Modelling with Discretized Ordered Choice Covariates Supplementary Materials

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

August 26, 2020

* Curtin University
** Central European University

1 Introduction

This document contains supplementary materials to 'Modelling with Discretized Ordered Choice Covariates'. Section 2 contains detailed research of the Monte Carlo experiments. Section 3 provides theoretical exposition of the Ordinary Least Squares (OLS) estimator on model with discretized data. The discussion covers cross section and panel data. Section 4 contains technical proofs of all the Propositions in the paper while Section 5 contains algorithms for implementing the two sub-sampling methods. Section 6 provides a list of notations used in the paper.

2 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* ($\bar{\beta}_{mc} = \sum_{mc}(\hat{\beta}_{mc} - \beta)/1000$), 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:

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

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

2.1 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 1: **Uniform distribution:** $\beta = 0.5, \sigma_\varepsilon^2 = 5$

From Table 1 the unbiasedness and consistency (in sample size) of the OLS 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.

2.2 Normal Distribution

From Table 2 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 note that the Kullback-Liebler index gives a good indication of the bias (see Table 3). 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 2: **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 3: **Kullback-Leibler ratio: Uniform vs. Truncated/Censored Normal with different σ_x^2 values,** $a = -1, b = 1$

2.3 Exponential Distribution and Weibull Distributions

We carried out a large number of simulations with different parametrisations for both distributions. In Table 4 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.

From Table 6, 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**
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 4: **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**
bias	**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
absbias	**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
se	**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 5: **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**
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 6: **Weibull distribution:** $\beta = 0.5, \sigma_\varepsilon^2 = 0.5, b = 1, c = 0.5$

2.4 Some Observations

- **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 values 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 replacement estimators, 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 sub-samples, we can*

decrease the absolute bias, but using too many sub-samples is counter-productive. $S/N < 0.01\%$ 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 sub-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 sub-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 sub-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, sub-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.

3 Properties of OLS using Discretized Data

Recall the data generating process is assumed to be

$$y_i = w_i'\gamma + x_i'\beta + u_i \tag{S.1}$$

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

$$y_i = w_i'\gamma + x_i^{*'}\beta + u_i \tag{S.2}$$

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 from Equation (1) be the class midpoint.[1]

[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.

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

$$
\begin{aligned}
C_1 &= \left[a_l, a_l + \frac{a_u - a_l}{M}\right) && z_1 = a_l + \frac{a_u - a_l}{2M}, \\
&\vdots \\
C_m &= \left[a_l + (m-1)\frac{(a_u - a_l)}{M}, a_l + m\frac{a_u - a_l}{M}\right) && z_m = a_l + (2m-1)\frac{a_u - a_l}{2M}, \\
&\vdots \\
C_M &= \left[a_l + (M-1)\frac{(a_u - a_l)}{M}, a_l + M\frac{a_u - a_l}{M}\right] && z_M = a_l + (2M-1)\frac{a_u - a_l}{2M}.
\end{aligned}
\tag{S.3}
$$

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 cummulative distribution cdf $F(\cdot)$,

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) &= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \left(\beta(x_i^* + \xi_i) + u_i\right)\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \left[\beta\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^* + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i\right) + \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i\right]}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i^*}{\sum_{m=1}^{M} N_m z_m^2}\right\} + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&\qquad + \mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} \xi_i}{\sum_{m=1}^{M} N_m z_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} z_m N_m \upsilon^m}{\sum_{m=1}^{M} N_m z_m^2}\right\}. \qquad \text{(S.4)}
\end{aligned}$$

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 last inference uses the fact that the errors ξ_i have the same conditional distribution over the class C_m, $\upsilon^m \overset{\mathrm{d}}{=} \xi_i | C_m$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. Importantly, the second term in Equation (S.4) 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 1, right panel). These would be sufficient assumptions for the OLS 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' OLS 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 1) 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 (S.4), we obtain that

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

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 OLS is unbiased.

