bm)} + i\nu, z_M^{(\zeta_i)}$. Thus for a specific survey we will have $M + 1$ questions, not M.



⊗: retained observations in each surveys

Figure 5: Shifting method

In the simplest case, we create similar amount of surveys. The big advantage is that one does not drop a large amount of observations. (The exact number depends on the underlying distribution.) The cost is that we will observe menu choice values only between $z_1^{(bm)}$ and $z_M^{(bm)}$, which is independent of EM. This leads to artificially truncated data, and the observed menu choices are only increases between the boundary menu choices as EM increases.[3]

[3] An other possibility is to use different amount of subsamples for each surveys. Then let $n^{(bm)}$ the subsample's observation for the benchmark case. We need max $\left(\frac{hn^{(bm)}}{i\nu}, \frac{hn^{(bm)}}{(EM-i)\nu}\right)$ observation for the ζ_i survey. Then

5.3 Instrumental Variables Estimation

We can use the sub-sampling methods as IV's instead of replacing x_{it}^* observations. Now a researcher needs two survey questions in the case of cross-sectional data, one is the original question, which gives the menu choice variable (x_{it}^*) and an other, which will be the IV. Usually it is not practical to ask the same questions with different possible menu choices, but one may refer to different time periods/locations/taste/etc., where the underlying distribution is the same. For example in the case of shifting method, one can ask 'How much have you used public transport in *this* week?' '0-20%, 20-40%, 40-60%, 60-80%, 80-100%' as the original menu choices, with a second question: 'How much have you used public transport in *last* week?' '0-10%, 10-30%, 30-50%, 50-70%, 70-90%, 90-100%' for IV.
There are some possible alternative specification, which are out of the scope for this paper, but worth noting. One is, when the researcher asks a more realistic question for the IV, which depends on the question of the original menu choice response, like: 'How much more or less have you used public transportation in last week?', with answers such '20% less, 10% less or equal, 10% more or equal, 20% more'. This is more realistic, however it may induce an autoregressive process, which must be modeled for proper inference.

In the case of panel data similar methods can be used as those outlined above for cross-sections. This, however, may give us some additional flexibility. Sub-sampling now can be used as follows: for the magnifying method, one can randomize the surveys assigned to each individual, this way ensuring variation in the response.[4] For the shifting method one can ask each individual with randomly changing shifts.
With respect to the use IVs, one can ask the menu choice question at some t points and the IV question at some other t time points (the same question, same M, but different class limits). The assumption needed in this case is that, the questions are paired such a way that they are belonging to the same underlying distribution and of course even number of type periods are needed.
Finally, for repeated cross-section the same procedures can be applied as for panel data.

5.4 Simulation Results

Next, we show the performance of our recommended solutions. As it turns out from the simulations, magnifying method performs well, if we replace original menu choices with the sub-sampling values, but not that well when we use it as an IV. On the contrary, the shifting method performs poorly when it is substituted with the original menu choices, and performs quite well when used as an IV. Which is the best, depend on the underlying distribution. Also it turns out truncation or censoring does not matter much when choosing the proper adjustment method.

we retain $n^{(bm)}$ observations between the boundary menu choices, $\frac{hn^{(bm)}}{i\nu}$ for the left boundary and $\frac{hn^{(bm)}}{(EM-i)\nu}$ for the right. In this way we achieve that as $EM \to \infty$ we cover the whole support. The drawback is we may drop again lots of observations in each survey, and the same properties must hold as in magnifying method on the boundaries.

[4] With magnifying method dropping observation may generate a missing data problem.

