Online supplement: Modelling with Sensitive Variables

Felix Chan[†], László Mátyás[‡] and Ágoston Reguly[*]

December 6, 2024

[†]Curtin University
[‡]Central European University
[*]Corvinus University of Budapest and Georgia Institute of Technology

1 Furhter supplements for shifting method

1.1 One-by-one discretization

A seemingly competitive alternative is to use discretization that discretizes each variable in $\mathbf{Z}$ one by one. However, this method will not ensure convergence in distribution for the joint $\mathbf{f}_{\mathbf{Z}}(\cdot)$ that is needed for point identification. To illustrate our argument let us use a simple illustrative example with $P = 2$, hence only two variables are discretized, $M = 3$ and $\mathbf{a}_l = [0, 0]$, $\mathbf{a}_u = [6, 6]$. For $s = 1$, $\mathcal{C}_{\mathbf{m}}^{(\mathbf{1})} = \{[(\mathbf{0}, \mathbf{0}), (\mathbf{2}, \mathbf{2})], [(\mathbf{2}, \mathbf{2}), (\mathbf{4}, \mathbf{4})], [(\mathbf{4}, \mathbf{4}), (\mathbf{6}, \mathbf{6})]\}$. Let us use $S = 6$, thus the shift size $\mathbf{h} = [0.5, 0.5]$. Discretizing the variables independently from each other will result in fixed M intervals along all other dimensions while learning more and more along one dimension. This will result in gaps in the domain of $\mathbf{f}_{\mathbf{Z}}(\cdot)$. Figure S.1 shows this case, where grey blocks show the mapped/learned parts of the distribution of $\mathbf{f}_{\mathbf{Z}}(\cdot)$.

[*]Corresponding author; e-mail: agoston.reguly@uni-corvinus.hu

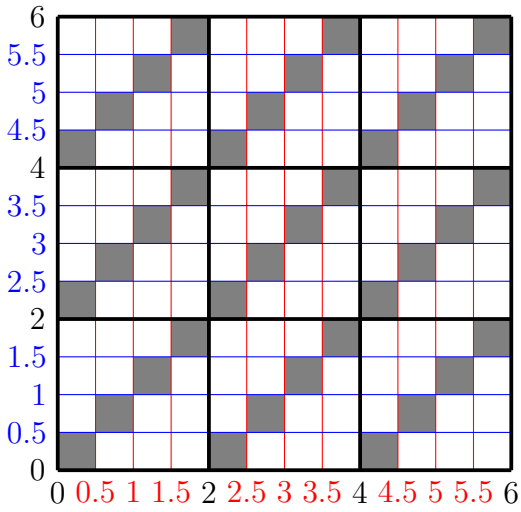

Figure S.1: One-by-one variable discretization

1.2 Algorithms for Shifting method

Algorithm A1 The shifting method - creation of split samples

1: For any given S and M, set $s = 1$

$$B = S(M-1)\ , \qquad h = \frac{a_u - a_l}{B}\ , \qquad \Delta = \frac{a_u - a_l}{M-1}\ .$$

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

3: If $s = 1$, set

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

else

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

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

4: If $s < S$ then $s := s + 1$ and goto Step 2.

Algorithm A2 The shifting method – creation of synthetic variable ($Z_i^\dagger$)

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

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

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

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

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

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

4: Draw observations from $\mathcal{A}_m^{(s)}$ and assign a new value (e.g. mid-value) to $\mathcal{V}_j$ based on a randomly selected interval of $\mathcal{C}_b^{WS} \in \mathcal{C}_m^{(s)}$, with probabilities given by,

$$\Pr\left(Z^\dagger \in \mathcal{C}_b^{WS} | Z^{(s)} \in \mathcal{C}_m^{(s)}\right) = \begin{cases} 1, & \text{if } s = 1 \text{ and } m = 1, \\ 1/(s-1), & \text{if } s \neq 1 \text{ and } m = 1, \\ 1/S, & \text{if } 1 < m < M, \text{ or} \\ 1/(S-s+1), & \text{if } m = M. \end{cases}$$

Example: Let $\mathcal{C}_3^{(2)} = [2.5, 4.5]$, $\mathcal{A}_m^{(s)} = \{3.5, 3.5, 3.5\}$, $N_m^{(s)} = 3$. The different intervals in the working sample is $C_{b_1}^{WS} = [2.5, 3]$, $C_{b_2}^{WS} = [3, 3.5]$, $C_{b_3}^{WS} = [3.5, 4]$ and $C_{b_4}^{WS} = [4, 4.5]$. Now we assign each member of $\mathcal{A}_m^{(s)}$ to one of the mid-value of C_b^{WS} with 1/4 probabilities which results in the values of $\mathcal{V}_j = \{2.75, 3.25, 3.75, 4.25\}$.

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

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

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

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

1.3 Algorithms for conditional mean estimators

Algorithm A3 describes the process for creating the working sample, when discretization happens with one or more right-hand side variables.

Algorithm A3 Discretization of explanatory variables – creation of working sample

1: Estimate $\hat{\boldsymbol{\kappa}}$ as defined in Equation (B.1) from the Appendix.

2: Set $c := 1, s := 1, \mathbf{m} := (1, \ldots, 1), k := 0,\ \{\mathbf{y}^{WS}, \mathbf{X}^{WS}, \mathbf{W}^{WS}\} = \emptyset$.

3: Assign the conditional mean for $\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(\mathbf{s})}$ from the c'th element of $\hat{\boldsymbol{\kappa}}$ to all discretized observations $\mathbf{X}^* \in \mathcal{C}_{\mathbf{m}}^{(\mathbf{s})}$. and the observed values $\mathbf{y}_j^{(s)}, \mathbf{W}_j^{(s)}$ to the working sample,

$$\left\{\mathbf{y}_i^{WS}, \mathbf{X}_i^{WS}, \mathbf{W}_i^{WS}\right\} := \left\{\mathbf{y}_i^{WS}, \mathbf{X}_i^{WS}, \mathbf{W}_i^{WS}, \bigcup_{j=1}^{N} \left(\mathbf{y}_j^{(s)}, \hat{\boldsymbol{\kappa}}(c), \mathbf{w}_j^{(s)} \mid \mathbf{X}_j^* \in \mathcal{C}_{\mathbf{m}}^{(\mathbf{s})}\right)\right\} .$$

4: Set $c := c + 1$.

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

6: If $s = S$, then $s := 1, k := k + 1$ and set $\mathbf{m} = \mathbf{m} + \mathbf{1}_{\{k\}}$, where $\mathbf{1}_{\{k\}}$ is a $(K \times 1)$ indicator function with value of 1 at element k, otherwise 0. Go to Step 3.

Algorithm A4 describes how to create in practice the working sample that can be used for estimation.

Algorithm A4 Discretization of the outcome variable – creation of working sample

1: Partition the data into L mutually exclusive sets based on the values of $\mathbf{X}$.

2: Estimate $\hat{\boldsymbol{\pi}}$ as defined in Equation (B.2) from the Appendix.

3: Set $c := 1, s := 1, m := 1, k := 1, \mathbf{l} := (1, \ldots, 1),\ \{\mathbf{y}^{WS}, \mathbf{X}^{WS}\} = \emptyset$.

4: Calculate the sample conditional mean: $\hat{\mathbb{E}}\left[\mathbf{X} | \mathbf{X} \in \mathcal{D}_{\mathbf{l}}\right]$ defined by the partition $\mathbf{l}$.

5: Add the c'th element of the conditional mean estimator $\hat{\boldsymbol{\pi}}$ and the calculated sample means of $\mathbf{X}$ to the working sample,

$$\left\{\mathbf{y}_i^{WS}, \mathbf{X}_i^{WS}\right\} := \left\{\mathbf{y}_i^{WS}, \mathbf{X}_i^{WS}, \bigcup_{j=1}^{N} \left(\hat{\boldsymbol{\pi}}(c), \hat{\mathbb{E}}\left[\mathbf{X} | \mathbf{X} \in \mathcal{D}_{\mathbf{l}}\right] \mid \mathbf{X}_j \in \mathcal{D}_{\mathbf{l}}\right)\right\} .$$

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

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

8: If $s = S$, then set $s := 1$, $m := 1$, $k := k + 1$,and set $\mathbf{l} = \mathbf{l} + \mathbf{1}_{\{k\}}$, where $\mathbf{1}_{\{k\}}$ is a $(K \times 1)$ indicator function with value of 1 at element k, otherwise 0. and go to Step 4.

2 Discretization on both sides

Let us investigate the case when discretization happens with both outcome $\mathbf{y}^*$, and with one or more explanatory variables $\mathbf{X}^*$. In this case, we do not need to partition the domain of $\mathbf{X}$, but can use the discretization grids $\mathcal{C}_{\mathbf{m}}^{(\mathbf{s})}$ for the explanatory variable. Note as $\mathbf{y}$ is discretized, we need to partition $\mathbf{W}$ with $\mathcal{D}_{\mathbf{l}}$, similarly to the case discussed in Section 4.2 from the main text. To simplify our derivations, let us assume the number of split samples and the number of classes are the same for both $\mathbf{y}$ and for $\mathbf{X}$, thus $S = S_Y = S_X$ and $M = M_Y = M_X$.[1]

Step 1: Identification with $\mathbf{y}^*$ and $\mathbf{X}^*$

Identification, when discretization happens on both sides, Equation (13) holds from the main text, but instead of $\mathbf{X} \in \mathcal{D}_{\mathbf{l}}$, one need to condition on $\mathbf{X} \in \mathcal{C}_{\mathbf{m_X}}$, correct conditioning for $\mathbf{y}$ to $\mathbf{y}^* \in \mathcal{C}_{m_Y}$ and add the conditioning $\mathbf{W} \in \mathcal{D}_{\mathbf{l}}$. This leads to

$$\begin{aligned} &\sum_{l}\sum_{m_X} \mathbb{E}[\mathbf{y}|\mathbf{y}^* \in \mathcal{C}_{m_Y}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}] \Pr[\mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}] = \\ &\mathbb{E}\left[\mathbf{X}|\mathbf{y}^* \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right] \boldsymbol{\beta} + \mathbb{E}\left[\mathbf{W}|\mathbf{y}^* \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right] \boldsymbol{\gamma} \end{aligned} \tag{S.1}$$

Note that $\Pr[\mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}]$ and $\mathbb{E}\left[\mathbf{W}|\mathbf{y}^* \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right]$ are identified as usual. The conditional expectations affected by discretization are given by

$$\begin{aligned} \psi_Y(s, \mathbf{m}, \mathbf{l}) :&= \mathbb{E}\left(\mathbf{y}|\mathbf{y}^* \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s})}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right) \\ &= \lim_{N,S\to\infty} \mathbb{E}\left(\mathbf{y}^\dagger|\mathbf{y}^\dagger \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^\dagger \in \mathcal{C}_{\mathbf{m,X}}^{(\mathbf{s})}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right) \ , \\ \psi_X(s, \mathbf{m}, \mathbf{l}) :&= \mathbb{E}\left(\mathbf{X}|\mathbf{y}^* \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s})}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right) \\ &= \lim_{N,S\to\infty} \mathbb{E}\left(\mathbf{X}^\dagger|\mathbf{y}^\dagger \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^\dagger \in \mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s})}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right) \ , \end{aligned}$$

where $\mathcal{C}_{m_Y}^{(s)}$ is the discretization intervals for $\mathbf{y}$, $\mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s})}$ is the discretization grid for $\mathbf{X}$. Note that, $s = \{s_Y, s_X\}$ and $m = \{m_Y, \mathbf{m}_X\}$ in this case, for $\mathcal{C}_{m_Y}^{(s)}$ and for $\mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s})}$.