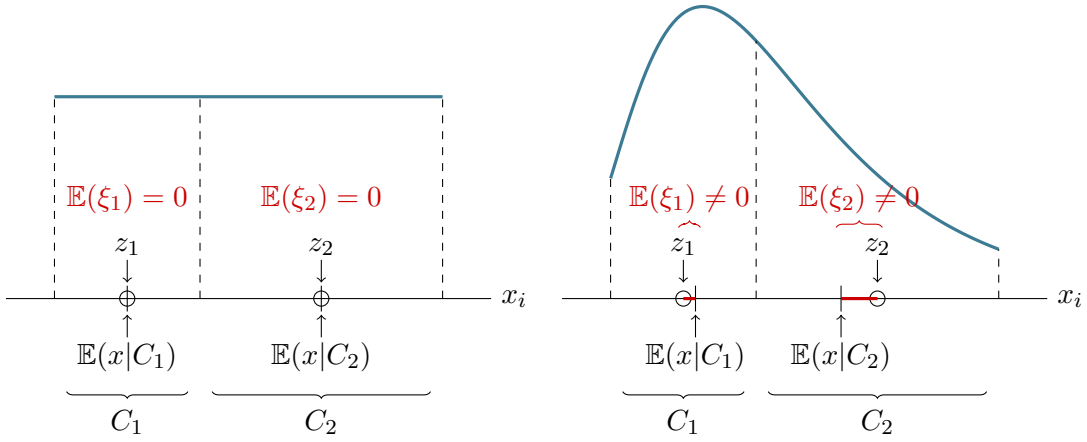

Figure 1: The difference between uniform (left panel) and general distributions (right panel)

3.1 N (in)consistency

This subsection considers the large sample properties of the estimator. First, assume that $\text{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} u_i = 0$, 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

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \hat{\beta}^*_{OLS} &= \operatorname*{plim}_{N\to\infty} \frac{\sum_{m=1}^{M} z_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (\beta x_i + u_i)\right]}{\sum_{m=1}^{M} N_m z_m^2} \\
&= \frac{\sum_{m=1}^{M} z_m \left[\text{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} (\beta x_i + u_i)\right]}{\sum_{m=1}^{M} z_m^2 \,\text{plim}_{N\to\infty} N_m} \\
&= \frac{\sum_{m=1}^{M} z_m \left[\text{plim}_{N\to\infty} \beta \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i\right]}{\sum_{m=1}^{M} z_m^2 \,\text{plim}_{N\to\infty} N_m} \\
&= \frac{\beta \sum_{m=1}^{M} z_m \left[\text{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_m\}} x_i\right]}{\sum_{m=1}^{M} z_m^2 \,\text{plim}_{N\to\infty} N_m}.
\end{aligned} \tag{S.5}$$

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, $x_m \overset{\text{d}}{=} x_i|C_m$, for all $m = 1, \ldots, M$, therefore its asymptotic distribution can be calculated by applying the Lindeberg-Levy Central Limit Theorem,

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

The $\hat{\beta}^*_{OLS}$ estimator is consistent if and only if the probability limit in Equation (S.5) equals β. To give a condition for consistency, first we rewrite the previous Equation (S.5) in terms

of the error terms ξ_i,

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

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

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

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^*_{OLS}-\beta\right) &= \frac{\operatorname{plim}_{N\to\infty}\beta\sum_{m=1}^{M} z_m\xi^m}{\operatorname{plim}_{N\to\infty}\sum_{m=1}^{M} z_m^2 N_m}\\ &= \frac{\operatorname{plim}_{N\to\infty}O(N)\beta\sum_{m=1}^{M} z_m\xi^m/N_m}{\operatorname{plim}_{N\to\infty}O(N)\sum_{m=1}^{M} z_m^2}\\ &= \frac{\beta\sum_{m=1}^{M} z_m\operatorname{plim}_{N\to\infty}\xi^m/N_m}{\sum_{m=1}^{M} z_m^2}O(N)\\ &= \frac{\beta\sum_{m=1}^{M} z_m\left\{\mathbb{E}(x_m)-z_m\right\}}{\sum_{m=1}^{M} z_m^2}O(N). \qquad \text{(S.6)}\end{aligned}$$

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 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 (S.6) 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 (S.6) zero, hence the *OLS* estimator is consistent.

Note that the consistency of the OLS 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 2).

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 (S.6). 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,

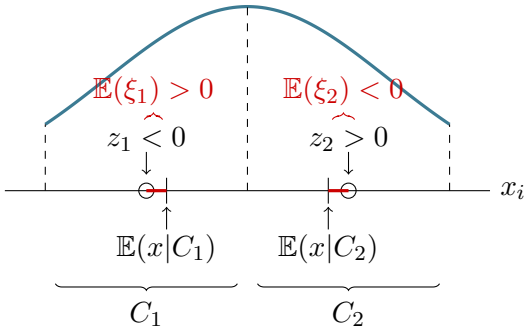

Figure 2: The estimator is inconsistent even in case of symmetric distributions (see Equation (S.6)).

the OLS estimator becomes inconsistent in N (see the size of the bias based on simulation results in the Appendix Supplement[2]).
So far we have focused on the estimation of β in Equation (S.2). 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.

3.2 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

[2]Available at: https://www.dropbox.com/s/ct7spjsozrwicy2/Menuchoice-JoE-Appendix.pdf?dl=0

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

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

where the index $i_m \in \{1,\ldots,N\}$ 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 $M = 10$.[3]

3.3 Some Remarks

The above results hold for much simpler cases as well. If instead of model (S.2) 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 $\bar{x}$.

