10.14754/CEU.2021.06

Three Essays on Econometrics

by

Ágoston Reguly

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy at
Central European University

Supervisors: Róbert Lieli, László Mátyás

Budapest, Hungary

© Copyright by Ágoston Reguly, 2021.
All rights reserved.

CEU eTD Collection

10.14754/CEU.2021.06

CENTRAL EUROPEAN UNIVERSITY
DEPARTMENT OF ECONOMICS AND BUSINESS

Author: Ágoston Reguly
Title: Three Essays on Econometrics
Degree: Ph.D.
Dated: October 6, 2021

Hereby I testify that this thesis contains no material accepted for any other degree in any other institution and that it contains no material previously written and/or published by another person except where appropriate acknowledgements is made.

Signature of the author:

Ágoston Reguly

CEU eTD Collection

10.14754/CEU.2021.06

Disclosure of coauthor contribution

Modelling with Discretized Continuous Covariate

and

Modelling with Discretized Dependent Variable

Co-authors: Felix Chan and László Mátyás

László Mátyás came up with the problem of biased estimates when one uses discretized variables as covariates in a linear model. All of the authors contributed substantially to develop the theoretical framework of split sampling to handle the problem. The split sampling idea came from László Mátyás, the development, and implementation of the magnifying and shifting methods and writing the paper were largely carried out by Ágoston Reguly, while the theoretical proofs were mainly derived by Felix Chan.

CEU eTD Collection

10.14754/CEU.2021.06

Abstract

The thesis consists of three chapters: the first, single-authored chapter proposes a supervised machine learning algorithm to discover heterogeneous treatment effects in regression discontinuity designs. The second and third chapter, co-authored with László Mátyás and Felix Chan proposes a new data gathering method, which allows the identification and consistent estimation of parameters in the linear regression model when variables are observed through a discretization process. The second chapter discusses the so called split sampling data gathering method in detail and investigates the properties of the least squares estimator when the discretized variable is on the right hand side. Chapter 3 discusses the identification and estimation when the discretized variable is on the left hand side.

Chapter 1: Heterogeneous Treatment Effects in Regression Discontinuity Designs

The paper proposes a supervised machine learning algorithm to uncover treatment effect heterogeneity in classical regression discontinuity (RD) designs. Extending Athey and Imbens (2016), I develop a criterion for building an honest "regression discontinuity tree", where each leaf of the tree contains the RD estimate of a treatment (assigned by a common cutoff rule) conditional on the values of some pre-treatment covariates. It is *a priori* unknown which covariates are relevant for capturing treatment effect heterogeneity, and it is the task of the algorithm to discover them, without invalidating inference. I study the performance of the method through Monte Carlo simulations, and apply it to the data set compiled by Pop-Eleches and Urquiola (2013) to uncover various sources of heterogeneity in the impact of attending a better secondary school in Romania.

Chapter 2: Modelling with Discretized Continuous Covariate

with Felix Chan and László Mátyás

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

CEU eTD Collection

10.14754/CEU.2021.06

Chapter 3: Modelling with Discretized Continuous Dependent Variable

with Felix Chan and László Mátyás

The paper deals with econometric models where the dependent variable is continuous but cannot be observed directly. Instead, it is observed through intervals or discretized ordered choice windows. Manski and Tamer (2002) show that the parameters in the conditional expectation cannot be point-identified using these discretized observations. Here we introduce a new sampling design, the so-called *split sampling*, which makes the point-identification of the parameters in regression models feasible. Split sampling yields point-identification through the way information is collected. The target sample set is split into multiple parts and data is collected in a differentiated way. We explore how split sampling affects statistical inference, and further Monte Carlo evidence is provided about its effect on estimation. Finally, we propose a simple formulation to deal with an eventual perception effect.

CEU eTD Collection

10.14754/CEU.2021.06

Acknowledgements

It is one of the most pleasing and one of the hardest tasks to write the Acknowledgements. It would have been impossible for me to write this thesis without the support of many people. I try to do my best to thank these people who helped me through this process, however, it is impossible to mention everybody.

First of all, I am indebted to my advisors Róbert Lieli and László Mátyás for their constant support and guidance they offered me throughout my studies. I am especially thankful to László Mátyás for his openness to write papers together, introducing me to Felix Chan, his trust and generosity with supporting my travel to the University College of London and conferences and his agile responses while we carried out the research. I am similarly grateful to Róbert Lieli for his patience and his meticulous supervision. He relentlessly pushed me forward towards my academic career and helped with many applications.

Working with Felix Chan was an experience that inspired me to pursue excellence in econometric theory. His attitude toward me, his encouragement, and being treated as equals at a young age gave me a lot.