Step 2: OLS estimators for conditional expectations

We propose conditional mean estimators, $\hat{\boldsymbol{\psi}}_Y$ and $\hat{\boldsymbol{\psi}}_X$ via OLS in the same spirit as for the other cases. To show the properties of the estimators, let us make our notation more tractable. First, let us define $\mathbf{s} = (s_Y, s_X)$ the split samples used for $\mathbf{y}$ and $\mathbf{X}$. The number of split samples S_Y and S_X can be the same or different. Let $\mathbf{m} = (m_Y, \mathbf{m}_X)$, vector for the

[1] One can extend the analysis, all results are the same.

intervals for $\mathbf{y}$ (scalar) and grids for $\mathbf{X}$ (vector with $(K \times 1)$ elements). The number of used intervals and grids may be the same or different. Finally, $\mathbf{l}$ represents the partition vector for the J variables from $\mathbf{W}$.

Overall for $\psi_Y(s, \mathbf{m}, \mathbf{l})$ and $\psi_X(s, \mathbf{m}, \mathbf{l})$, we have $S_Y \times S_X \times M_Y \times M_X^K \times L^J$ different cases. We use vectorized versions for both denoted by $\boldsymbol{\psi}_Y$ and $\boldsymbol{\psi}_X$, that contains these elements iterating in the order of $s_y, s_m, m_y, \mathbf{m}_X, \mathbf{l}$. We propose estimators $\widehat{\boldsymbol{\psi}}_Y$ and $\widehat{\boldsymbol{\psi}}_X$ via OLS. To simplify our notation, let $\mathbf{1}'_{\{\psi\}} = \mathbf{1}'_{\{\mathbf{y}^\dagger \in \mathcal{C}^{(s)}_{m,Y}, \mathbf{X}^\dagger \in \mathcal{C}^{(\mathbf{s})}_{\mathbf{m},\mathbf{X}}, \mathbf{W} \in \mathcal{D}_\mathbf{l}\}}$. The estimators are,

$$\begin{aligned} \widehat{\boldsymbol{\psi}}_Y &= \left(\mathbf{1}'_{\{\psi\}}\mathbf{1}_{\{\psi\}}\right)^{-1}\mathbf{1}'_{\{\psi\}}\mathbf{y}^\dagger\,, \\ \widehat{\boldsymbol{\psi}}_X &= \left(\mathbf{1}'_{\{\psi\}}\mathbf{1}_{\{\psi\}}\right)^{-1}\mathbf{1}'_{\{\psi\}}\mathbf{X}^\dagger\,, \end{aligned} \tag{S.2}$$

Under same assumptions as in Appendix B.1. and B.2. from the main text $\widehat{\boldsymbol{\psi}}_Y \to \boldsymbol{\psi}_Y$ and $\widehat{\boldsymbol{\psi}}_X \to \boldsymbol{\psi}_X$. To be specific, we utilize the weak law of large numbers in both cases and $\lim_{S_Y\to\infty} F_{\mathbf{y}^\dagger}(\cdot) = F(\cdot), \lim_{S_X\to\infty} F_{\mathbf{X}^\dagger}(\cdot) = F_{\mathbf{X}}(\cdot)$ under Assumptions 2.a), 2.b) and $\Pr(\mathbf{y} \in \mathcal{S}_s) = 1/S_Y, \Pr(\mathbf{X} \in \mathcal{S}_s) = 1/S_X$ as shown in Section 3.2. Note that K, J are the number of (discretized) regressors and fixed, as well as L the number of partitions. The asymptotic distribution of the estimator can be derived, similarly. Let us write,

$$\mathbf{y}^\dagger = \mathbf{1}_{\{\psi\}}\boldsymbol{\psi}_Y + \boldsymbol{\eta}_{\psi_Y}\,, \qquad \mathbf{X}^\dagger = \mathbf{1}_{\{\psi\}}\boldsymbol{\psi}_X + \boldsymbol{\eta}_{\psi_X},$$

where $\boldsymbol{\eta}_{\psi_Y} = (\eta_{\psi_Y,1}, \ldots, \eta_{\psi_Y,N})'$ and $\boldsymbol{\eta}_{\psi_X} = (\eta_{\psi_X,1}, \ldots, \eta_{\psi_X,N})'$ are the corresponding idiosyncratic terms. Under standard OLS assumption, we have

$$\sqrt{N}\left(\widehat{\boldsymbol{\psi}}_Y - \boldsymbol{\psi}_Y\right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\psi}_Y}\right)\,, \qquad \sqrt{N}\left(\widehat{\boldsymbol{\psi}}_X - \boldsymbol{\psi}_X\right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\psi}_X}\right)\,.$$

The variance of the OLS estimators is given by

$$\boldsymbol{\Omega}_{\boldsymbol{\psi}_Y} = V\left(\boldsymbol{\eta}_{\psi_Y}\right)\left(\mathbf{1}'_{\{\psi\}}\mathbf{1}_{\{\psi\}}\right)^{-1}\,, \qquad \boldsymbol{\Omega}_{\boldsymbol{\psi}_X} = V\left(\boldsymbol{\eta}_{\psi_X}\right)\left(\mathbf{1}'_{\{\psi\}}\mathbf{1}_{\{\psi\}}\right)^{-1}$$

where $V\left(\boldsymbol{\eta}_{\psi_Y}\right), V\left(\boldsymbol{\eta}_{\psi_X}\right)$ are the variances of the idiosyncratic term. Algorithm A5 describes how to create in practice a working sample when discretization happens on both sides of the regression equation.

Algorithm A5 Discretization on both sides – creation of working sample

1: Partition the $\mathbf{W}$ into L mutually exclusive sets.
2: Estimate $\hat{\boldsymbol{\psi}}_Y$ and $\hat{\boldsymbol{\psi}}_X$ as defined in Equation (S.2)
3: Set $c := 1, s_Y := 1, s_X := 1, m_Y := 1, \mathbf{m}_X := (1, \ldots, 1), k := 0, \mathbf{l} := (1, \ldots, 1)$, $j := 0$, $\{\mathbf{y}^{WS}, \mathbf{X}^{WS}, \mathbf{W}^{WS}\} = \emptyset$.
4: Calculate the sample conditional mean: $\bar{\mathbf{w}} := \hat{\mathbb{E}}\left[\mathbf{W} | \mathbf{y}^* \in \mathcal{C}_{m_y}^{(s_y)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s_x})}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right]$
5: Add the c'th element of the conditional mean estimators $\widehat{\boldsymbol{\psi}}_Y$, $\widehat{\boldsymbol{\psi}}_X$ and the calculated sample means of $\mathbf{W}$ to the working sample,

$$\left\{\mathbf{y}_i^{WS}, \mathbf{X}_i^{WS}, \mathbf{W}_i^{WS}\right\} := \left\{\mathbf{y}_i^{WS}, \mathbf{X}_i^{WS}, \mathbf{W}_i^{WS} \right. \\ \left. \bigcup_{j=1}^{N} \left(\widehat{\boldsymbol{\psi}}_Y(c), \widehat{\boldsymbol{\psi}}_X(c), \bar{\mathbf{w}} \mid \mathbf{y}_j^* \in \mathcal{C}_{m_y}^{(s_y)}, \mathbf{X}_j^* \in \mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s_x})}, \mathbf{W_j} \in \mathcal{D}_{\mathbf{l}}\right)\right\} .$$

6: If $s_Y < S_Y$, then $s_Y := s_Y + 1$ and go to Step 4.
7: If $s_Y = S_Y$, then set $s_Y := 1$, $s_X := s_X + 1$ and go to Step 4.
8: If $s_X = S_X$, then set $s_Y := 1, s_X := 1$, $m_Y := m_Y + 1$ and go to Step 4.
9: If $m_y = M_Y$, then set $s_Y := 1, s_X := 1, m_Y := 1, k := k + 1$, $\mathbf{m}_X := \mathbf{m}_X + \mathbf{1}_{\{k\}}$ and go to Step 4.
10: If $k = K - 1$, then set $s_Y := 1, s_X := 1, m_Y := 1, k := 0$, $\mathbf{m}_X := (1, \ldots, 1), j := j + 1$, $\mathbf{l} := \mathbf{l} + \mathbf{1}_{\{j\}}$ and go to Step 4.

Step 3: OLS estimator for $\boldsymbol{\beta}$

To get a consistent estimator for $\boldsymbol{\beta}$, let us define $\breve{\mathbf{y}}$ and $\breve{\mathbf{X}}$ that takes the corresponding values from $\hat{\boldsymbol{\psi}}_Y$, and $\hat{\boldsymbol{\psi}}_X$ based on the discretized values of $\mathbf{y}^*$ and $\mathbf{X}^*$. Note that if there are additional controls $\mathbf{W}$, one needs to replace it with $\breve{\mathbf{W}}$ that takes the sample conditional means of $\mathbb{E}\left(\mathbf{W} | \mathbf{y}^* \in \mathcal{C}_{m_Y}^{(s)}, \mathbf{X}^* \in \mathcal{C}_{\mathbf{m_X}}^{(\mathbf{s})}, \mathbf{W} \in \mathcal{D}_{\mathbf{l}}\right)$.

$$\breve{\mathbf{y}} = \breve{\mathbf{X}}\boldsymbol{\beta} + \breve{\mathbf{W}}\boldsymbol{\gamma} + \breve{\boldsymbol{\nu}} ,$$

where $\breve{\boldsymbol{\nu}} = (\nu_1, \ldots, \nu_N)$. The OLS estimator for $\boldsymbol{\beta}$ is,

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

where $\mathbf{M}_{\breve{\mathbf{W}}}$ is the usual residual maker, using $\breve{\mathbf{W}}$. Note that $\mathbb{E}[\breve{\nu}_i] = 0$ for all $i = 1, \ldots, N$ since $\breve{\mathbf{y}}$, $\breve{\mathbf{X}}$ and $\breve{\mathbf{W}}$ are consistent estimates of the corresponding conditional expectations.

Moreover, $\mathbb{E}[\breve{\nu}_i\breve{\nu}_j] = 0$ for $i \neq j$ due to $\mathcal{D}_{\mathbf{l}}$, and $\mathcal{C}_{\mathbf{m}}^{(\mathbf{s})}$ in m and k are mutually exclusive. $\mathbb{E}[\breve{\nu}_i|\breve{\mathbf{X}}, \breve{\mathbf{W}}] = 0$ since the discretization does not affect the sampling error. Furthermore,if the discretization is mean independent from the discretization scheme, then, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$. Under the these assumptions, $\mathbb{E}\left[\breve{\boldsymbol{\nu}}\breve{\boldsymbol{\nu}}'|\breve{\mathbf{X}}, \breve{\mathbf{W}}\right] = \sigma_\nu^2\mathbf{I}$, the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is given by $\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \overset{\text{a}}{\sim} \mathcal{N}\left(\mathbf{0}, \sigma_{\breve{\boldsymbol{\nu}}}^2\left(\breve{\mathbf{X}}'\mathbf{M}_{\breve{\mathbf{W}}}\breve{\mathbf{X}}\right)^{-1}\right)$.

3 Magnifying Method

Here, we discuss another split sampling method that also converges in distribution to the unknown distribution, but has different discretization scheme thus different properties. The idea for magnifying method is to magnify specific parts of the underlying variable's domain. The interval size for each split sample depends on the number of split samples (S) and the number of original intervals (M). As the number of split samples increases, the interval widths decrease, which is the main mechanism for uncovering the unknown distribution. Figure S.2 shows the main idea of the magnifying method for the case of $M = 3, S = 4$.