		Magnifying method - used as x_{it}^*							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0182	-0.0004	-0.0093	0.0074	0.1341	-0.0058	0.0163	-0.0097
	N=100,000	-0.0185	-0.0046	-0.0009	-0.0041	0.1342	0.0006	0.0032	0.0045
	N=500,000	-0.0190	-0.0039	0.0011	-0.0015	0.1339	-0.0025	-0.0073	0.0040
absbias	**N=10,000**	0.0415	0.1186	0.2897	0.4159	0.1342	0.1403	0.3127	0.4462
	N=100,000	0.0208	0.0376	0.0883	0.1317	0.1342	0.0442	0.0948	0.1353
	N=500,000	0.0191	0.0161	0.0391	0.0625	0.1339	0.0211	0.0441	0.0613
se	**N=10,000**	0.0489	0.1476	0.3618	0.5278	0.0445	0.1765	0.3914	0.5627
	N=100,000	0.0163	0.0460	0.1117	0.1670	0.0137	0.0554	0.1206	0.1705
	N=500,000	0.0073	0.0200	0.0508	0.0765	0.0061	0.0262	0.0550	0.0789
numObs	**N=10,000**	10000	1250	208	102	10000	817	171	86
	N=100,000	100000	12501	2083	1019	100000	8169	1709	860
	N=500,000	500000	62501	10422	5106	500000	40830	8550	4304
		Shifting method - used as IV							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10000**	-0.0182	0.0074	0.0061	0.0066	0.2453	0.0174	0.0074	0.0087
	N=100000	-0.0185	0.0016	0.0004	0.0006	0.2443	0.0346	0.0237	0.0229
	N=500000	-0.0190	0.0019	0.0009	0.0009	0.2447	0.0308	0.0133	0.0134
absbias	**N=10000**	0.0415	0.1156	0.1229	0.1235	0.2453	0.4364	0.4401	0.4489
	N=100000	0.0208	0.0366	0.0379	0.0376	0.2443	0.1466	0.1384	0.1397
	N=500000	0.0191	0.0169	0.0168	0.0172	0.2447	0.0682	0.0647	0.0672
se	**N=10000**	0.0489	0.1457	0.1537	0.1550	0.0792	0.5511	0.5569	0.5620
	N=100000	0.0163	0.0456	0.0477	0.0477	0.0251	0.1805	0.1733	0.1755
	N=500000	0.0073	0.0208	0.0213	0.0219	0.0119	0.0817	0.0818	0.0832
$corr(x_{it}^*, z_{it})$	**N=10000**	1.0000	0.8282	0.8124	0.8105	1.0000	0.5661	0.5710	0.5716
	N=100000	1.0000	0.8283	0.8125	0.8105	1.0000	0.5663	0.5715	0.5721
	N=500000	1.0000	0.8283	0.8125	0.8105	1.0000	0.5660	0.5711	0.5717
numObs	**N=10000**	10000	6575	6288	6253	10000	1886	1837	1831
	N=100000	100000	65744	62878	62521	100000	18850	18361	18300
	N=500000	500000	328675	314332	312553	500000	94268	91793	91494

Table 2: $Exp\,[0.5]\,, Supp = [0, 1]$, **M=3**; BM: Benchmark, see Table 9.

		Magnifying method - used as x_{it}^*							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0811	-0.0075	-0.0049	-0.0049	-0.0563	-0.0088	0.0066	0.0154
	N=100,000	-0.0810	-0.0087	-0.0007	-0.0001	-0.0554	-0.0083	0.0003	0.0005
	N=500,000	-0.0811	-0.0085	-0.0004	-0.0005	-0.0557	-0.0078	0.0009	-0.0015
absbias	**N=10,000**	0.0811	0.0523	0.1376	0.2002	0.0563	0.0644	0.1407	0.1980
	N=100,000	0.0810	0.0183	0.0426	0.0612	0.0554	0.0210	0.0431	0.0615
	N=500,000	0.0811	0.0104	0.0194	0.0286	0.0557	0.0113	0.0187	0.0272
se	**N=10,000**	0.0224	0.0655	0.1730	0.2470	0.0226	0.0790	0.1744	0.2483
	N=100,000	0.0071	0.0212	0.0527	0.0764	0.0069	0.0254	0.0546	0.0757
	N=500,000	0.0033	0.0093	0.0245	0.0349	0.0030	0.0116	0.0239	0.0341
numObs	**N=10,000**	10000	1251	208	102	10000	1160	202	99
	N=100,000	100000	12504	2083	1021	100000	11588	2020	993
	N=500,000	500000	62497	10425	5110	500000	57908	10077	4947
		Shifting method - used as IV							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10000**	-0.0811	-0.0045	0.0034	0.0048	-0.0563	0.0002	-0.0168	-0.0192
	N=100000	-0.0810	-0.0054	0.0030	0.0041	-0.0554	-0.0030	-0.0189	-0.0210
	N=500000	-0.0811	-0.0049	0.0034	0.0043	-0.0557	-0.0010	-0.0177	-0.0196
absbias	**N=10000**	0.0811	0.0286	0.0297	0.0300	0.0563	0.0753	0.0759	0.0759
	N=100000	0.0810	0.0102	0.0099	0.0102	0.0554	0.0258	0.0289	0.0295
	N=500000	0.0811	0.0059	0.0050	0.0056	0.0557	0.0107	0.0184	0.0201
se	**N=10000**	0.0224	0.0352	0.0371	0.0374	0.0226	0.0945	0.0940	0.0938
	N=100000	0.0071	0.0115	0.0122	0.0122	0.0069	0.0319	0.0311	0.0307
	N=500000	0.0033	0.0053	0.0054	0.0054	0.0030	0.0134	0.0121	0.0123
$corr(x_{it}^*, z_{it})$	**N=10000**	1.0000	0.7406	0.7209	0.7185	1.0000	0.5018	0.5074	0.5083
	N=100000	1.0000	0.7407	0.7209	0.7185	1.0000	0.5017	0.5075	0.5083
	N=500000	1.0000	0.7407	0.7210	0.7185	1.0000	0.5019	0.5075	0.5085
numObs	**N=10000**	10000	8716	8546	8524	10000	4303	4176	4160
	N=100000	100000	87149	85446	85228	100000	43031	41750	41585
	N=500000	500000	435745	427232	426136	500000	215173	208765	207952