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 (S.2). Consider a simple example where there is an unobserved variable x_i with an observed discretized choice version:

$$x_i^* = \begin{cases} z_1 & \text{if } c_0 \le x_i < c_1, \\ z_2 & \text{if } c_1 \le x_i < c_2, \end{cases} \tag{S.7}$$

[3]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).

and

$$y_i = x_i\beta + \varepsilon_i. \tag{S.8}$$

Using the discretized choice variable means:

$$y_i = x_i^*\beta + (x_i - x_i^*)\beta + u_i \tag{S.9}$$

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

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

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.

3.4 Estimation Reconsidered

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{S.10}$$

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 $\mathbf{z}_i = (z_{i1}, \dots, z_{iM})'$ contains the 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 $\mathbf{X}^* = \mathbf{E}'\mathbf{Z}$.

3.5 The OLS Estimator

From Equation (S.10), consider the regression based on the observed data:

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + (\mathbf{X} - \mathbf{X}^*)\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{S.11}$$

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 standard derivation shows that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\boldsymbol{\varepsilon}. \tag{S.12}$$

This implies OLS 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:

$$\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{X} \tag{S.13}$$

$$\mathbf{M}_{\mathbf{W}}\mathbf{X}^{*} = \mathbf{X}^{*}. \tag{S.14}$$

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 OLS 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 OLS 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,

$$\mathbb{E}\left(\mathbf{1}_m^{ki}\mathbf{1}_n^{li}\right) = \int_{\Omega} f(x_k, x_l)dx_k dx_l\,, \tag{S.15}$$

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 (S.15) 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,

$$\begin{aligned}\mathbb{E}\left[\sum_{m=1}^{M} z_{km}\mathbf{1}_m^{ki}x_{li}\right] &= \sum_{m=1}^{M} z_{km}\mathbb{E}\left[\mathbf{1}_m^{ki}x_{li}\right] \\ &= \sum_{m=1}^{M} z_{km}\int_{\Omega_1} x_l f(x_k, x_l)dx_k dx_l\,,\end{aligned} \tag{S.16}$$

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 (S.16) under the usual WLLN.

In the case when Equations (S.13) and (S.14) do not hold, the analysis becomes more tedious algebraically, but it does not affect the result that OLS is biased. Recall Equation (S.12), and let ω_{ij} be the (i, j) element in $\mathbf{M}_{\mathbf{W}}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, then following the same argument as above, $\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'$ 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

$$g_{kl} = \sum_{j=1}^{N}\sum_{i=1}^{N} \omega_{ij}\mathbf{1}_m^{ki}\mathbf{1}_n^{li} \tag{S.17}$$

with its expected value being

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{\Omega} \omega_{ij} f\left(x_k, x_l, \mathbf{w}\right) dx_k dx_k d\mathbf{w}, \tag{S.18}$$

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 (S.18) is complicated. However, under the assumption that the integral in Equation (S.18) exits, then $N^{-1}g_{kl}$ should converge to Equation (S.18) under the usual WLLN. It is also worth noting that $\mathbb{E}\left[\mathbf{M}_{\mathbf{W}}\mathbf{X}\right] = \mathbb{E}\left[\mathbf{M}_{\mathbf{W}}\right]\mathbb{E}\left[\mathbf{X}\right] = \mathbb{E}\left[\mathbf{X}\right]$ and $\mathbb{E}\left[\mathbf{M}_{\mathbf{W}}\mathbf{X}^*\right] = \mathbb{E}\left[\mathbf{M}_{\mathbf{W}}\right]\mathbb{E}\left[\mathbf{X}^*\right] = \mathbb{E}\left[\mathbf{X}^*\right]$ under the assumption of independence, which reduces Equation (S.18) to Equation (S.15).

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

$$h_{kl} = \sum_{m=1}^{M}\sum_{i=1}^{N} z_{km}\mathbf{1}_m^{ki}u_{li}, \tag{S.19}$$

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

$$\sum_{m=1}^{M} z_{km}\int_{\Omega_m} u_l f(x_k, x_l, \mathbf{w})dx_k dx_l d\mathbf{w}, \tag{S.20}$$

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 (S.20) to Equation (S.16).

3.6 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 (S.2) to

$$y_{it} = w'_{it}\gamma + x^{*'}_{it}\beta + \varepsilon_{it}, \tag{S.21}$$

and adjust the DGP, based on Equation (S.1)

$$y_{it} = w'_{it}\gamma + x'_{it}\beta + u_{it}, \tag{S.22}$$

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 (S.11) 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 $\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{Z} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^* \neq \mathbf{M}_{\mathbf{D}_N}\mathbf{X}$.

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

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.