I am grateful to Gábor Békés, who put faith in me and made it possible for me to teach data analysis related courses at CEU. His confidence and trust was a truly life-changing experience. He has not only led me into how and what to teach, but supported me through the whole process. I have gained a tremendous amount of experience in teaching and learned a lot from my students. Also, Gábor was always open for discussion in various topics in econometrics all of which helped to widen my perspective.

I am thankful to the faculty of the Department of Economics and Business at CEU, they have set up an environment, where my research could open up and I got much valuable feedback. I am grateful for all the comments and discussion with Arieda Muço, Andrea Weber, Sergey Lychagin, Tímea Molnár, Miklós Koren, Botond Kőszegi, Marc Kaufmann, and from the participants of Brown Bag Seminars. I am especially thankful for Péter Szilágyi for his support in teaching and guidance in pursuing an academic career.

I am grateful to the examiners of my thesis, Gábor Békés, and Michael Knaus, who gave constructive comments that helped me a lot to improve the quality of my papers. I also appreciate the work of the other members of the Committee, Andrea Weber (chair) and Gábor Körösi (external member) who made my defense possible.

I learned a lot from the discussion of my fellow Ph.D. students and they always provided me a helpful hand during my studies. Among many, Jenő Pál, János Divényi, István Boza, Luca Drucker, Olivér Kiss, Balázs Kertész, Balázs Krusper, Boldizsár Juhász, Lili Márk, Balázs Vonnák, Ceyda Ustun, Gábor Révész, András Kollarik and Ákos Aczél. I would like to particularly mention Lajos T. Szabó whose friendship and continuous support enabled me to do my Ph.D. with such enthusiasm.

Veronika Orosz, Márt Jombach, Melinda Molnár, Katalin Szimler, Judit Lafferthon, Eszter Fuchs, and Zsuzsanna Bordás, staff members at the Department of Economics and

CEU eTD Collection

Business, have always been very helpful, kind and supported me to tackle many administrative burdens.
I could not start my Ph.D. at CEU without the support of my former professors or colleagues. I am thankful for Meyer Dietmar, who showed me the academic world and the way of carrying out research I am also thankful for Zsuzsanna Mosolygó for her openness, trust, and encouragement at the Hungarian Government Debt Management Agency, where I have worked for more than three years. I am also grateful for the recommendation and support of Mihály Ormos, László Borbély and Ádám Török to pursue a Ph.D. degree at CEU.
The loving environment of my family, where I have grown up was essential for me to be in this position, defending my Ph.D. My parents, my brothers, my sister, grandparents and in-laws encouraged and inspired me during the process and stood behind me at all times for which I am indebted. I am delighted when I think of the many creative discussion with my family on various topics, the help of my father and older brother at the early stages of my research, and my mother's passionate and creative upbringing. I am also grateful for the community of Regnum Marianum, where I have grown up, found true friendships and could acquire true values.
The most important companion in this journey has been my wife. Her unconditional love helped me through all the challenges that I have faced. Her support and encouragement gave me the confidence to finish what I have started and remained strong during times of doubt. She always trusted my decisions, listened to my ideas, and lend an ear to my presentations.
Finally, I am indebted to God, without whom none of this would be possible. His providence and guidance let me meet and work with many wonderful people, his goodness gave me some talent in which I tried to remain trustworthy. I offer him my work let it be as he wishes.

CEU eTD Collection

10.14754/CEU.2021.06

Contents

CEU eTD Collection

CEU eTD Collection

10.14754/CEU.2021.06

CEU eTD Collection

List of Tables

CEU eTD Collection

CEU eTD Collection

10.14754/CEU.2021.06

List of Figures

CEU eTD Collection

10.14754/CEU.2021.06

Chapter 1

Heterogeneous Treatment Effects in Regression Discontinuity Designs

1.1 Introduction

In regression discontinuity (RD) designs one identifies the *average* treatment effect from a jump in the regression function caused by the change in treatment assignment (or the probability of treatment assignment) as a running variable crosses a given threshold. Identification is based on comparing outcomes on the two sides of the cutoff, assuming that all other factors affecting the outcome change continuously with the running variable, which is not manipulable (see, e.g., Hahn et al., 2001, Imbens and Lemieux (2008), Lee and Lemieux (2010), Calonico et al., 2014). From 2005 regression discontinuity has become extremely popular in theoretical and empirical works, resulting in a large number of extensions.[1]

This paper contributes to the literature by proposing a machine learning algorithm designed to discover heterogeneity in the average treatment effect (ATE) estimated in an RD setup. The subpopulations that the algorithm searches over are defined by the values of a set of additional pre-treatment covariates. Analysis of treatment effect heterogeneity is important for at least two reasons. Firstly, researchers and policy makers gain a more detailed understanding of the treatment by learning the extent to which the treatment works differently in different groups. Indeed, the overall average effect may not be very informative if there is substantial heterogeneity. For example, the treatment may have no impact in one group while a large one in another, or there may