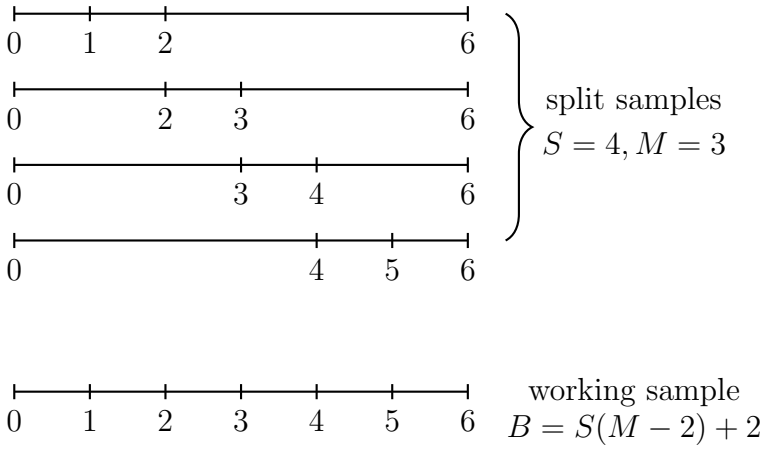

Figure S.2: The magnifying method

As seen in Figure S.2, there are two types of intervals: i) Around the boundaries of Z, the first and last interval widths are changing and in general they do not decrease as we increase the number of split sizes. Observations that fall into these intervals are called "*non-directly-transferable observations*" or NDTOs. ii) All remaining intervals have widths that are decreasing as S increases. Observations falling into these categories are called "*directly-transferable observations*" or DTOs. These DTOs can then be used to map the original distribution.[2]

[2]Note that, the first and last split samples are slightly different. Their observations from first/last intervals

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

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

Note that we have 2 split samples (first and last respectively), where we magnify intervals around the boundary of the domain. Here, we capture $M-1$ intervals of equal size, for all the other $S-2$ split samples, there are $M-2$ intervals.

The widths of the intervals in the working sample are given by,

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

Intuitively, the magnifying method works as if we increase the number of split samples, the magnified interval widths go to zero ($h \to 0$ as $S \to \infty$). Note however, to have this property we will fix the upper and lower bounds of the support for the split samples ($a_l = c_0^{WS} = c_0^{(s)};\ a_u = c_B^{WS} = c_M^{(s)},\ \forall s$), thus make sure that the discretized domain does not have infinite support. Next, we derive the boundary points for each magnified split sample and show how and under what assumptions the working sample converges in distribution to the underlying unknown distribution.

Algorithm A6 Magnifying method – creation of the split samples

1: For any given S and M. Set

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

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

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

else set

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

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

($c_1^{(1)}$ and $c_M^{(S)}$) are also DTOs, as the interval widths are decreasing in S similarly to the intervals that are not at the boundary points ($c_m^{(s)}$, $\forall\ 1 < s < S, 1 < m < M$).

The boundary points for each split sample can be derived as

$$c_m^{(s)} = \begin{cases} a_l \text{ or } -\infty & \text{if } m = 0, \\ a_l + mh & \text{if } 0 < m < M \text{ and } s = 1, \\ a_l + h\left[(s-2)(M-2) + M + m - 2\right] & \text{if } 0 < m < M \text{ and } s > 1, \\ a_u \text{ or } \infty & \text{if } m = M. \end{cases} \tag{S.3}$$

The intuition behind this is that on the boundaries of the support, the split samples take the values of the lower and upper bounds. For the first split sample, one needs to shift the boundary points m times. However, for the other split samples, one needs to push by $h(M-1)$ times to shift through the first questionnaire and then $h(M-2)$ to shift through each split sample in between $s = 2$ and $s = S-1$, $s-2$ times. Deriving this process algebraically will result in the above expression.[3] Algorithm A6 shows how to create the boundary points for the split samples in the case of the magnifying method.

The working sample's boundary points are given by the unique boundary points from the split samples, which leads to

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

To show how and why the magnifying method works, let us derive the different interval widths for each split samples' interval, defined by Equation (S.3). Let $||\mathcal{C}_m^{(s)}|| = c_m^{(s)} - c_{m-1}^{(s)}$ be the m-th interval width, then for the split samples which are in-between the boundaries $(1 < s < S)$ and substituting for h, we can write

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

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

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

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

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

Note that $||\mathcal{C}_m^{(s)}|| \leq ||\mathcal{C}_1^{(s)}||$ and $||\mathcal{C}_m^{(s)}|| \leq ||\mathcal{C}_M^{(s)}||$. Formally, let us define $\zeta := \{\mathcal{C}_m^{(s)} \mid 1 < m < M, 1 < s < S, \mathcal{C}_m^{(1)} \mid 1 \leq m < M, \mathcal{C}_m^{(S)} \mid 1 < m \leq M\}$ as the set of intervals which have the interval width $\frac{a_u - a_l}{S(M-2)+2}$.

An interesting insight that we can write $\Pr\left((z - z^{(s)})^2 | z \in \zeta \leq (z - z^{(s)})^2 \mid z \notin \zeta\right) = 1$, which is true if and only if, $\mathbb{E}[Z] = \mathbb{E}\left[Z^{(s)}\right], \forall Z$. One example is when Z is uniformly distributed.

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

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

and for the first and last split sample

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

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

This formulation takes a_l as the starting point and expresses the boundary points given a_l. However, we can use a_u as the starting point as well to shift the boundary point. This implies that the convergences on the bounds $(||\mathcal{C}_1^{(s)}||, ||\mathcal{C}_M^{(s)}||)$ will change, resulting in those parts not converging to 0 in general.

Now, it is clear that there are two types of observations: The first type is $Z_i^{(s)} \in \zeta$. These observations are the closest to the underlying unknown observations, as these have the feature of $\lim_{S \to \infty} ||\mathcal{C}_m^{(s)}|| = 0$. Moreover, these observations have the same interval width as the working sample's intervals and each of them can be directly linked to a certain working sample interval by design. Formally, $\exists \mathcal{C}_m^{(s)} \cong \mathcal{C}_b^{WS}$ such that $c_m^{(s)} = c_b^{WS}$, $c_{m-1}^{(s)} = c_{b-1}^{WS}$. We call these values "*directly transferable observations*", as we can directly transfer and use them

in the working sample. These observations are denoted by $Z_i^{DTO} := Z_i^{(s)} \in \zeta$, $\forall s$, and the related random variable by Z^{DTO}.

The second type of observation is all others for which none of the above is true. We call them "*non–directly transferable observations*". Algorithm A7 describes how to construct the working sample when using only the directly transferable observations.

Algorithm A7 Magnifying method - creation of the "DTO" working sample

1: Set $m = 1, s = 1$ and $Z_i^{DTO} = \emptyset$.

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

$$Z_i^{DTO} := \left\{ Z_i^{DTO}, \bigcup_{j=1}^{N} \left(Z_j^{(s)} \in \mathcal{C}_m^{(s)} \mid \mathcal{C}_m^{(s)} \in \zeta \right) \right\}$$

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

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

As a next step let us derive the probability that a *directly transferable observation* lies in a given interval of the working sample. Based on Equation (7) from the main text,

$$\Pr\left(Z \in \mathcal{C}_b^{WS}\right) = \Pr(Z \in \mathcal{S}_s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z) \mathrm{d}z \,.$$

Note, here we can use the fact that individual i being assigned to a split sample s is independent of i choosing the interval with class value $Z_m^{(s)}$.

An important requirement of the magnifying method is that we want to ensure that in each interval in the working sample, there are directly transferable observations. For each split sample, the expected number of directly transferable observations is

$$\begin{aligned} \mathbb{E}(N_b^{WS}) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{Z_i \in \mathcal{C}_b^{WS}\}} \right) \\ &= N \Pr(Z \in \mathcal{S}_s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z) \mathrm{d}z. \end{aligned} \tag{S.4}$$

Following from Equation (S.4), consider the following assumptions,

Assumption M1. *Let Z be a continuous random variable with probability density function $f_Z(z)$ with S, N and $\mathcal{C}_m^{(s)}$ follow the definitions above. We require that all split samples will have non-zero respondents,* $\Pr\left(Z \in \mathcal{S}_s\right) > 0$.

Assumption M1 ensures utilization of all split samples, i.e. each split sample will have non-zero respondents. Similarly, for the shifting case, we need assumption 1.a), which ensures that the number of respondents will always be higher than the number of split samples, and assumption 1.b), that imposes a mild assumption on the underlying distribution. (The support of the random variable is not disjoint, thus $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z)\mathrm{d}z > 0$.) These assumptions allow us to establish proposition S.1, which establishes convergence in distribution.

Proposition S.1. *Under Assumptions 1a, 1b from the main text and M1,*

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

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

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

Proof:
The probabilities of the unobserved variable to fall into class $\mathcal{C}_b^{WS}$,

$$\Pr\left(Z \in \mathcal{C}_b^{WS}\right) = \Pr(Z \in \mathcal{S}_s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z)\mathrm{d}z\,.$$

where, $\mathcal{S}_s$ is the set for split sample s, and we used the fact that individual i being assigned to a split sample s is independent of i. This is satisfied if the discretization schemes are randomly assigned to observations. To ensure that in each interval from the working sample, there are directly transferable observations, let us write

$$\begin{aligned}\mathbb{E}(N_b^{WS}) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{Z_i \in \mathcal{C}_b^{WS}\}}\right) \\ &= N \Pr(Z \in \mathcal{S}_s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z)\mathrm{d}z.\end{aligned} \tag{S.5}$$

We can reformulate Equation (S.5) by considering the number of observations up to a certain boundary point, rather than the number of observations in a particular class. That is

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

This gives the possibility to replace $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f_Z(z)\mathrm{d}z$ with $\int_{c_0^{WS}}^{c_b^{WS}} f_Z(z)\mathrm{d}z$. Since this is a CDF, and hence a non-decreasing function, it effectively shows that each interval has non-empty observations:

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

Next, we need to show that this is an increasing function in $\mathcal{C}_b^{WS}$. As $N\to\infty$, under the assumption that $\Pr(Z\in\mathcal{S}_s)=1/S$ and $S/N\to d$ with $d\in(0,1)$ – which is satisfied when $S=dN$,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}\left(\sum_{i=1}^{b} N_i^{WS}\right) =& N\Pr(Z_i<\mathcal{C}_b^{WS})\\ =& \frac{1}{d}\int_{\mathcal{C}_0^{WS}}^{\mathcal{C}_b^{WS}} f_Z(z)dz.\end{aligned}$$

Note that the derivative with respect to $\mathcal{C}_b^{WS}$ is $\frac{1}{d}f_Z\left(\mathcal{C}_b^{WS}\right)>0$, so the expected number of observations in each class is not 0. This completes our proof in the univariate case. We leave the proof for multivariate cases to future research.

3.1 Derivation for Estimation –using NDTOs

Let us consider the placement of the *non-directly transferable observations*. We have seen that these observations belong to intervals, where the interval widths do not converge to zero. One way to proceed is to remove them completely so that they do not appear in the working sample. In practice, it seems that too many could fall into this category, resulting in a large efficiency loss. Another approach is to use the information available for these observations namely, the known boundary points for these values. Then we could use all the *directly transferable observations* from the working sample to calculate specific conditional averages for all *non-directly transferable observations* and replace them with those values. Let us denote a new variable $Z_{i,ALL}^{WS}$ that represents all the directly transferable observations and the replaced values for non-directly transferable observations.

For simplicity let us consider the case, when discretization happens with the explanatory variable. In this case, we simply need to calculate the conditional expectation that the underlying variable falls into the interval where NDTOs are. The other two cases follow the same logic but use different conditioning.

Let us formalize the non-directly transferable observations as $Z_i^{(s)} \in \mathcal{C}_\chi$, where

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

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

$$\hat{\nu}_\chi = \left(\sum_{i=1}^{N} \mathbf{1}_{\{Z_i^{DTO} \in \mathcal{C}_\chi\}} \right)^{-1} \sum_{i=1}^{N} \mathbf{1}_{\{Z_i^{DTO} \in \mathcal{C}_\chi\}} Z_i^{DTO}.$$

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

Under the WLLN and the same assumptions needed for the magnifying method, it is straightforward to show, $\hat{\nu}_\chi \to \mathbb{E}(Z | Z \in \mathcal{C}_\chi)$, as $N, S \to \infty$. Algorithm A8 shows how to replace NDTO values with the appropriate conditional expectation estimators.