Table 3: $\mathcal{N}\,(0, 0.2)\,, Supp = [-1, 1]$, **M=3**; BM: Benchmark, see Table 7.

Creating artificial observations, with the magnifying method is only good, when the bias itself is small, like in the exponential case. In other cases it does not help, only makes the standard errors smaller. Also note that as mentioned in section 5.1, as we increase S the number of observations becomes smaller. This leads to a moderation in the decrease of the bias and increase in the standard errors, if the original sample size is large (eg., $N = 10,000$). As the tables show, the effective number of observations are heavily depending on the underlying distribution, for the exponential case it is only around 1% of the data when $S = 100$, and

for the normal distribution it is more than 50%. This leads us to the recommendation not to use too much types of surveys, while significant decrease in the bias can be achieved by only ~ 10 surveys. More than that, will not add too much information.
The choice of methods is crucial. How well it performs depends on the underlying distribution and as it can be seen from the simulations, the magnifying method will guarantee that the bias vanishes as one increases S, at the cost of dropping many observations. Thus we recommend using it when one has more observation than 100,000 and a completely unknown underlying distribution. Shifting method seems to deliver better estimator, if number of surveys is small, observations are limited and the distribution is close to symmetric.

When one can use larger number of menu choices, eg., $M = 5$, it naturally gives better results: bias decreases faster, and one does not need to drop that much observations.

		Magnifying method - used as x^*_{it}							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0074	-0.0025	-0.0027	-0.0034	0.1304	-0.0058	-0.0077	-0.0004
	N=100,000	-0.0072	-0.0026	-0.0016	0.0007	0.1307	-0.0010	0.0008	0.0028
	N=500,000	-0.0078	-0.0026	-0.0018	-0.0003	0.1303	-0.0021	-0.0027	-0.0047
absbias	**N=10,000**	0.0394	0.0667	0.1629	0.2390	0.1305	0.0823	0.1792	0.2582
	N=100,000	0.0145	0.0223	0.0523	0.0746	0.1307	0.0239	0.0568	0.0781
	N=500,000	0.0090	0.0095	0.0250	0.0347	0.1303	0.0115	0.0246	0.0364
se	**N=10,000**	0.0489	0.0829	0.2085	0.3008	0.0437	0.1014	0.2265	0.3233
	N=100,000	0.0165	0.0277	0.0651	0.0937	0.0135	0.0302	0.0710	0.0976
	N=500,000	0.0073	0.0116	0.0318	0.0432	0.0059	0.0146	0.0305	0.0439
numObs	**N=10,000**	10000	3549	628	308	10000	2434	517	260
	N=100,000	100000	35486	6285	3072	100000	24348	5169	2593
	N=500,000	500000	177462	31438	15382	500000	121707	25860	12978
		Shifting method - used as IV							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10000**	-0.0074	0.0013	0.0003	0.0005	0.2132	0.0006	-0.0057	-0.0044
	N=100000	-0.0072	-0.0024	-0.0027	-0.0027	0.2132	0.0034	0.0040	0.0041
	N=500000	-0.0078	-0.0016	-0.0022	-0.0021	0.2133	0.0017	0.0028	0.0026
absbias	**N=10000**	0.0489	0.0886	0.0909	0.0908	0.2132	0.1639	0.1690	0.1681
	N=100000	0.0165	0.0288	0.0297	0.0300	0.2132	0.0540	0.0544	0.0548
	N=500000	0.0073	0.0134	0.0129	0.0130	0.2133	0.0228	0.0222	0.0217
se	**N=10000**	0.0843	0.1096	0.1134	0.1140	0.0739	0.2070	0.2141	0.2126
	N=100000	0.0279	0.0359	0.0370	0.0374	0.0233	0.0672	0.0677	0.0682
	N=500000	0.0131	0.0166	0.0161	0.0163	0.0109	0.0283	0.0275	0.0272
$corr(x^*_{it}, z_{it})$	**N=10000**	1.0000	0.9413	0.9376	0.9372	1.0000	0.9018	0.9000	0.8998
	N=100000	1.0000	0.9413	0.9376	0.9372	1.0000	0.9017	0.9000	0.8998
	N=500000	1.0000	0.9413	0.9376	0.9372	1.0000	0.9017	0.9000	0.8998
numObs	**N=10000**	10000	7840	7656	7633	10000	3941	3878	3870
	N=100000	100000	78379	76539	76308	100000	39386	38762	38685
	N=500000	500000	391864	382659	381511	500000	196934	193827	193441