[1] E.g., Becker et al. (2013) defines heterogeneous local treatment effects in RD, where heterogeneity comes from a known covariate; Calonico et al. (2019) analyze the effect of using additional covariates; Xu (2017) extend the analysis with categorical variables as outcome; Cattaneo et al. (2016) concern multiple thresholds; Caetano et al. (2017) uses covariates to generate over-identifying restrictions in case of multiple treatment variable; Robson et al. (2019) proposes decomposition of ATE and CATE using covariate(s) with non-parametric methods; Toda et al. (2019) uses multiple groups with multiple threshold values to estimate CATE given by these pre-specified groups; Toda et al. (2019) uses machine learning to find discontinuity when there are many (potential) running variables and thresholds – but no heterogeneity in the treatment effect. Cattaneo et al. (2019) gives a great overview of recent developments in RD.

CEU eTD Collection

even be groups where the average treatment effect has opposite signs. Secondly, uncovering treatment effect heterogeneity – with strong external validity – can lead to a more efficient allocation of resources. If the budget for implementing a treatment is limited, decision makers can design future policies to focus on treating those groups where the expected treatment effects are the largest.

Of course, heterogeneity analysis is routinely undertaken in applied work, typically by repeating the main RD estimation within different groups defined by the researcher. Nevertheless, ad-hoc (or even pre-specified) selection of sub-samples has disadvantages: i) when there are many candidate groups defined by pre-treatment covariates, searching across these groups presents a multiple testing problem and without correction, it leads to invalid inference. ii) The relevant groups may have a complicated non-linear relationship with the treatment effect and discovering the non-linear pattern is cumbersome or impossible "by hand." For example, searching along the interactions of the pre-treatment covariates is usually infeasible and the researcher only checks few interactions motivated by theoretical considerations.[2]

By contrast, the method proposed in this paper allows discovering treatment effect heterogeneity based on pre-treatment covariates in a systematic way while offering a solution to the aforementioned challenges. At present, I know of no other paper that accomplishes these goals specifically in an RD setup. The closest paper with an RD focus is perhaps Hsu and Shen (2019), who develop *tests* for possible heterogeneity in the treatment effect based on the null hypothesis that the conditional average treatment effect (CATE) function is equal to a constant (the overall average treatment effect). Their proposed tests reveal whether there are groups defined in terms of observed characteristics for which the ATE deviates from the overall average, but they leave the discovery and the estimation of the conditional average treatment effect function as an open question. I address precisely this problem by proposing a data-driven machine learning method, which discovers groups with different treatment effects, using many candidate pre-treatment variables, without invalidating inference. The method provides discovery in the sense that the researcher does not need to specify the sources of heterogeneity (the relevant variables) in a pre-analysis plan, but can use many potentially relevant pre-treatment variables. The task of the algorithm is to find the relevant variables and the functional form from the many possible combinations. The end result gives groups with differences in the treatment effects. The implementation of the algorithm assumes that the standard RD identification conditions hold in the potentially relevant subpopulations; e.g., one cannot consider groups in which the running variable is always above or below the cutoff.

The paper also builds on and extends the more recent literature of discovering heterogeneous treatment effects with machine learning methods. There is a growing number of papers (e.g., Imai et al. (2013), Athey and Imbens (2016), Wager and Athey (2018),

[2]Hsu and Shen (2019) carry out a small survey of top publications in economics in 2005 that use the RD design. They find that 15 out of 17 papers check for heterogeneity and only 2 address the issue with interaction terms. The rest use subsample techniques without correcting for multiple testing.

CEU eTD Collection

Athey et al. (2019) Bargagli and Gnecco (2020), Friedberg et al. (2020), Knaus (2021) or Knaus et al. (2021)) using causal supervised machine learning (ML) techniques for this purpose.[3] All of these works are concerned with i) randomized experiments and/or ii) observational studies with the unconfoundedness assumption or iii) using instruments to estimate the local average treatment effect (LATE). Imai et al. (2013) use lasso with two sparsity constraints to identify heterogeneous treatment effects. The idea is to formulate heterogeneity as a variable selection problem in randomized experiments or observational studies with the unconfoundedness assumption. Athey and Imbens (2016) also focus on randomized experiments or observational studies with unconfoundedness, but use what they call *honest* regression trees to find heterogeneity in the treatment effect. The honest approach means that independent samples are used for growing the tree and estimating the average treatment effect in the resulting leaves. This ensures that traditional confidence intervals constructed for the estimates have the proper coverage rate. Bargagli and Gnecco (2020) follow the Athey and Imbens (2016) approach and extend it with instrumental variable setting to estimate conditional local average treatment effects (CLATE). Finally, the rest of the aforementioned papers and references therein go beyond regression trees[4] and use random forests or other machine learning methods to estimate conditional treatment effects in settings i), ii) and iii).[5] As these methods are already developed, one may argue that CATE function in RD can be estimated by using these causal supervised machine learning methods. However, these methods require some restrictive and unnecessary assumptions when identifying the ATE parameter. For example, in observational studies with unconfoundedness, it is impossible to construct a treatment-control contrast, without overlap. The lack of overlap causes the propensity score to take either the value of 1 or 0. This is indeed a problem as in many causal supervised machine learning methods this is typically excluded by assumption. E.g., for the propensity score weighted outcomes to make the transformation possible, one needs to assume that the propensity score values are away from the boundaries. Another caveat is that these methods