Algorithm A8 The magnifying method - creation of "ALL" working sample

1: Let, $Z_i^{ALL} := \{Z_i^{DTO}\}$
2: Set, $m = 1, s = 1$
3: If $\mathcal{C}_m^{(s)} \in \mathcal{C}_\chi$, then calculate $\hat{\nu}_\chi$ and expand the working sample as,

$$Z_i^{ALL} := \left\{ Z_i^{ALL}, \bigcup_{j=1}^{N} \hat{\nu}_\chi \mid \left(Z_j^{(s)} \in \mathcal{C}_m^{(s)} \mid \mathcal{C}_m^{(s)} \in \mathcal{C}_\chi \right) \right\}$$

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

We can obtain the asymptotic standard errors of this estimator as if these are large, the replacement might not be favorable, as it induces more uncertainty relative to the potential loss of efficiency by not including all the observations. To obtain the standard errors, one can think of $\hat{\nu}_\chi$ as an LS estimator, regressing $\mathbf{1}_{\{Z_{i,DTO}^{WS} \in \mathcal{C}_\chi\}}$ on Z_i^{DTO}. Here $\mathbf{1}_{\{Z_i^{DTO} \in \mathcal{C}_\chi\}}$ is a vector of indicator variables, created by $2(S-1)$ indicator functions: It takes the value of one for the directly transferable observations, which are within $\mathcal{C}_\chi$.[4] We can now write the following:

$$Z_i^{DTO} = \boldsymbol{\nu}_\chi \mathbf{1}_{\{Z_i^{DTO} \in \mathcal{C}_\chi\}} + \eta_i,$$

[4]The indicator variables are not independent of each other, while the non-transferable observation intervals ($\mathcal{C}_\chi$) overlap each other.

where $\boldsymbol{\nu}_\chi$ stands for the vector of $\nu_\chi, \forall\chi$. The LS estimator of $\boldsymbol{\nu}_\chi$ is

$$\hat{\boldsymbol{\nu}}_\chi = \left(\mathbf{1}'_{\{Z_i^{DTO}\in\mathcal{C}_\chi\}}\mathbf{1}_{\{Z_i^{DTO}\in\mathcal{C}_\chi\}}\right)^{-1}\mathbf{1}'_{\{z_i^{DTO}\in\mathcal{C}_\chi\}}Z_i^{DTO}\,,$$

and under the standard LS assumptions, we can write

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

where $\boldsymbol{\nu}_\chi = \mathbb{E}(Z|Z\in\mathcal{C}_\chi), \forall\chi$. The variance of the OLS estimator is

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

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

4 Estimation with discretized regressors[5]

In this section, first, we analyze the bias and consistency of OLS estimation for β in the univariate case, when the explanatory variable is discretized. We investigate properties when $N\to\infty$ and when $M\to\infty$. The last implies we observe each value directly resulting in a consistent OLS estimator as we outlined in Section 2 from the main text. These exercises are helpful to see how these results generalize in the multivariate case discussed in Section 4.4. As a last subsection, we investigate the bias in the panel set up in Section 4.6.

Recall the data-generating process is assumed to be

$$Y_i = X_i'\beta + u_i \tag{S.6}$$

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

$$Y_i = X_i^{*'}\beta^* + u_i \tag{S.7}$$

It is also assumed there is a known support $[a_l, a_u]$ for X_i with known boundaries $(\mathcal{C}_m)$, and let v_m from Equation (1) from the main text be any value X_i^* take, typically the mid-point.

Let N_m be the number of observations in each class $\mathcal{C}_m$, that is $N_m = \sum_{i=1}^N \mathbf{1}_{\{X_i\in\mathcal{C}_m\}}$, where $\mathbf{1}_{\{X\in\mathcal{C}_m\}}$ denotes the indicator function. When X has a cumulative distribution (cdf) $F_X(\cdot)$,

[5]We acknowledge the work of Balázs Kertész from this section on the expected value of $\hat{\beta}^*_{OLS}$ and on N and M (in-)consistency results.

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

using the independence assumption. Note, when X has a uniform distribution, we have $\mathbb{E}(N_m) = N/M$ for all $m = 1, \ldots, M$.
The OLS estimator can expanded as,

$$\begin{aligned}\hat{\beta}^*_{OLS} &= \left(X^{*'}X^*\right)^{-1}(X^{*'}Y) \\ &= \frac{v_1\left(\sum_{i=1}^{N_1} Y_i\right) + v_2\left(\sum_{i=N_1+1}^{N_1+N_2} Y_i\right) + \cdots + v_M\left(\sum_{i=N-N_M+1}^{N_M} Y_i\right)}{N_1 v_1^2 + N_2 v_2^2 + \cdots + N_M v_M^2} \\ &= \frac{v_1\left(\sum_{i=1}^{N_1} \beta X_i + u_i\right) + \cdots + v_M\left(\sum_{i=N-N_M+1}^{N_M} \beta X_i + u_i\right)}{N_1 v_1^2 + \cdots + N_M v_M^2} \\ &= \frac{v_1\left[\sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_1\}}(\beta X_i + u_i)\right] + \cdots + v_M\left[\sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_M\}}(\beta X_i + u_i)\right]}{N_1 v_1^2 + \cdots + N_M v_1^2} \\ &= \frac{\sum_{m=1}^{M} v_m\left[\sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}}(\beta X_i + u_i)\right]}{\sum_{m=1}^{M} N_m v_m^2}\end{aligned}$$

Using the expression above, we can get the following general formula for the expected value

of the OLS estimator,

$$
\begin{aligned}
\mathbb{E}\left(\hat{\beta}^*_{OLS}\right) &= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} \left(\beta(X_i^* + \xi_i) + u_i\right)\right]}{\sum_{m=1}^{M} N_m v_m^2}\right\} \\
&= \mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m \left[\beta\left(\sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} X_i^* + \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} \xi_i\right) + \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} u_i\right]}{\sum_{m=1}^{M} N_m v_m^2}\right\} \\
&= \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} X_i^*}{\sum_{m=1}^{M} N_m v_m^2}\right\} + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} \xi_i}{\sum_{m=1}^{M} N_m v_m^2}\right\} \\
&\qquad + \mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} u_i}{\sum_{m=1}^{M} N_m v_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m \sum_{i=1}^{N} \mathbf{1}_{\{X_i \in \mathcal{C}_m\}} \xi_i}{\sum_{m=1}^{M} N_m v_m^2}\right\} \\
&= \beta + \beta\mathbb{E}\left\{\frac{\sum_{m=1}^{M} v_m N_m \upsilon^m}{\sum_{m=1}^{M} N_m v_m^2}\right\}. \qquad \text{(S.8)}
\end{aligned}
$$

where the discretization 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 $\mathcal{C}_m$, $\upsilon^m \overset{\mathrm{d}}{=} \xi_i | \mathcal{C}_m$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. Importantly, the second term in Equation (S.8) does not vanish in general, since $\upsilon^m | \mathcal{C}_m$ is not independent of $N_m | \mathcal{C}_m$, $\upsilon^m | \mathcal{C}_m \not\perp\!\!\!\perp N_m | \mathcal{C}_m$ nor $\mathbb{E}(\xi_i | \mathcal{C}_m) = \mathbb{E}(\upsilon^m) = 0$ (see Figure S.3, right panel for illustrative explanation). The former issue can be eliminated by conditioning on the underlying distribution of X_i. Conditional on the distribution X_i and the class $\mathcal{C}_m$, the number of observations in the class and assuming that the errors are independent of each other, $N_m | X_i, \mathcal{C}_m \perp\!\!\!\perp \upsilon^m | X_i, \mathcal{C}_m$, but knowing the underlying distribution makes the problem trivial. Nonetheless, because of both issues, the "naive" OLS estimator is biased.

Note that 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 S.3) if we are using the class midpoints. The first problem is resolved, because, in the case of the uniform distribution, both the number of observations N_m in each class $\mathcal{C}_m$ 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

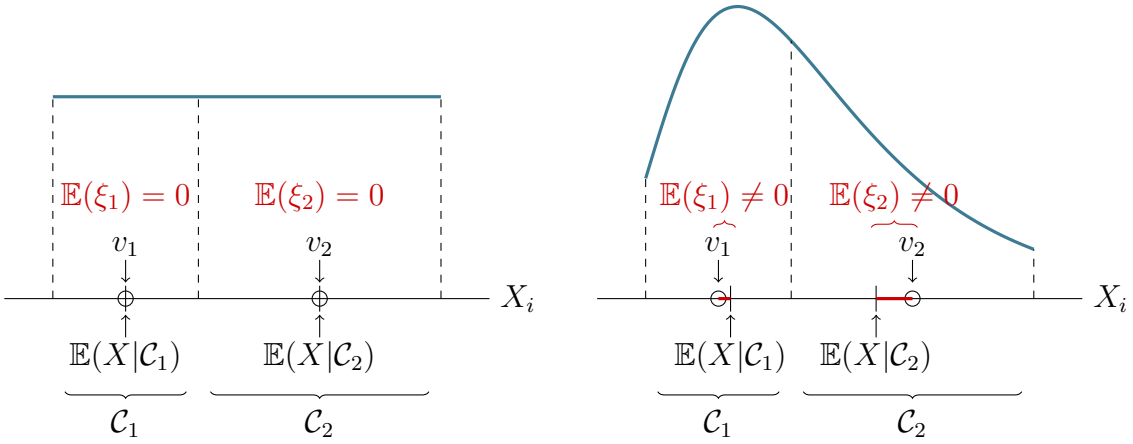

Figure S.3: The difference between uniform (left panel) and general distributions (right panel)

regressor's X_i expected value in the class $\mathcal{C}_m$. From Equation (S.8), we obtain that

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

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.

4.1 N (in)consistency

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

$$
\begin{aligned}
\operatorname*{plim}_{N\to\infty} \hat{\beta}^*_{OLS} &= \operatorname*{plim}_{N\to\infty} \frac{\sum_{m=1}^{M} v_m \left[\sum_{i=1}^{N} \mathbf{1}_{\{X_i\in\mathcal{C}_m\}}(\beta X_i + u_i)\right]}{\sum_{m=1}^{M} N_m v_m^2} \\
&= \frac{\sum_{m=1}^{M} v_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{X_i\in\mathcal{C}_m\}}(\beta X_i + u_i)\right]}{\sum_{m=1}^{M} v_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \frac{\sum_{m=1}^{M} v_m \left[\operatorname{plim}_{N\to\infty} \beta \sum_{i=1}^{N} \mathbf{1}_{\{X_i\in\mathcal{C}_m\}} X_i\right]}{\sum_{m=1}^{M} v_m^2 \operatorname{plim}_{N\to\infty} N_m} \\
&= \beta\, \frac{\sum_{m=1}^{M} v_m \left[\operatorname{plim}_{N\to\infty} \sum_{i=1}^{N} \mathbf{1}_{\{X_i\in\mathcal{C}_m\}} X_i\right]}{\sum_{m=1}^{M} v_m^2 \operatorname{plim}_{N\to\infty} N_m}.
\end{aligned} \tag{S.9}
$$

Define $X^m = \sum_{i=1}^{N} \mathbf{1}_{\{X_i\in\mathcal{C}_m\}} X_i$, then X^m sums the truncated version of the original random variables X_i on the class $\mathcal{C}_m$, $X_m \overset{\text{d}}{=} X_i|\mathcal{C}_m$, 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.9) equals β. To give a condition for consistency, first, we rewrite the previous Equation (S.9) in terms of the error terms ξ_i,

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

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

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

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^{*}_{OLS}-\beta\right) &= \frac{\operatorname{plim}_{N\to\infty}\beta\sum_{m=1}^{M}v_m\xi^m}{\operatorname{plim}_{N\to\infty}\sum_{m=1}^{M}v_m^2N_m} \\
&= \frac{\operatorname{plim}_{N\to\infty}O(N)\beta\sum_{m=1}^{M}v_m\xi^m/N_m}{\operatorname{plim}_{N\to\infty}O(N)\sum_{m=1}^{M}v_m^2} \\
&= \frac{\beta\sum_{m=1}^{M}v_m\operatorname{plim}_{N\to\infty}\xi^m/N_m}{\sum_{m=1}^{M}v_m^2}O(N) \\
&= \frac{\beta\sum_{m=1}^{M}v_m\left\{\mathbb{E}(z_m)-v_m\right\}}{\sum_{m=1}^{M}v_m^2}O(N). \qquad \text{(S.10)}
\end{aligned}$$

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.10) equals 0 to satisfy a sufficient condition.