Table 4: $Exp\,[0.5]\,, Supp = [0, 1]$, **M=5**; BM: Benchmark see Table 9.

		Magnifying method - used as x^*_{it}							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10,000**	-0.0318	-0.0079	-0.0013	0.0034	-0.0105	-0.0095	0.0013	0.0056
	N=100,000	-0.0319	-0.0085	-0.0009	-0.0003	-0.0101	-0.0079	0.0008	0.0027
	N=500,000	-0.0322	-0.0085	-0.0015	-0.0005	-0.0101	-0.0077	-0.0002	-0.0010
absbias	**N=10,000**	0.0337	0.0320	0.0764	0.1082	0.0200	0.0361	0.0776	0.1086
	N=100,000	0.0319	0.0119	0.0237	0.0356	0.0105	0.0133	0.0249	0.0353
	N=500,000	0.0322	0.0089	0.0113	0.0163	0.0101	0.0081	0.0114	0.0168
se	**N=10,000**	0.0233	0.0388	0.0956	0.1393	0.0232	0.0443	0.0982	0.1393
	N=100,000	0.0074	0.0122	0.0299	0.0444	0.0070	0.0145	0.0313	0.0444
	N=500,000	0.0036	0.0057	0.0144	0.0207	0.0030	0.0062	0.0143	0.0210
numObs	**N=10,000**	10000	3778	607	302	10000	3539	588	293
	N=100,000	100000	37782	6064	3009	100000	35387	5878	2930
	N=500,000	500000	188926	30327	15083	500000	176882	29360	14625
		Shifting method - used as IV							
		Truncated				Censored			
		BM	**S=10**	**S=50**	**S=100**	**BM**	**S=10**	**S=50**	**S=100**
bias	**N=10000**	-0.0318	-0.0009	0.0004	0.0006	-0.0105	-0.0005	0.0000	-0.0001
	N=100000	-0.0319	-0.0015	-0.0002	0.0000	-0.0101	-0.0002	0.0001	0.0001
	N=500000	-0.0322	-0.0013	-0.0002	0.0000	-0.0101	-0.0003	0.0005	0.0005
absbias	**N=10000**	0.0337	0.0222	0.0229	0.0230	0.0200	0.0327	0.0328	0.0327
	N=100000	0.0319	0.0074	0.0075	0.0075	0.0105	0.0104	0.0106	0.0106
	N=500000	0.0322	0.0036	0.0034	0.0034	0.0101	0.0041	0.0041	0.0042
se	**N=10000**	0.0233	0.0283	0.0290	0.0291	0.0232	0.0409	0.0409	0.0410
	N=100000	0.0074	0.0092	0.0094	0.0094	0.0070	0.0128	0.0131	0.0131
	N=500000	0.0036	0.0044	0.0044	0.0044	0.0030	0.0052	0.0052	0.0052
$corr(x^*_{it}, z_{it})$	**N=10000**	1.0000	0.9120	0.9086	0.9083	1.0000	0.8691	0.8664	0.8661
	N=100000	1.0000	0.9120	0.9086	0.9082	1.0000	0.8692	0.8666	0.8663
	N=500000	1.0000	0.9120	0.9086	0.9082	1.0000	0.8692	0.8665	0.8663
numObs	**N=10000**	10000	9483	9422	9415	10000	7730	7668	7661
	N=100000	100000	94830	94216	94138	100000	77279	76661	76582
	N=500000	500000	474152	471082	470691	500000	386355	383282	382890

Table 5: $\mathcal{N}\left(0, 0.2\right), Supp = [-1, 1]$, **M=5**; BM: Benchmark, see Table 7.