[3]There is another, distinct, strand of the broader causal inference literature where ML techniques are used for estimating high-dimensional nuisance parameters, while the parameter of interest is still the average treatment effect or a reduced dimensional version of CATE. See e.g., Chernozhukov et al. (2018), Semenova and Chernozhukov (2020) or Fan et al. (2020).

[4]Let me note here that in this paper I discuss building only one tree, which is known to be less stable in case of variables are (highly) correlated with each other. Extension to forest methods would generate a more robust estimator from this perspective. This extension is left for future research.

[5]Wager and Athey (2018), introduces causal (random) forests and shows that using honest trees to construct the forest, yields asymptotic normality for the conditional treatment effect estimator. They implement their theoretical results for causal forests in randomized experiments or observational studies with unconfoundedness. Friedberg et al. (2020) uses 'generalized random forests' as an adaptive weighting function to express heterogeneity. Friedberg et al. (2020) improves the asymptotic rates of convergence for generalized random forests with smooth signals by using local linear regressions, where the weights are given by the forests. Their method applies to randomized experiments and shows an application with observational study with unconfoundedness assumption. Knaus (2021) synthesize different methods using double machine learning with a focus on program evaluation under unconfoundedness assumption. He also proposes a normalized DR-learner to estimate individual average treatment effects. Knaus et al. (2021) provide a great overview about the Empirical Monte Carlo Study performances of the different machine learning methods, which are available and used in practice.

CEU eTD Collection

assume that the unobservable factors must be the same for treated and control units for every value of the running variable, thus the treated and control conditional expectation functions (CEFs) are on top of each other. In contrast, RD only assumes continuity for CEFs around the threshold, and CEFs can be anything away from the cutoff. An other approach is to use instrumental variables which relaxes the assumption of unobservable factors to be the same for both treated and control groups by using instruments. With introducing instrument(s) the core assumption is the exclusion restriction, thus instrumental variables enter only to the selection equation, but not the outcome equation, and are uncorrelated with the unobservables. Usually, it is hard to find such variable(s) in the context of RD. Furthermore, when the instrument is binary the ATE or CATE can be identified without further assumption. However, if this is not the case one needs to use the "identification at infinity" assumption. In contrast, with classical RD design, the researcher can avoid taking such strong assumption(s) while relying on the observed running variable and the continuity assumptions. For a more detailed discussion on how to estimate ATE with different types of models, see (Lee and Lemieux, 2010) Section 3.5.

Finally, let me mention two closely related papers, which work out a general framework for estimating heterogeneous treatment effects. (Athey et al., 2019) generalize the method of random forests and offer a method, based on local moment conditions to estimate the parameter of interest. In their paper, they work out local moment conditions for nonparametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables, but do not account for regression discontinuity designs. From their perspective, this paper is a special case of their general method, tailored for RD. The other paper is a working paper by (Nekipelov et al., 2019), which uses moment based models when constructing trees, called "moment forest". They use regression discontinuity design as an application for their method, where they make some strong assumptions on the functional form of conditional expectation functions, when estimating the CATE function, that I do not impose in this paper.

I contribute to the causal machine learning literature by introducing a specialized machine learning method to search for and estimate conditional average treatment effects in an RD setup. Following Athey and Imbens (2016) I capture heterogeneity by building an honest "regression discontinuity tree", where each leaf of the tree contains a parametric RD regression (to be estimated over an independent sample) rather than a simple difference between two means.[6] Similarly, an expected mean squared error criterion used to build the tree is modified appropriately to account for the more complicated statistic to be computed within each candidate leaf. Furthermore, the tree building algorithm also needs modifications to accommodate RD estimation and the new criterion. From a strictly technical standpoint, these are the main contributions of the paper. With the proposed algorithm, one can achieve unbiased estimates for the

[6]My future research agenda includes allowing for nonparametric RD estimation where the search for heterogeneity and the choice of the appropriate bandwidth is handled simultaneously.