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

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.

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.10) 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

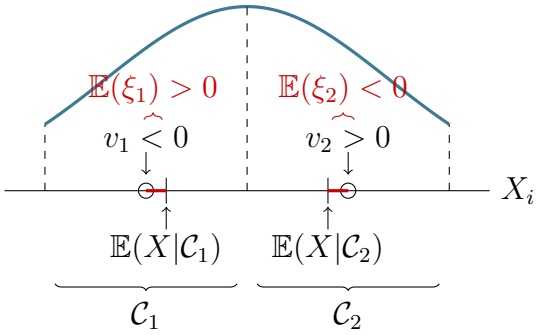

Figure S.4: The estimator is inconsistent even in case of symmetric distributions (see Equation (S.10)).

bias in N (see Figure S.4).

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.10). As a result, the difference between the expectation and the class midpoints, in general, is not zero for all m, hence the formula cannot be made arbitrarily small. Therefore, the OLS estimator becomes inconsistent in N.

So far we have focused on the estimation of β in Equation (S.7). But how about γ? It can be shown that the bias and inconsistency presented above are 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 interval-type variable in a model is going to infect the estimation of all variables of the model.

4.2 M Consistency

Let us see the case when N is fixed but $M \to \infty$. Now, we may have some intervals that do not contain any observations, while others still do. Omitting, however, empty intervals does not cause any bias because of our iid assumption. Furthermore, while we increase the number of intervals, the size of the intervals 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 interval boundaries. To see that, we derive the consistency formula in the number of intervals M assuming that $\operatorname{plim}_{M\to\infty} \sum_{\{m:\mathcal{C}_m\neq\emptyset, m=1,\ldots,M\}} v_m u_{i_m} = 0$, or with re-indexation $\operatorname{plim}_{M\to\infty} \sum_{i=1}^{N} v_{m_i} u_i = \sum_{i=1}^{N} z_i u_i = 0$, which should hold in the sample and is a stronger

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

$$\begin{aligned}
\operatorname*{plim}_{M\to\infty}\left(\hat{\beta}^{*}_{OLS}-\beta\right) &= \operatorname*{plim}_{M\to\infty}\frac{\sum_{m=1}^{M} v_m\left[\sum_{i=1}^{N}\mathbf{1}_{\{X_i\in\mathcal{C}_m\}}(\beta X_i+u_i)\right]}{\sum_{m=1}^{M} N_m v_m^2}-\beta \\
&= \operatorname*{plim}_{M\to\infty}\frac{\sum_{\{m:\mathcal{C}_m\neq\emptyset,m=1,\ldots,M\}} v_m\left[\sum_{i=1}^{N}\mathbf{1}_{\{X_i\in\mathcal{C}_m\}}(\beta X_i+u_i)\right]}{\sum_{\{m:\mathcal{C}_m\neq\emptyset,m=1,\ldots,M\}} N_m v_m^2}-\beta \\
&= \operatorname*{plim}_{M\to\infty}\frac{\sum_{\{m:\mathcal{C}_m\neq\emptyset,m=1,\ldots,M\}} v_m(\beta X_{i_m}+u_{i_m})}{\sum_{\{m:\mathcal{C}_m\neq\emptyset,m=1,\ldots,M\}} v_m^2}-\beta \\
&= \operatorname*{plim}_{M\to\infty}\beta\left\{\frac{\sum_{\{m:\mathcal{C}_m\neq\emptyset,m=1,\ldots,M\}} v_m X_{i_m}}{\sum_{\{m:\mathcal{C}_m\neq\emptyset,m=1,\ldots,M\}} v_m^2}-1\right\} \\
&= \operatorname*{plim}_{M\to\infty}\beta\left\{\frac{\sum_{i=1}^{N} v_{m_i} X_i}{\sum_{i=1}^{N} v_{m_i}^2}-1\right\} \\
&= \beta\left\{\frac{\sum_{i=1}^{N}\operatorname{plim}_{M\to\infty} v_{m_i} X_i}{\sum_{i=1}^{N}\operatorname{plim}_{M\to\infty} v_{m_i}^2}-1\right\} \\
&= \beta\left\{\frac{\sum_{i=1}^{N} X_i X_i}{\sum_{i=1}^{N} {X_i}^2}-1\right\} \\
&= 0,
\end{aligned}$$

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.

4.3 Some Remarks

The above results hold for much simpler cases as well. If instead of model (S.7) 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 variables, however, not only induces a 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.7). Consider a simple exam-

ple where there is an unobserved variable X_i with an observed discretized version:

$$X_i^* = \begin{cases} v_1 & \text{if } c_0 \leq X_i < c_1, \\ v_2 & \text{if } c_1 \leq X_i < c_2, \end{cases} \tag{S.11}$$

and

$$Y_i = X_i\beta + \varepsilon_i. \tag{S.12}$$

Using the discretized variable means:

$$Y_i = X_i^*\beta + (X_i - X_i^*)\beta + u_i \tag{S.13}$$

and

$$\begin{aligned} \mathbb{E}\left[X_i - X_i^*\right] =& \mathbb{E}(X_i) - \mathbb{E}(X_i^*) \\ =& \mathbb{E}(X_i) - \mathbb{E}\left[X_1\mathbf{1}(c_0 \leq X_i < c_1) + X_2\mathbf{1}(c_1 \leq X_i < c_2)\right] \\ =& \mathbb{E}(X_i) - v_1 \Pr(c_0 \leq X_i < c_1) - X_2 \Pr(c_1 \leq X_i < c_2). \end{aligned}$$

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

4.4 Estimation in multivariate regression

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}\,, \tag{S.14}$$

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 version $\mathbf{X}^*$ is. Define the $MK \times K$ matrix as

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

where $\mathbf{V}_i = (v_{i1}, \dots, v_{iM})'$ contains the values for variable i. Let $\mathbf{E} = \{\mathbf{e}_{ki}\}$, where $k =$

$1, \dots K$ and $i = 1, \dots, N$ such that

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

where x_{ki} denotes the value of the i^{th} observation from the explanatory variable $\mathbf{x}_k$.

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

4.5 The OLS Estimator

From Equation (S.14), consider the regression based on the observed data:

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + (\mathbf{X} - \mathbf{X}^*)\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{S.15}$$

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

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

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

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{V}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{V}\right)^{-1}\mathbf{V}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{V}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\mathbf{E}'\mathbf{V}\right)^{-1}\mathbf{V}'\mathbf{E}\mathbf{M}_{\mathbf{W}}\boldsymbol{\varepsilon}. \tag{S.16}$$

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

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

$$\mathbf{M}_{\mathbf{W}}\mathbf{X} = \mathbf{X} \tag{S.17}$$

$$\mathbf{M}_{\mathbf{W}}\mathbf{X}^* = \mathbf{X}^*. \tag{S.18}$$

In other words, we assume independence between $\mathbf{X}$ and $\mathbf{W}$, as well as its discretized 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 later stage.

The OLS estimator in this case becomes:

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

The OLS is unbiased if $(\mathbf{V}'\mathbf{E}\mathbf{E}'\mathbf{V})^{-1}\mathbf{V}'\mathbf{E}\mathbf{X} = \mathbf{I}$. Note that $\mathbf{V}'$ and $\mathbf{E}$ are of size $K \times MK$ and $MK \times N$, respectively. This means $\mathbf{V}'\mathbf{E}\mathbf{E}'\mathbf{V}$ are invertible as long as $N > K$, which is a standard assumption in classical regression analysis. Let us consider a typical element in $\mathbf{V}'\mathbf{E}\mathbf{E}'\mathbf{V}$ first. Since $\mathbf{V}$ is non-stochastic as it contains only all the pre-defined interval 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 $\mathbf{1}_m^{ki} = \mathbf{1}\left(c_{km-1} \leq z_{ki} < c_{km}\right)$, 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,

$$\mathbb{E}\left(\mathbf{1}_m^{ki}\mathbf{1}_n^{li}\right) = \int_{\Omega} f(x_k, x_l) dx_k dx_l \,, \tag{S.19}$$

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.19) under the usual WLLN.

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

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

Now,

$$\begin{aligned}\mathbb{E}\left[\sum_{m=1}^{M} v_{km}\mathbf{1}_m^{ki}x_{li}\right] &= \sum_{m=1}^{M} v_{km}\mathbb{E}\left[\mathbf{1}_m^{ki}x_{li}\right] \\ &= \sum_{m=1}^{M} v_{km}\int_{\Omega_1} x_l f(x_k, x_l)dx_k dx_l\,,\end{aligned} \tag{S.20}$$

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, $N^{-1}a_{kl}$ converge into Equation (S.20) under the usual WLLN.

In the case when Equations (S.17) and (S.18) do not hold, the analysis becomes more tedious algebraically, but it does not affect the result that OLS is biased. Recall Equation (S.16), and let ω_{ij} be the (i,j) element in $\mathbf{M_W}$ for $i = 1,\ldots,N$ and $j = 1,\ldots,J$, then following the same argument as above, $\mathbf{EM_W E}'$ 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

$$g_{kl} = \sum_{j=1}^{N}\sum_{i=1}^{N} \omega_{ij}\mathbf{1}_m^{ki}\mathbf{1}_n^{li} \tag{S.21}$$

with its expected value being

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{\Omega} \omega_{ij} f\left(x_k, x_l, \mathbf{W}\right) dw_k dw_k d\mathbf{W}, \tag{S.22}$$

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.22) is complicated. However, under the assumption that the integral in Equation (S.22) exits, then $N^{-1}g_{kl}$ should converge to Equation (S.22) 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 (S.22) to Equation (S.19).

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

$$h_{kl} = \sum_{m=1}^{M}\sum_{i=1}^{N} x_{km}\mathbf{1}_m^{ki}u_{li}, \tag{S.23}$$

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

$$\sum_{m=1}^{M} v_{km} \int_{\Omega_m} u_l f(x_k, x_l, \mathbf{W}) dx_k dx_l d\mathbf{W}, \tag{S.24}$$

where u_l denotes the random variable corresponding to the i^{th} column of $\mathbf{M_W X}$ and $\Omega_m = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}} \times \Omega_{\mathbf{X}}$ with $\Omega_{\mathbf{X}}$ denotes the sample space of $\mathbf{W}$. Note that $u_l = w_l$ under the assumption of independence, which reduces Equation (S.24) to Equation (S.20).

4.6 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. We can extend our DGP based on Equation (S.14), to

$$Y_{it} = X'_{it}\beta + \varepsilon_{it}, \tag{S.25}$$

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

Now, the most important problem is identification. If the interval for an individual does not change over the time periods covered, the individual effects in the panel and the parameter associated with the class variable cannot be identified separately. The within transformation would wipe out the interval variable as well. When the interval 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 the estimator can be easily shown. Let us re-write Equation (S.14) in a panel data context without further control variables $\mathbf{W}$.