6 Some Further Extensions: Random Midpoints and Class Limits

In this section we are trying to blur the lines between menu choice type observations and continuous ones. Let us take again our example in model 4. But ask the question directly, say by moving an indicator on the screen between 0 and 100. The "usual" way to approach these types of observations is to consider them with a measurement error, i.e., by adding a white noise random term. We arguing here that in some cases another approach may be more realistic. Say that in our example one observation is 63%. This means that it can be considered as a menu choice type observation that falls into the [65%–70%] class (if we assume 20 classes, i.e., 5% "precision" for the observations), with random class boundaries and random class midpoints. That is the "real" answer to the question in this case is "around 65%–70%".

6.1 Stochastic Class Midpoints

Let us go back to equation (10). Instead of assuming that each menu/class is represented by its midpoint z_m, $m = 1, \ldots, M$, we assume, that the responses are randomly distributed with a on the domain of each class C_m, $\zeta_m \sim f_{\zeta_m}$. We assume that this distribution is known, it is in fact the expectation on what kind of bias the respondents might make when answering. Proceeding to equation (11), we now have

$$\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) = \frac{\sum_{m=1}^{M} \mathbb{E}(\zeta_m) \left[\mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}}(\beta x_i + u_i)\right)\right]}{\sum_{m=1}^{M} \mathbb{E}(N_m)\mathbb{E}^2(\zeta_m)}$$
$$= \beta \frac{\sum_{m=1}^{M} \mathbb{E}(N_m)\mathbb{E}(\zeta_m)\mathbb{E}\left(x \,|\, x \in C_m\right)}{\sum_{m=1}^{M} \mathbb{E}(N_m)\mathbb{E}^2(\zeta_m)}.$$

From this expression, we see that, that the OLS will only be unbiased if we assume that random variable ζ_m have the same expected value as the underlying random variable x, conditional on each class C_m, $m = 1, \dots, M$. This, in general, is quite unlikely scenario.

6.2 Stochastic Class Boundaries

Let us now return to the definition of the class boundaries in equation (9) and consider the case when they all are random variables on disjoint subintervals of $[a, b]$ rather than constants. Let $\delta_m \sim f_{\Delta_m}$ for $m = 0, \dots, M$ be the independent random boundaries and the expected value of the intermediate boundaries be $\mathbb{E}(\Delta_m) = c_m$ for $m = 1, \dots, M-1$. Therefore now we have random classes of the following form $C_1 = [\Delta_0, \Delta_1), C_2 = [\Delta_1, \Delta_2), \dots, C_M = [\Delta_{M-1}, \Delta_M]$. Note that if the distribution of Δ_0 and Δ_M is not trivial (Dirac delta), then their expected value does not match the lower and upper bound of the whole domain a and b. For simplicity, let the class value be given by the average of the corresponding two class boundaries, $Z_m = (\Delta_m + \Delta_{m-1})/2$ for $m = 1, \dots, M$ as in section 3. Furthermore, we also assume that the disturbance term ε in the regression equation (2) is independent of the class boundaries Δ. Equation (11) now reads as

$$\mathbb{E}(\hat{\beta}^*_{OLS}) = \beta \frac{\sum_{m=1}^{M} \mathbb{E}(Z_m)\mathbb{E}(N_m)\mathbb{E}(x \mid x \in C_m)}{\sum_{m=1}^{M} \mathbb{E}(N_m)\mathbb{E}^2(Z_m)},$$

where

$$\mathbb{E}(N_m) = N \int_{\mathrm{dom}(Z_{m-1})} \int_{\mathrm{dom}(Z_m)} \int_{\zeta_{m-1}}^{\zeta_m} f(x) f(\zeta_m) f(\zeta_{m-1}) \, \mathrm{d}x \, \mathrm{d}\zeta_m \, \mathrm{d}\zeta_{m-1}.$$

Therefore, similarly as in the case of stochastic class midpoint, $\hat{\beta}^*_{OLS}$ is only unbiased, in the unlikely case when the expected value of the class value Z_m matches the expected value of the underlying random variable x conditional on each class C_m, $m = 1, \dots, M$.