CEU eTD Collection

group-level (conditional) average treatment effects and their variance.

I present Monte Carlo simulations to demonstrate that the algorithm successfully discovers and estimates heterogeneity in a variety of settings — at least with suitably large samples. In addition, I use the well-known and investigated dataset of Pop-Eleches and Urquiola (2013) on the Romanian school system. Pop-Eleches and Urquiola (2013) study the average treatment effect on Baccalaureate examination outcomes of going to a better school, and undertake some additional ad-hoc heterogeneity analysis. Hsu and Shen (2019) use their proposed test and show some evidence on the heterogeneity in the treatment effect without identifying the sources of the heterogeneity. I show that using the algorithm I can refine their results, discovering important treatment heterogeneity along with the level of school average transition scores[7] and number of schools in town. The algorithm reveals groups that have different treatment effects, but were missed by Pop-Eleches and Urquiola (2013). Furthermore, with a more extensive survey dataset with many socio-economic variables (but with fewer observations), I find that the estimated intention-to-treat effect varies among other covariates with having internet access at home, gender of the student, the education of the mother, and the proportion of novice teachers in school.

The paper is organized as follows. Section 1.2 introduces the concept of a sharp RD, a regression tree, and defines the conditional average treatment effect for the regression discontinuity tree. Section 1.3 develops the honest criterion for RD trees, which governs the discovery of the partitions. It also overviews the specifics of the algorithm for RD trees along with some practical guidance on bandwidth and order of polynomial selection. Section 1.4 shows the Monte Carlo simulation results with sharp regression discontinuity design for linear and nonlinear in running variable cases. Section 1.5 demonstrates the usefulness of the algorithm on datasets, collected by Pop-Eleches and Urquiola (2013). Section 1.6 extends the method to fuzzy RD designs. Section 1.7 concludes.

1.2 Regression Discontinuity Tree

With classical regression discontinuity design, researchers are interested in the causal effect of a binary treatment. Let $Y(1)$ denote the potential outcome, when a unit gets the treatment and $Y(0)$ if no treatment takes place. The observed outcome corresponding to the actual treatment status can be written as

$$Y = Y(D) = \begin{cases} Y(0), & \text{if} \quad D = 0, \\ Y(1), & \text{if} \quad D = 1\,. \end{cases}$$

[7] This is the average score within schools for incoming students. The transition score is calculated based on students performance on the national test(s) and by their previous grades during classes 5-8.

CEU eTD Collection

Treatment assignment in sharp RD[8] is a deterministic function of a scalar variable, called the *running variable*, which is denoted by X. This paper considers the standard case, in which the treatment D is determined solely by whether the value of the running variable is above or below a *fixed* and *known* threshold c :

$$D = \mathbb{1}_c(x) = \mathbb{1}_{[c,\infty)}(x) \begin{cases} 1, & \text{if} \quad x \geq c \\ 0, & \text{otherwise} \end{cases}$$

Treatment heterogeneity comes in the form of additional characteristics. Let Z be a set of K random variables referring to the possible sources of heterogeneity. Z are pre-treatment variables, therefore they must not have any effect on the value of the running variable. Following the machine learning terminology, call these variables *features*.
This paper proposes a method to estimate, or in some cases approximate, the conditional average treatment effect function given by

$$\tau(z) = \mathbb{E}\left[Y(1) - Y(0) | X = c, Z = z\right] \tag{1.1}$$

This function can be continuous, discrete or a mixture in Z. The proposed regression tree algorithm does not allow for such flexibility in each case, but gives a step-function approximation, when this CATE function is continuous in z. I will now introduce the basics of regression trees.

1.2.1 CATE in regression discontinuity tree

Regression trees – sometimes referred to as a partitioning scheme – allows one to construct a simple, intuitive and easy-to-interpret step-function approximation to the CATE. A tree Π corresponds to a partitioning of the feature space. Partitioning is carried out by recursive binary splitting: 1) Split the sample into two sub-samples along one feature with a split value. If a unit has a larger value for the selected feature than the split value, then it goes to the first sub-sample, otherwise to the second sub-sample. 2) If needed, one repeats the split, but now one considers the already split sub-samples for the next split. This way the feature space is partitioned into mutually exclusive rectangular regions. These final regions are called '*leaves*' or '*partitions*', denoted by ℓ_j. A regression tree, Π has $\#\Pi$ leaves, $j = 1, \ldots, \#\Pi$, whose union gives back the complete feature space $\mathbb{Z}$.

$$\Pi = \{\ell_1, \ldots, \ell_j, \ldots, \ell_{\#(\Pi)}\}, \quad \text{with} \quad \bigcup_{j=1}^{\#\Pi} \ell_j = \mathbb{Z}$$

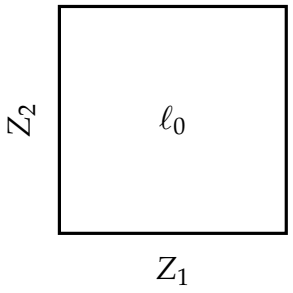
For illustrative purposes, consider only two features Z_1 and Z_2. Figure 1.1 shows three different trees with two representations.