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

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 to form $\boldsymbol{\alpha}$. The Within estimator is

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

or equivalently

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

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 $\mathbb{E}(\hat{\boldsymbol{\beta}}^*_X) \neq \boldsymbol{\beta}$, because $\mathbf{M}_{\mathbf{D}_N}\mathbf{E}'\mathbf{V} = \mathbf{M}_{\mathbf{D}_N}\mathbf{X}^* \neq \mathbf{M}_{\mathbf{D}_N}\mathbf{X}$.

5 Split sampling and perception effect

To extend our method towards fixed effect type of estimators, let us discuss a further phenomenon that we call the *perception effect*. The perception effect is relevant if the discretization happens through surveys. There is much evidence in the behavioral literature that the answers to a question may depend on the way the question is asked (see, e.g., Diamond and Hausman, 1994, Haisley et al., 2008 and Fox and Rottenstreich, 2003).[6] Note, that this is present regardless of whether split sampling has been performed or not. However, with split sampling, there is a way to tackle this issue, much akin to the approach a similar problem has been dealt with in the panel data literature.

5.1 Outcome variable

To sketch out the idea in the univariate case, let us define the *perception effect* B_s for split sample s, as

$$Z_i^{**} = \begin{cases} v_1^{(s)} & \text{if} \quad c_0^{(s)} < Z_i + B_s < c_1^{(s)} \\ \vdots & \vdots \\ v_m^{(s)} & \text{if} \quad c_{m-1}^{(s)} < Z_i + B_s < c_M^{(s)}\,. \end{cases} \tag{S.26}$$

Let Z_i^* be the same quantity, but $B_s = 0\,,\ \forall s$, thus no perception effect. $\tilde{Z}_i^*$ and $\tilde{Z}_i^{**}$ denote the replaced observations in the working sample that derived from Z_i^* and Z_i^{**}, respectively. Following the construction of the working sample,

$$\tilde{Z}_i^{**} = \tilde{Z}_i^* + B_s\,.$$

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

For example, in the case of discretization happening with the outcome variable, one can use a similar approach as outlined in Section 4.2 from the main text, and use the redefined equation,

$$\tilde{Y}_i^{**} = \tilde{Y}_i^{*} + B_s = \beta \tilde{X}_i + u_i \,.$$

Re-write the above in multivariate case,

$$\tilde{\mathbf{y}}^{*} + \mathbf{TB} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u},$$

where $\mathbf{B} = (B_1, \ldots, B_S)'$ and $\mathbf{T}$ is a $(N \times S)$ zero-one matrix that extracts the appropriate elements from $\mathbf{B}$.

Note, the estimator for $\mathbb{E}(\mathbf{y}|\mathbf{X} \in \mathcal{D}_\mathbf{l})$ needs to be adjusted for the above to hold if identification of $\boldsymbol{\beta}$ is based on the conditional expectation. The main challenge is to keep track of the perception effect. This means we need to identify each split sample and observation when estimating the conditional averages. Specifically,

$$N_l^{-1} \sum_{\mathbf{X} \in \mathcal{D}_\mathbf{l}, \mathbf{X} \in \mathbf{s}} \tilde{\mathbf{y}}^{**} - \mathbb{E}(\mathbf{y}|\mathbf{X} \in \mathcal{D}_\mathbf{l}) + B_s = o_p(1) \ ,$$

where N_l is the number of observations in the partitioned interval $\mathcal{D}_\mathbf{l}$.

Now, the estimation of $\boldsymbol{\beta}$ can be done in the spirit of a fixed effect estimator. Define the usual residual maker, $\mathbf{M_T} = \mathbf{I}_N - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'$, then

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\mathbf{M_T}\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'\mathbf{M_T}\tilde{\mathbf{y}}^{**} \tag{S.27}$$

is a consistent estimator of $\boldsymbol{\beta}$ given the results presented in this paper and similar arguments for the consistency of the standard fixed effect estimator in the panel data literature (see, e.g., Matyas 2024).

5.2 Explanatory variable

Let us continue when the explanatory variable is discretized. The discretization of X_i, is the same as defined in Equitation S.26, with B_s perception effect for split sample s. Let $\tilde{X}_i^{**} = \tilde{X}_i^{*} + B_s$ denote the observations in the working sample that derived from X_i^{*} and X_i^{**}, respectively. The model with perception effects is,

$$Y_i = \beta \tilde{X}_i^{**} + u_i = \beta \tilde{Z}_i^{*} + \beta B_s + u_i \,.$$

Extending to a multivariate case, requires specifying $\mathbf{B_X}$ the perception effect matrix with $S \times K$ dimensions, that allows for different effects for each split sample s and variable k in $\mathbf{X}$. Rewrite in matrix form results in

$$\mathbf{y} = \tilde{\mathbf{X}}^* \boldsymbol{\beta} + \mathbf{D}\mathbf{B}_\mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where $\mathbf{D}$ is a $N \times S$ zero-one matrix that extracts the appropriate elements from $\mathbf{B_X}$.

We need to modify the replacement estimator $\boldsymbol{\kappa}$ for the above to hold. We need to keep track of the perception effects, thus from which split sample each observation comes from when estimating the conditional averages. This implies,

$$\hat{\boldsymbol{\kappa}}_s = \left(\mathbf{1}'_{\{\mathbf{X} \in \mathcal{C}_\mathbf{m}^{(\mathbf{s})}, \mathbf{X} \in \mathbf{s}\}} \mathbf{1}_{\{\mathbf{X} \in \mathcal{C}_\mathbf{m}^{(\mathbf{s})}, \mathbf{X} \in \mathbf{s}\}} \right)^{-1} \mathbf{1}'_{\{\mathbf{X} \in \mathcal{C}_\mathbf{m}^{(\mathbf{s})}, \mathbf{X} \in \mathbf{s}\}} \tilde{\mathbf{X}}^{**} .$$

As $S, N \to \infty$

$$\hat{\boldsymbol{\kappa}}_s = \operatorname{vec}\left(\mathbb{E}\left[\mathbf{X} | \mathbf{X} \in \mathcal{C}_\mathbf{m}^{(\mathbf{s})} \right) + \mathbf{B}_\mathbf{X} \right] \right) + o_p(1) ,$$

where $\operatorname{vec}(\cdot)$ vectorize the conditional expectations similarly as in Appendix B.1. Note that to identify $\mathbf{B_X}$, we require variation along s and k, thus individual i shall face different split sample discretization for different variables. This is a mild condition and can be satisfied if the survey is constructed accordingly.

Now, the estimation of $\boldsymbol{\beta}$ can be done in the spirit of a fixed effect estimator. Define the usual residual maker, $\mathbf{M_D} = \mathbf{I}_N - \mathbf{D}\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'$, then

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}^{*\prime} \mathbf{M_D} \tilde{\mathbf{X}}^* \right)^{-1} \tilde{\mathbf{X}}^{*\prime} \mathbf{M_D} \mathbf{y} \tag{S.28}$$

is a consistent estimator of $\boldsymbol{\beta}$ following the similar argument.

Perhaps a more interesting question is the presence of perception effects over different m. In principle, this can also be incorporated by replacing B_s with B_{sm} for $s = 1, \ldots, S$ and $m = 1, \ldots, M$. Therefore, this particular setup does not just allow for perception effects due to different split samples, but rather, it provides a framework to investigate different types of perception effects. This would be an interesting avenue for future research in this area.

5.3 Both variables

Similarly, as before, let us define the observed discretized variables with perception effects in the univariate case,

$$\tilde{Y}_i^{**} = \tilde{Y}_i^* + B_Y \qquad \tilde{X}_i^{**} = \tilde{X}_i^* + B_X .$$

The model without further controls is,

$$\tilde{Y}_i^{**} = \beta \tilde{X}_i^{**} + u_i \,,$$

that is equivalent to

$$\begin{aligned} \tilde{Y}_i^{*} + B_Y &= \beta \tilde{X}_i^{*} + \beta B_X + u_i \\ \tilde{Y}_i^{*} &= \beta \tilde{X}_i^{*} + \beta B_X - B_Y + u_i \,. \end{aligned} \tag{S.29}$$

Equation (S.29) is interesting, as it allows for different perception effects in both Y_i and X_i variables in the univariate case. However, in such a setup, the parameters can not be identified in general. To show this, let us consider the matrix formulation,

$$\begin{aligned} \tilde{\mathbf{y}}^{*} + \mathbf{T}\mathbf{B}_Y &= \tilde{\mathbf{X}}^{*}\boldsymbol{\beta} + \mathbf{D}\mathbf{B}_{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u} \\ \tilde{\mathbf{y}}_1^{*} &= \tilde{\mathbf{X}}^{*}\boldsymbol{\beta} + \mathbf{D}\mathbf{B}_{\mathbf{X}}\boldsymbol{\beta} - \mathbf{T}\mathbf{B}_Y + \mathbf{u} \end{aligned}$$

where $\mathbf{B}_Y = (B_{Y,1}, \ldots, B_{Y,S_Y})'$. A unique solution for $\boldsymbol{\beta}$ exists if and only if $\mathbf{T}$ is orthogonal to $\mathbf{D}$. If so, one can use the corresponding residual maker $\mathbf{M_{TD}} = \mathbf{M_T}\mathbf{M_D}$, that yields in $\hat{\beta} = \left(\tilde{\mathbf{X}}^{*\prime}\mathbf{M_{TD}}\tilde{\mathbf{X}}^{*}\right)^{-1}\tilde{\mathbf{X}}^{*\prime}\mathbf{M_{TD}}\tilde{\mathbf{y}}^{*}$. Note however this assumption requires a careful survey design. Lastly, note that when constructing $\hat{\boldsymbol{\psi}}_Y$ and $\hat{\boldsymbol{\psi}}_X$, one needs to keep track of the split sample as well to ensure convergence.

5.4 Test for perception effect

It is theoretically possible to test the impacts of the perception effects on the estimator. Since $\hat{\boldsymbol{\beta}}$ as defined in Equation (S.27) is consistent regardless of the presence of perception effects. As,

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

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

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

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

6 Further Monte Carlo evidence

We extend the Monte Carlo simulations in five different ways. The basic setup is the same as in Section 5.1 from the main text, and we change each time one parameter compared to the basic setup. First, we investigate the effect of sample size on our shifting method, and how the magnitude of the bias changes when we use $N = 1,000$. As a second exercise, we investigate how the bias changes if the generated distributions are symmetric. As a third exercise, we check how the bias changes if instead of $M = 5$ we use only $M = 3$ intervals representing 'low-mid-high' categories. As the last exercise, we show some results on how the bias vanishes as we increase N and S, and the inconsistency of the alternative(s). All the following tables show the Monte Carlo average bias (or distortion) of $\hat{\beta}$ from $\beta = 0.5$. In parenthesis, we report the Monte Carlo standard deviation of the estimated parameter.

6.1 Explanatory variable

First, we investigate the case, when X_i is discretized, hence we only observe X_i^*.

6.1.1 Moderate sample size

For moderate sample size set $N = 1,000$. Table S.1 shows the results which are similar to the results with $N = 10,000$ as reported in the paper.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Mid-point regression	-0.0251	-0.0100	-0.0170	0.0003	0.0009	-0.0414
	(0.0175)	(0.0143)	(0.0159)	(0.0126)	(0.033)	(0.0229)
Shifting ($S = 10$)	-0.0002	0.0001	-0.0005	0.0002	0.0008	-0.0004
	(0.0179)	(0.0140)	(0.0155)	(0.0120)	(0.0287)	(0.0228)

Table S.1: Monte Carlo average bias and standard deviation with moderate sample size, $N = 1,000$, when discretization happens to the explanatory variable

Shifting method always outperforms the alternatives, except in the case of uniform and exponential, where there is no bias or small.