Let us remark here that this case covers the “rounding up” problem as well, when an answer, say 65% is in fact a rounded up value by the respondent. This 65 can be considered as a stochastic class midpoint, with random class boundaries, where the width of a class is dependent on the researcher’s confidence in the answer.

7 Conclusion

This paper investigated the effects of using menu choice variables in a linear regression model when the underlying variable is not observed. This situation arises often in survey data when the continuous variables, such as income for example, are not captured, but rather, being replaced by a set of M menu choices. Unlike other studies in the literature, our approach considered the more realistic case where the underlying distributions of the unobserved explanatory variables are unknown and the values of each menu choice can be arbitrary assigned. With fixed M, the results showed that using the menu choice counterparts as explanatory variables in a linear regression will lead to biased parameters estimates for the OLS and panel Within estimators in general. While the results provided analytical forms of the bias, the it is unfortunately not practically possible to obtain a bias-correction since it requires information from the distributions of the underlying explanatory variables, which are presumed to be unknown.

Under the assumption that no further information can be obtained for the unobserved explanatory variables, this paper proposed a novel survey construction by sub-sampling. Utilising the fact the menu choice variables approach their unobserved counterparts when M approaches infinity, the proposed approach essentially replaces the requirement of M being sufficiently large to the more standard scenario where the number of individuals, N is very large. Monte Carlo simulations showed that the proposed methods work reasonably well and they may have significant implications on the future of survey design.

Appendix: Some Monte Carlo Simulation Results on the Bias

Let us use the same very simple model as in Section 3.
The basic setup of the Monte Carlo experiment was: $\varepsilon_{it} \sim \mathcal{N}(0,1)$, $t = 1$, $\beta = 0.5$, x was 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^2 variances. When generating x^*, observation outside the support, whenever relevant, would be discarded (truncated approach), or assigned to the limit of the menu/class (censored approach). We report the *average bias* ($\bar{\beta}_{mc} = \sum_{mc}(\hat{\beta}_{mc} - \beta)/1000$), the *average absolute bias* ($\sum_{mc} |\hat{\beta}_{mc} - \beta|/1000$), and the *standard error* of the $\hat{\beta}$ estimated parameter ($\sqrt{\sum_{mc}(\hat{\beta}_{mc} - \bar{\beta}_{mc})^2/1000}$). 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:

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

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

Uniform Distribution

		Uniform[-1,1]				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**
bias	**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
absbias	**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
se	**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**
bias	**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
absbias	**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
se	**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**
bias	**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
absbias	**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
se	**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:** $\beta = 0.5, \sigma^2 = 5$

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

Let us turn next our attention to some other distributions.

Normal Distribution

From Table 7 it is clear that the OLS 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 notice that 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.

	Bias					
	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
	Abs. Bias					
	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
	SE					
	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,** $M = 5, \beta = 0.5, \sigma^2 = 1, Supp = [-1, 1]$

	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,** $a = -1, b = 1$

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 menu choices are censored the bias increases dramatically.

From Table 10, 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.

		$Exp\left[\lambda\right], Supp = [0,1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
bias	**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
absbias	**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
se	**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:** $\beta = 0.5, \sigma^2 = 5, \lambda = 0.5$

		$Weibull\,[b,c]\,,Supp=[0,1]$									
		Truncated					Censored				
		M=3	**M=5**	**M=10**	**M=20**	**M=50**	**M=3**	**M=5**	**M=10**	**M=20**	**M=50**
bias	**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
absbias	**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
se	**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 10: **Weibull distribution:** $\beta = 0.5, \sigma^2 = 0.5, b = 1, c = 0.5$

References

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

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

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

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

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

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

Frey, B. S. and Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic Literature*, 40(2):402–435.

Heath, Y. and Gifford, R. (2002). Extending the theory of planned behavior: Predicting the use of public transportation. *Journal of Applied Social Psychology*, 32(10):2154–2189.

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.

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.

Stutzer, A. (2004). The role of income aspirations in individual happiness. *Journal of Economic Behavior & Organization*, 54(1):89–109.

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

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

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

To be done
Most of the tasks to be done relate to Sections 5.1 and 5.2:

- Write up analytically the 2 methods on Section 5.1 and then they use in 5.2
- Derive for all the above cases the consistency and the speed of convergence.
- If feasible and not overly complicated, how the increase in the sample size N and S AND given N the increase in S reduce the bias of the OLS expressed in (12) and (13), as this would be the most important guideline for practitioners.