[8] For fuzzy design, see Section 1.6

CEU eTD Collection

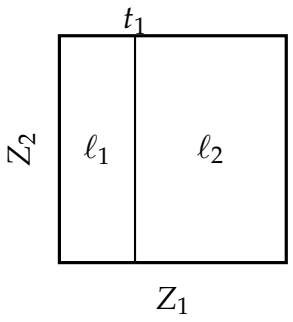
(1) **(2)**

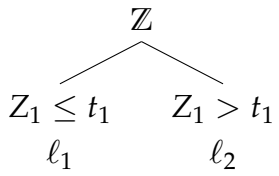
Partitioning **Tree structure**



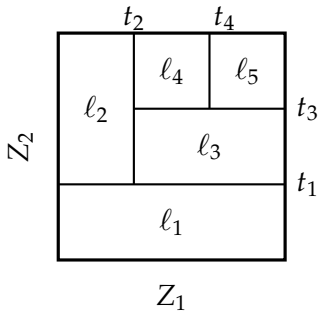
$\ell_0 = \mathbb{Z}$

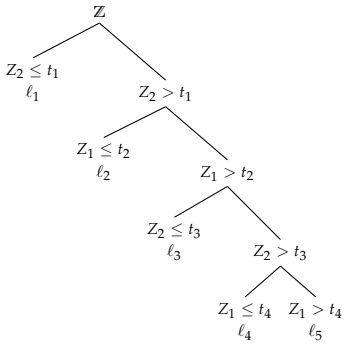
a) Π_0: homogeneous treatment effect, no partitioning





b) Π_1: two leaves, two treatment effects





c) Π_2: five leaves, five treatment effects

Figure 1.1: Different trees and their conditional average treatment effects

Column (1) shows the partitioning scheme: how the different partitions (or leaves) are split along the two features. Column (2) shows the tree structure: an intuitive interpretation using yes or no decisions, depending on the feature values and on the splitting values. Figure 1.1a) shows a tree, where there is only one leaf ℓ_0 containing all the units. This tree corresponds to a homogeneous treatment effect: no matter which value Z_1 or Z_2 takes, the treatment effect is always the same. In this case the conditional

CEU eTD Collection

average treatment effect is the same as the simple average treatment effect. Figure 1.1b) has two leaves: ℓ_1 and ℓ_2 resulting in two different treatment effects. Leaf ℓ_1 contains values with $Z_1 \leq t_1$ and ℓ_2 contains $Z_1 > t_1$, where t_1 is the splitting value. Note that Z_2 does not affect the partitioning and irrelevant with respect to treatment heterogeneity. Finally Figure 1.1c) shows a tree with five different leaves, resulting in five different treatment effects depending on both Z_1 and Z_2. In this case if one wants to find the treatment effect for a unit with $Z_1 = z_1$ and $Z_2 = z_2$, one needs to go through the decisions given by the tree. *Example*: $z_1 > t_3$ and $t_2 < z_2 \leq t_4$, corresponds to leaf ℓ_4. Note that the splitting values must satisfy $t_3 > t_1$, $t_1, t_3 \in Supp(Z_1)$ and $t_2, t_4 \in Supp(Z_2)$.

Recursive splitting provides rectangular regions for the different treatment effects, but never a continuous function. In case of a continuous CATE a simple tree offers only a step-function approximation. However, the tree structure ensures an intuitive decision-based interpretation for the treatment effects. Until Section 1.3, let us assume that the (true) tree Π is given. Using this known tree, the average treatment effect for leaf ℓ_j is defined as

$$\tau_j = \mathbb{E}\left[Y(1) - Y(0) | X = c, Z \in \ell_j(\Pi)\right] \tag{1.2}$$

To state the regression discontinuity tree approximation to the whole CATE function, let me introduce the indicator function for leaf ℓ_j.

$$\mathbb{1}_{\ell_j}(z;\Pi) = \begin{cases} 1\,, & \text{if} \quad z \in \ell_j(\Pi) \\ 0\,, & \text{otherwise} \end{cases}$$