6.1.2 Symmetric boundaries

Next, we investigate symmetric boundary cases. We set the domain of the explanatory variable to $a_l = -2, a_u = 2$ and keep ε_i generated in the same way. For the log-normal,

exponential, and weibull cases, we truncate at 3 and subtract 1 from the generated distribution.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Mid-point regression	-0.0312	-0.0228	-0.0051	-0.0169	-0.0015	-0.0172
	(0.0067)	(0.0062)	(0.0092)	(0.006)	(0.0285)	(0.0174)
Shifting ($S = 10$)	0.0001	-0.0001	0.0002	0.0000	0.0006	0.0002
	(0.0066)	(0.0062)	(0.0087)	(0.0059)	(0.0242)	(0.0157)

Table S.2: Monte Carlo average bias and standard deviation with symmetric boundary points: $a_l = -2, a_u = 2$, when discretization happens to the explanatory variable

In this case the bias is even more severe for the mid-point regression than in the assymetric case. This relates to the distance of the midpoints and the actual excepted values within the intervals. Shifting performs in this setting well and the bias vanishes.

6.1.3 Number of intervals (M)

Another question is how the number of intervals (M) affects the bias. In this exercise, we investigated the $M = 3$ case, where interval defines (known) low-mid-high ranges. In general, the bias increases for the methods, however, it shows up in a larger standard deviation.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Mid-point regression	-0.0252	-0.0101	-0.0174	0.0002	0.0005	-0.0422
	(0.0057)	(0.0046)	(0.0051)	(0.0040)	(0.0102)	(0.0073)
Shifting ($S = 10$)	0.0002	0.0001	0.0000	0.0002	0.0005	0.0002
	(0.0056)	(0.0044)	(0.0049)	(0.0038)	(0.009)	(0.0072)

Table S.3: Monte Carlo average bias and standard deviation with small number of interval options, $M = 3$, when discretization happens to the explanatory variable

6.1.4 Convergence in N

Table S.4 shows the (asymptotic) reduction in the bias with the split sampling method. We use now only normal distribution's setup for ε_i. The shifting method decreases the bias towards zero, but one needs higher N and S as well. Mid-point regression remains biased regardless of N.

		$N = 1,000$	$N = 10,000$	$N = 100,000$
Midpoint regression		-0.1053	-0.105	-0.1044
		(0.0521)	(0.0166)	(0.0052)
Shifting	$S = 3$	-0.0252	-0.0269	-0.0259
		(0.0610)	(0.0193)	(0.0059)
	$S = 5$	-0.0158	-0.0162	-0.0155
		(0.0594)	(0.0188)	(0.0060)
	$S = 10$	-0.0115	-0.0101	-0.0097
		(0.0586)	(0.0189)	(0.0058)
	$S = 25$	-0.0085	-0.0073	-0.0067
		(0.0587)	(0.0188)	(0.0058)
	$S = 50$	-0.0066	-0.0053	-0.0049
		(0.0577)	(0.0189)	(0.0058)
	$S = 100$	-0.0048	-0.0041	-0.0037
		(0.0579)	(0.0186)	(0.0058)

Table S.4: Bias reduction for split sampling methods: different sample sizes and number of split samples, when discretization happens to the explanatory variable

6.2 Outcome variable

In the following, we provide further evidence on the bias reduction, when discretization happens on the left-hand side, thus to the outcome variable Y_i.

Notes: In the case of *"Set identification"* † shows that we can only estimate the lower and upper boundaries for the valid parameter set. We report these bounds subtracted with the true parameter, therefore it should give a (close) interval around zero. For ordered choice models * shows we report the distortion from the true β is reported. Ordered probit and logit models' maximum likelihood parameters do not aim to recover the true β parameter, therefore it is not appropriate to call it bias.

6.2.1 Moderate sample size

First, we investigate the magnitude of the bias, when the sample size is moderate, namely $N = 1,000$. Table S.5 shows the results which are similar to the results with $N = 10,000$ as reported in the paper.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.1, 1.15]$	$[-1.09, 1.15]$	$[-1.09, 1.16]$	$[-1.07, 1.17]$	$[-1.06, 1.18]$	$[-1.09, 1.15]$
	(0.06),(0.07)	(0.08),(0.08)	(0.07),(0.07)	(0.09),(0.09)	(0.08),(0.09)	(0.05),(0.06)
Ordered probit*	0.1978	0.0690	0.2138	0.0181	0.0965	0.4484
	(0.0810)	(0.0797)	(0.0827)	(0.0763)	(0.0795)	(0.0908)
Ordered logit*	0.6523	0.3828	0.6967	0.2419	0.4309	1.2109
	(0.1479)	(0.1431)	(0.1561)	(0.1364)	(0.1455)	(0.1682)
Interval regression	0.0254	0.0329	0.0398	0.0512	0.0638	0.0396
	(0.0618)	(0.0784)	(0.0694)	(0.0882)	(0.0825)	(0.0505)
Midpoint regression	0.0209	0.0293	0.0310	0.0453	0.2029	0.0275
	(0.0643)	(0.0786)	(0.0733)	(0.0895)	(0.0426)	(0.0526)
Shifting ($S = 10$)	-0.0043	-0.0021	-0.0036	-0.0014	-0.0019	-0.0019
	(0.0611)	(0.0758)	(0.0685)	(0.0869)	(0.0389)	(0.0475)

Table S.5: Monte Carlo average bias and standard deviation with moderate sample size, $N = 1,000$, when discretization happens to the outcome variable

Shifting method always outperforms the alternatives.

6.2.2 Symmetric boundaries

Next, we investigate symmetric boundary cases. We set the domain of the outcome variable to $a_l = -3, a_u = 3$ and keep X_i generated in the same way. ε_i is generated/truncated such that its lower and upper bound is -2 and 2. In the normal, logistic, and uniform cases, it means the lower and upper bounds are -2 and 2. For the log-normal, exponential, and weibull cases, we truncate at 4 and subtract 2 from the generated distribution.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.11, 1.13]$	$[-1.15, 1.10]$	$[-1.09, 1.16]$	$[-1.07, 1.17]$	$[-1.06, 1.19]$	$[-1.09, 1.15]$
	(0.02),(0.02)	(0.02),(0.02)	(0.02),(0.02)	(0.03),(0.03)	(0.03),(0.03)	(0.02),(0.02)
Ordered probit*	0.0890	0.0029	0.2085	0.0158	0.0986	0.4461
	(0.0252)	(0.0243)	(0.0262)	(0.0234)	(0.0241)	(0.0295)
Ordered logit*	0.4513	0.3198	0.6862	0.2379	0.4338	1.2085
	(0.0446)	(0.0427)	(0.0499)	(0.0422)	(0.044)	(0.0546)
Interval regression	0.0085	-0.0267	0.0371	0.0491	0.0663	0.0397
	(0.022)	(0.0234)	(0.0221)	(0.0271)	(0.0249)	(0.0166)
Midpoint regression	0.0070	0.0240	0.0362	0.0490	0.2077	0.0314
	(0.0211)	(0.0242)	(0.0216)	(0.0273)	(0.0128)	(0.0157)
Shifting ($S = 10$)	-0.0001	0.0004	-0.0001	-0.0007	-0.0015	-0.0001
	(0.0199)	(0.0232)	(0.0204)	(0.0262)	(0.0115)	(0.0140)

Table S.6: Monte Carlo average bias and standard deviation with symmetric boundary points: $a_l = -3, a_u = 3$, when discretization happens to the outcome variable

As we expected the maximum likelihood methods, have a closer fit to the assumed distribution the distortion is somewhat smaller in the case of ordered probit model[7]. This is the case with the normal and logistic distributions for the disturbance term. However, the distortion remains with the same magnitude for all the other misspecified cases. The shifting method outperforms all other methods.

6.2.3 Number of intervals (M)

We investigated the $M = 3$ case, where interval defines (known) low-mid-high ranges. In general, the bias increases for the methods. Interesting exceptions are interval regression and midpoint regression, where the results become more volatile: in some cases, they give better results, while in others even worse. The shifting method gives fairly accurate estimates.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.83, 1.90]$	$[-1.85, 1.88]$	$[-1.85, 1.89]$	$[-1.87, 1.87]$	$[-1.89, 1.85]$	$[-1.81, 1.93]$
	(0.03),(0.03)	(0.03),(0.03)	(0.03),(0.03)	(0.03),(0.03)	(0.03),(0.03)	(0.02),(0.02)
Ordered probit*	0.1062	-0.0220	0.0197	-0.1028	-0.0752	0.2347
	(0.0272)	(0.0266)	(0.0278)	(0.0250)	(0.0253)	(0.0302)
Ordered logit*	0.5193	0.3220	0.3916	0.1700	0.2169	0.7246
	(0.0462)	(0.0457)	(0.0472)	(0.0423)	(0.0428)	(0.0509)
Interval regression	0.0124	0.0124	0.0122	-0.0044	-0.0243	-0.0061
	(0.0224)	(0.0281)	(0.0268)	(0.0306)	(0.0280)	(0.0200)
Midpoint regression	0.0336	0.0168	0.0229	-0.0011	-0.2026	0.0647
	(0.0233)	(0.0274)	(0.0267)	(0.0307)	(0.0170)	(0.0216)
Shifting ($S = 10$)	-0.0274	-0.0114	-0.0009	-0.0027	0.0011	-0.0008
	(0.0237)	(0.0256)	(0.0226)	(0.0277)	(0.0135)	(0.0151)

Table S.7: Monte Carlo average bias and standard deviation with small number of interval options, $M = 3$, when discretization happens to the outcome variable

Also note that with the shifting method, the average bias is within 1 standard deviation, which is not true for the other methods, especially when the underlying distribution is exponential or weibull.

[7]Note that ordered probit and logit uses different scaling (depending on the assumed distribution), which results in different parameter estimates. In our case it means ordered logit has higher average distortions than ordered probit, but this is only a matter of scaling. One can map one to the other with the scaling factor, $\hat{\beta}^{ML}_{probit} \approx \hat{\beta}^{ML}_{logit} \times 0.25/0.3989$. This is why we use the term distortion rather than bias for these methods.

6.2.4 Convergence in N

Table S.8 shows the (asymptotic) reduction in the bias with the split sampling methods. We use now only normal distribution's setup for ε_i. As Table S.8 suggests, as we increase the number of observations the bias vanishes for the shifting method. Also if we increase the number of split samples the bias tends to decrease. It is important to highlight the other methods' bias/distortion remains the same as we increase the number of observations, therefore they give inconsistent estimates.

		$N = 1,000$	$N = 10,000$	$N = 100,000$
Set identification†		$[-1.1, 1.15]$ ((0.06),(0.07))	$[-1.1, 1.15]$ ((0.02),(0.02))	$[-1.1, 1.15]$ ((0.01),(0.01))
Ordered probit*		0.1978 (0.0810)	0.1971 (0.0256)	0.1968 (0.0080)
Ordered logit*		0.6523 (0.1479)	0.6509 (0.0464)	0.6502 (0.0146)
Interval regression		0.0254 (0.0618)	0.0268 (0.0198)	0.0266 (0.0062)
Midpoint regression		0.0257 (0.0635)	0.0251 (0.0195)	0.0251 (0.0061)
Shifting	$S = 3$	-0.0019 (0.0635)	0.0014 (0.0197)	-0.0008 (0.0062)
	$S = 5$	-0.0016 (0.0614)	-0.0007 (0.0189)	-0.0005 (0.0060)
	$S = 10$	-0.0067 (0.0605)	-0.0025 (0.0190)	-0.0006 (0.0059)
	$S = 25$	0.0052 (0.0602)	0.0008 (0.0185)	-0.0001 (0.0057)
	$S = 50$	-0.0027 (0.0587)	-0.0011 (0.0185)	-0.0004 (0.0058)
	$S = 100$	-0.0006 (0.0596)	-0.0002 (0.0183)	-0.0002 (0.0057)

Table S.8: Bias reduction for split sampling methods: different sample sizes and number of split samples, when discretization happens to the outcome variable

6.3 Both side

In our final simulations, we investigate the properties of bias when both outcome and explanatory variables are discretized.