4 Technical Proofs

Proof of Proposition 1

Recall

$$\begin{aligned}\mathbb{E}(N_b^{WS}) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_b^{WS}\}}\right) \\ &= N \Pr(x \in s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x.\end{aligned} \tag{S.23}$$

We can reformulate Equation (S.23) 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:

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

Next, we need to show that this is an increasing function in C_b^{WS}. Now as $N \to \infty$, under the assumption that $\Pr(x \in s) = 1/S$ 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.

Proof of Proposition 2

Recall

$$\Pr\left(x \in C_b^{WS}\right) = \sum_{s=1}^{S} \Pr(x \in s) \sum_{m=1}^{M} \Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,. \tag{S.24}$$

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 $\Pr(x \in S) = 1/S$ and using equation (S.24) 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 (S.24) 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.

4.1 Speed of Convergence for the Shifting Method

Recall

$$\Pr\left(x_i^\dagger \in C_b^{WS}\right) = \begin{cases} 0, & \text{if } s = 1 \text{ and } m = 1, \\ \frac{1}{S} \sum_{s=2}^{S} \frac{1}{s-1} \int_{C_1^{(s)} | C_b^{WS} \in C_1^{(s)}} f(x)\mathrm{d}x, & \text{if } s \neq 1 \text{ and } m = 1, \\ \frac{1}{S^2} \sum_{s=1}^{S} \int_{C_m^{(s)} | C_b^{WS} \in C_m^{(s)}} f(x)\mathrm{d}x, & \text{if } 1 < m < M, \\ \frac{1}{S} \sum_{s=1}^{S} \frac{1}{S-s+1} \int_{C_M^{(s)} | C_b^{WS} \in C_M^{(s)}} f(x)\mathrm{d}x, & \text{if } m = M\,. \end{cases} \tag{S.25}$$

For each of the conditions in Equation (S.25), 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 $1 < m < M$,

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

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

5 Algorithms

Algorithm 1 Magnifying method – creation of the sub-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.

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:

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

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.

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.

Algorithm 4 The shifting method - creation of sub-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 $c^{(s)}_0 = a_l$ and $c^{(s)}_M = a_u$.
3: If $s = 1$, set

$$c^{(s)}_m = c^{(s)}_{m-1} + \Delta, \qquad m = 2, \ldots, M-1$$

else

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

Note: $c^{(1)}_1$ does not exist.
4: If $s < S$ then $s := s + 1$ and goto Step 2.

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 sub-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,

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

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.

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.

6 Summary of the Notation Used in the Paper

Scalars:
N – number of individuals in the sample
T – number of time period in the sample (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 variable
γ – parameter for control variables
K – number of DOC variables (matrix notations)
J – number of control variables (matrix notations)
B – number of working sample classes
S – number of sub-samples
$N^{(s)}$ – number of observations in sub-sample s
$z_m^{(s)}$ – choice value of class m in sub-sample s
$c_m^{(s)}$ – s'th sub-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 shfting 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 – sub-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 choices with distribution $X \sim f(\cdot)$ (unknown)
X^* – discretized choice (DOC), with distribution $\psi(X)$ (observed)
$\hat{\beta}$ – parameter estimate for β with OLS (estimate)
$\hat{\gamma}$ – parameter estimate for γ with OLS (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: $\upsilon^m \overset{d}{=} \xi_i | C_m$ (not observed)
x_m – conditional distribution for x_i within class m, formally: $x_m \overset{d}{=} x_i | C_m$ (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 sub-sample s (observed)
N^{WS} – number of observations in the working-sample (observed)
$N_m^{(s)}$ – number of observations in sub-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. (ob-

served)
$x^{WS}_{i,NDTO}$ – 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)
$x^{\dagger}_i$ – artificial variable created during the shifting method (constructed)
$x^{WS}_{i,Shifting}$ – 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 sub-sample s
$\Psi(\cdot)$ – merging function
$||\cdot||$ – width of a class (or euclidean distance)
$Z(s,m)$ – set 'creator' function: given a sub-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 $x^{(s)}_i \in C^{(s)}_m$, with a given (uniform) probability
$\mathcal{F}^{WS}$ – assign estimated values $\hat{\pi}_\tau$ to each observation $x^{(s)}_i \in C^{(s)}_m$

Intervals
C_m – m'th class
$C^{(s)}_m$ – s sub-sample's, m'th class
C^{WS}_b – working sample's, b'th class

Sets
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
$\mathcal{A}^{(s)}_m$ – set for observations $x^{(s)}_i$ which are in class $C^{(s)}_m$

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 $\mathbf{Z}'\mathbf{EX}$
$\Omega_{\mathbf{X}}$ – 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
Sub-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 $\hat{\boldsymbol{\pi}}_\tau$

References

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.

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.

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.

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