The approximated conditional average treatment effect function provided by the regression discontinuity tree is given by

$$\tau(z;\Pi) = \sum_{j=1}^{\#\Pi} \tau_j \mathbb{1}_{\ell_j}(z;\Pi) \tag{1.3}$$

This CATE function – which incorporates the tree structure – links the treatment effects for each individual leaf. As the leaves represent rectangular partitions, this function is a step-function approximation to the continuous CATE function. By the law of iterated expectation, this approximation has the property of $\mathbb{E}\left[\tau(Z) \mid \mathbb{1}_{\{c\}}(X), \mathbb{1}_{\ell_1}(Z), \ldots, \mathbb{1}_{\ell_{\#\Pi}}(Z)\right] = \tau(Z;\Pi)$. This means that at the threshold value $(X = c)$ with the given tree structure, the expected value of the continuous CATE function over the leaves, is equal to the step-approximated CATE.

1.2.2 Identification of CATE in the sharp RD

To identify the conditional average treatment effect function for trees in sharp RD, the following assumptions are needed:

CEU eTD Collection

Identification assumptions

i) $\mathbb{E}\left[Y(1)|X=x, Z \in \ell_j(\Pi)\right]$ and $\mathbb{E}\left[Y(0)|X=x, Z \in \ell_j(\Pi)\right]$, exists and continuous at $x = c$ for all leaves in the tree.

ii) Let $f_j(x)$ denote the conditional density of x in leaf j. In each leaf j, c is an interior point of the support of $f_j(x)$.

Assumption i) states that the expected value of the potential outcomes conditional on the running variable in each leaf exists and continuous. It is required to identify the average treatment effects for all leaves. This assumption is similar to the classical RD assumption (see e.g., Imbens and Lemieux (2008)), but somewhat stronger, due to extension to the tree.[9] Assumption ii) ensures that the density for the running variable is well behaved: it has positive probability below or above the threshold value within each leaf. This excludes cases when there are no values of the running variable on both sides of the threshold in a given leaf. Finally, in the RD literature it is common to require the continuity of the conditional distribution functions – in this case it extends to $f_j(x)$ to be continuous in x [10] – which is an implication of *"no precise control over the running variable"* (see e.g., Lee and Lemieux (2010)). In case, when local randomization around the threshold holds, the algorithm does not need this assumption. [11]

[9] But less restrictive if one assumes continuity in $Z = z$ as in e.g., Hsu and Shen (2019).

[10] One need to use the Bayes' Rule to show this, along with assumption i)

[11] *Note:* Although the used conditional average treatment effect function here is a step-function approximation, it can be a building block of a causal forest for sharp RD, which produces continuous condition average treatment effect. In this case one needs further modification on the assumption for the conditional expectation and densities. Causal forests for RD is out of the scope of this current paper.

CEU eTD Collection

If these assumptions hold, the (step-function approximated) conditional average treatment effect given by a regression discontinuity tree is identified as

$$\begin{aligned}\tau(z;\Pi) &= \sum_{j=1}^{\#\Pi} \tau_j \mathbb{1}_{\ell_j}(z;\Pi) \\ &= \sum_{j=1}^{\#\Pi} \left\{ \mathbb{E}\left[Y(1)|X=c, Z\in\ell_j(\Pi)\right] - \mathbb{E}\left[Y(0)|X=c, Z\in\ell_j(\Pi)\right] \right\} \mathbb{1}_{\ell_j}(z;\Pi) \\ &= \sum_{j=1}^{\#\Pi} \left\{ \lim_{x\downarrow c} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] - \lim_{x\uparrow c} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] \right\} \mathbb{1}_{\ell_j}(z;\Pi) \\ &= \mu_+(c,z;\Pi) - \mu_-(c,z;\Pi)\end{aligned} \tag{1.4}$$

where

$$\begin{aligned}\mu_+(x,z;\Pi) &= \sum_{j=1}^{\#\Pi} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] \mathbb{1}_{\ell_j}(z;\Pi) \\ \mu_-(x,z;\Pi) &= \sum_{j=1}^{\#\Pi} \mathbb{E}\left[Y(0)|X=x, Z\in\ell_j(\Pi)\right] \mathbb{1}_{\ell_j}(z;\Pi)\end{aligned} \tag{1.5}$$

refers to the conditional expectation function for (μ_+) above the threshold (treated) and (μ_-) below the threshold (untreated) units. That is, each τ_j is identified within its leaf in the usual way.

1.2.3 Interpretation of the estimand conditional on (un)observables

The conditional average treatment effect estimand in regression discontinuity designs is not as straightforward as in experimental designs or observational studies with the unconfoundedness assumption. To interpret the estimand, let me formalize the individual treatment effect as in (Lee and Lemieux, 2010),

$$Y(1) = Y(0) + \tau(Z, U)$$

where Z are the known observed covariates and U is unobserved heterogeneity in the individual treatment effect. In classical sharp RD setup[12], $\tau(Z,U)$ does not depend directly on the running variable X. Here, I will consider this simple case. Note that X, Z and U can be correlated in this setup, thus individuals with characteristics of Z and U can have typical X values, but X does not directly influence the magnitude of the treatment effect.