6.3.1 Moderate sample size

For moderate sample size set $N = 1,000$. Table S.9 shows the results which are similar to the pattern with $N = 10,000$ as reported in the paper.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Mid-point regression	-0.0856	-0.0797	-0.0759	-0.0647	0.0809	-0.0771
	(0.0571)	(0.0691)	(0.0616)	(0.0790)	(0.0377)	(0.0456)
Shifting ($S = 10$)	0.0139	0.0094	0.0042	0.0064	0.0001	0.0035
	(0.0607)	(0.0768)	(0.0668)	(0.0862)	(0.0364)	(0.0452)

Table S.9: Monte Carlo average bias and standard deviation with moderate sample size, $N = 1,000$, when discretization happens on both sides

Shifting method always outperforms the alternative mid-point regression.

6.3.2 Symmetric boundaries

For the symmetric boundary case, we set the domain of the disturbance term to $a_l = -2, a_u = 2$ and keep X_i generated in the same way. Now the outcome variable's domain is between -3 and 3. For the log-normal, exponential, and weibull cases, we truncate at 3 and subtract 1 from the generated distribution.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Mid-point regression	-0.0975	-0.0838	-0.0752	-0.0635	0.0797	-0.0759
	(0.0189)	(0.0217)	(0.019)	(0.0243)	(0.0116)	(0.0137)
Shifting ($S = 10$)	0.0088	0.0075	0.0056	0.0057	0.0026	0.0047
	(0.0205)	(0.0237)	(0.02211)	(0.0264)	(0.0118)	(0.0142)

Table S.10: Monte Carlo average bias and standard deviation with symmetric boundary points: $a_l = -2, a_u = 2$, when discretization happens on both sides

Results are similar to the reported table in the main paper.

6.3.3 Number of intervals (M)

We investigated the $M = 3$ case, where interval defines (known) low-mid-high ranges. In general, the bias increases for the methods. Shifting gives closer results to zero bias.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Mid-point regression	-0.1998	-0.2091	-0.2221	-0.2143	-0.3552	-0.2012
	(0.0177)	(0.0210)	(0.0198)	(0.0235)	(0.0108)	(0.0154)
Shifting ($S = 10$)	-0.1062	-0.0168	-0.0123	0.0142	-0.0049	0.0092
	(0.0269)	(0.0301)	(0.0257)	(0.0304)	(0.0137)	(0.0162)

Table S.11: Monte Carlo average bias and standard deviation with small number of interval options, $M = 3$, when discretization happens on both sides

6.3.4 Convergence in N

Table S.12 shows the (asymptotic) reduction in the bias with the split sampling methods. We use now only normal distribution's setup for ε_i. Here the bias is not decreasing as quickly as in the other cases.

		$N = 1,000$	$N = 10,000$	$N = 100,000$
Midpoint regression		-0.1562	-0.1547	-0.1543
		(0.0517)	(0.0162)	(0.0051)
Shifting	$S = 3$	-0.0694	-0.0609	-0.0597
		(0.0633)	(0.0205)	(0.0064)
	$S = 5$	-0.0543	-0.0515	-0.0505
		(0.0606)	(0.0198)	(0.0062)
	$S = 10$	-0.0539	-0.0481	-0.0463
		(0.0594)	(0.0194)	(0.0059)
	$S = 25$	-0.0506	-0.0471	-0.0462
		(0.0593)	(0.0188)	(0.0060)
	$S = 50$	-0.0520	-0.0482	-0.0470
		(0.0570)	(0.0186)	(0.0058)
	$S = 100$	-0.0520	-0.0487	-0.0470
		(0.0564)	(0.0183)	(0.0057)

Table S.12: Bias reduction for split sampling methods: different sample sizes and number of split samples, when discretization happens on both sides

7 Gender wage gap in detail

To demonstrate how our method works in practice, we need a dataset where we can measure the difference between the parameter estimated on a non-discretized variable and the parameter(s) using some discretized version of the data. The Australian Tax Office's (ATO) individual sample files record income and some basic socio-economic variables for a 2% sam-

ple of the whole population.[8] To contrast this non-discretized income variable with practice, we use different discretization processes. We employ a simple equally distanced discretization method and a specific method, which is used in the Household, Income, and Labour Dynamics in Australia (HILDA) Survey [9]. HILDA is an annual survey and dataset, which is well known and widely used in Australia for economic research [10]. This way we can estimate parameters on the complete sample and the parameter estimates on "what if it is observed through a discretization process".

7.1 Data

The Australian Tax Office (ATO) dataset is a confidentialised 2% sample of the whole population. It records individual income tax returns for various income for separate years. We use data from 2016-17, which contains overall 277,202 records. Our outcome variable is yearly earned wage in the Australian dollar. Our parameter of interest is the coefficient for the gender variable. Further variables are,

- Expected age for each age group
  - We have calculated the expected age conditional on the age groups. This is necessary, while the ATO dataset only uses age groups: $0-20, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70+$. To circumvent this discretization process, we are using the Australian Bureau of Statistics on demographic statistics[11], which contains the number of males and female for each age. Based on this we calculate the conditional expected values for the year 2016-17, conditioning on gender.
- occupation code (as a series of dummies)
- spouse (dummy)
- region (dummies)
- lodgment method (dummy - via tax agent or self-prepared return)
- private health insurance (PHI) indicator (dummy)

[8] For details see ATO's website: https://www.ato.gov.au/about-ato/research-and-statistics/in-detail/taxation-statistics/taxation-statistics-previous-editions/taxation-statistics-2016-17 .

[9] https://melbourneinstitute.unimelb.edu.au/hilda

[10] see:https://melbourneinstitute.unimelb.edu.au/hilda/publications

[11] File,31010$DO002_201906$, available at https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3101.0Jun%202019?OpenDocument

	Mean	Median	Std	Min	Max
wage	77,808	68,864	35,369	34,981	22,4951
total income	83,069	71,337	55,609	-13,5989	3,001,331
exp. age	46.77	47.02	10.78	27.00	61.95

Positive income	Total income>0	Total income≤ 0
	0.9998	0.0002
Gender	Male	Female
	0.5598	0.4402
Spouse	No	Yes
	0.3867	0.6133
LM	Tax agent	Self preparer
	0.7304	0.2696
PHI	No	Yes
	0.3562	0.6438

Occ. codes*	0	1	2	3	4
	0.0005	0.1537	0.2759	0.1320	0.0774
	5	6	7	8	9
	0.1392	0.0485	0.0681	0.0714	0.0333

Region. codes†	New South Wales				
	Capital	Other Urban	R.H.U.	R.L.U.	Rural
	0.1971	0.0412	0.0388	0.0169	0.0250
	Queensland				
	Capital	Other Urban	R.H.U.	R.L.U.	Rural
	0.0970	0.0522	0.0131	0.0081	0.0290
	Tasmania				
	Capital	Other Urban	R.H.U.	R.L.U.	Rural
	0.0055	0.0026	0.0026	0.0018	0.0061
	Victoria				
	Capital	Other Urban	R.H.U.	R.L.U.	Rural
	0.1830	0.0128	0.0193	0.0120	0.0249
	Western Australia				
	Capital	Other Urban	R.H.U.	R.L.U.	Rural
	0.0716	0.0030	0.0159	0.0066	0.0125
	ACT	South Australia			
		Capital	R.H.U.	R.L.U.	Rural
	0.0218	0.0472	0.0060	0.0043	0.0079
	Northern Territory			Overseas/	
	Capital	R.H.U.	R.L.U.	invalid	
	0.0057	0.0033	0.0026	0.0026	

** Occupation codes: 0 - Occupation not listed/ Occupation not specified, 1 - Managers, 2 - Professionals, 3 - Technicians and Trades Workers, 4 - Community and Personal Service Workers, 5 - Clerical and Administrative Workers, 6 - Sales workers, 7 - Machinery operators and drivers, 8 - Labourers, 9 - Consultants, apprentices, and type not specified or not listed*

† : R.H.U: regional high urbanization, R.L.U: regional low urbanization.

Table S.13: Summary table for used variables

We restrict our sample to the working population (older than 25 and younger than 65 years old) and we remove individuals whose wage is lower than the 2017 minimum wage (weekly minimum wage was 627.7 AUD) and whose wage is more than 225.000AUD (that is the top 1%). Table S.13 shows basic descriptive statistics on the used variables in our sample.

7.2 Discretization process and model results

The discretization process influences the magnitude of the bias. We have fixed the lower bound to zero and the upper bound to $225,000$ for the wages. We use two different discretization processes:

- $M = 10$ with equal distances for mid-point regression and shifting as it was the closer in our reported main result in Table 1 from the main text.
- HILDA's household questionnaire, which uses $M = 12{+}1$ categories in 2017[12]: $1-9.999$, $10.000-19.999$, $20.000-29.999$, $30.000-39.999$, $40.000-49.999$, $50.000-59.999$, $60.000-79.999$, $80.000-99.999$, $100.000-124.999$, $125.000-149.999$, $150.000-199.999$ and 200.000 or more. Three extra options are added: negative or zero refused, and don't know.

We need to note that HILDA is aiming for total income and not for wages/salaries, thus we discretize the income jointly with the wages. For the shifting method, we use equal distances and check for $S = 10$ during the modeling.

Our outcome variable is yearly wage in Australian dollar and our parameter of interest is the coefficient for the gender dummy. We compare the conditional average outcome based on gender, using three different linear models, all of them estimated by OLS,

- Model 1: $y_i = \alpha + \beta \times gender_i + \epsilon_i$
- Model 2: $y_i = \alpha + \beta \times gender_i + \gamma_1 age_i + \gamma_2 age_i^2 + \varepsilon_i$
- Model 3: adding further controls for design 2, occupation code (dummies), having a spouse (dummy), region (dummies), and whether having private health insurance (dummy)
- Model 4: Model 3 and interaction terms of occupation and age

[12] https://melbourneinstitute.unimelb.edu.au/__data/assets/pdf_file/0005/2409674/HouseholdQuestionnaireW17M.pdf

Discretization	Model 1	Model 2	Model 3	Model 4
Non-discretized	-0.2027	-0.2025	-0.2277	-0.2261
	(0.0023)	(0.0023)	(0.0023)	(0.0023)
HILDA	-0.2012	-0.2010	-0.2258	-0.2242
	0.0023	0.0023	0.0023	0.0023
Mid-point regression	-0.2124	-0.2122	-0.2381	-0.2364
	(0.0025)	(0.0024)	(0.0025)	(0.0025)
Shifting method (S=10)	-0.2040	-0.2038	-0.2296	-0.2280
	0.0024	0.0024	0.0024	0.0024

Non-discretized row uses the actual wage data. HILDA uses its special 13-category discretization outlined above. Mid-point regression and shifting method use $M = 10$ intervals. Shifting uses $S = 10$ split samples. Standard errors are in parenthesis.

Table S.14: Estimated $\hat{\beta}$ parameters for gender dummy with different model specifications

Table S.14 shows that the shifting method gives close and statistically non-distinguishable estimates of the non-discretized parameter value. HILDA performs similarly well, while mid-point regression gives statistically different parameter values at 5% in almost all cases.

References

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

Fox, C.R., Rottenstreich, Y., 2003. Partition priming in judgment under uncertainty. Psychological Science 14, 195–200.

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

Matyas, L. (Ed.), 2024. The econometrics of multi-dimensional panels. Springer Nature.