Naturally, individual treatment effects can not be observed as one can not assign the same unit to be treated and non-treated at the same time. Instead, one can identify a type of conditional average treatment effect , where Z and X are fixed and U is aver-

[12]Simple assignment rule: $X \geq c$, the individual gets the treatment, otherwise not.

CEU eTD Collection

aged out:

$$\tau(z) = \mathbb{E}\left[\tau(Z,U) \mid X = c, Z = z\right] = \mathbb{E}\left[Y(1) - Y(0) \mid X = c, Z = z\right] .$$

For purposes of interpretation, I consider the case when Z and U are discrete, but a similar argument applies to the continuous case. First, focus on the general case when no tree structure is used. CATE is identified through

$$\tau(z) = \lim_{x \downarrow c} \mathbb{E}\left[Y \mid X = x, Z = z\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y \mid X = x, Z = z\right] .$$

For the identifying equality to hold, the following extension of standard continuity conditions must hold:

i) $\mathbb{E}\left[Y(1)|X = x, Z = z\right]$ and $\mathbb{E}\left[Y(0)|X = x, Z = z\right]$, exists and continuous at $x = c$.

ii) Let $f(x \mid Z = z)$ denote the conditional density of x given $Z = z$. For each value of $z \in \text{Supp}(Z)$, c is an interior point of the support $f(x \mid Z = z)$.

Under these conditions, the CATE function is equal to,

$$\begin{aligned}
\mathbb{E}\left[Y(1) - Y(0) \mid X = c, Z = z\right] &= \mathbb{E}\left[\tau(Z,U) \mid X = c, Z = z\right] \\
&= \sum_u \tau(z,u) \mathbb{P}\left[U = u \mid X = c, Z = z\right] \\
&= \sum_u \tau(z,u) \frac{f(c \mid U = u, Z = z)}{f(c \mid Z = z)} \mathbb{P}\left[U = u \mid Z = z\right]
\end{aligned}$$

where $\mathbb{P}\left[\cdot \mid \cdot\right]$ denotes conditional probability and $f(\cdot \mid \cdot)$ denotes conditional density function. This formula is the exact analog of equation (5) in (Lee and Lemieux, 2010).

Thus, the CATE function is a particular kind of average treatment effect across individuals with covariate values $Z = z$. If the term $f(c|U = u, Z = z)/f(c \mid Z = z)$ were equal to 1, it would be the treatment effect for individuals with observed $Z = z$ averaged over the unobserved $U = u$ values. This is the case if the unobserved heterogeneity U is independent of the running variable X conditional on the covariates Z. More generally, the presence of the ratio $f(c|U = u, Z = z)/f(c \mid Z = z)$ implies the regression discontinuity estimand is instead a weighted average treatment effect. Within the subgroup $Z = z$, the weight is larger for individuals whose X value is ex-ante more likely to be close to the threshold c based on their unobserved characteristics. The weights may be relatively similar across individuals, in which case the individual treatment effects would be closer to the CATE but, if the weights are highly varied and also related to the magnitude of the treatment effect, then the individualized treatment effects would be very different from the CATE. However, the weights across individuals is ultimately unknown, since we do not observe U. Thus, it is not possible to know how close the individualized treatment effects are to the CATE and it remains the case that the treatment effect estimated using an RD design is averaged over a larger population than one would have anticipated from a purely "cutoff" interpretation.

CEU eTD Collection

Finally, let me discuss the impact of using a regression tree representation in the interpretation of the CATE function. Following equation (1.2) from the paper, the leaf-by-leaf treatment effect can be similarly decomposed as

$$\begin{aligned}\tau_j &= \mathbb{E}\left[Y(1) - Y(0) \mid X = c, Z \in \ell_j\right] \\ &= \mathbb{E}\left[\tau(Z,U) \mid X = c, Z \in \ell_j\right] \\ &= \sum_{z,u} \tau(z,u)\mathbb{P}\left[Z = z, U = z \mid X = c, Z \in \ell_j\right] \\ &= \sum_{z \in \ell_j, u} \tau(z,u)\frac{f(c \mid Z = z, U = u, Z \in \ell_j)}{f(c \mid Z \in \ell_j)}\mathbb{P}\left[Z = z, U = u \mid Z \in \ell_j\right] .\end{aligned}$$

The interpretation remains similar, but with tree structure one needs to average over not only the unobserved characteristics $(U = u)$, but over the observed characteristics within each leaf j as well.

Remarks:

i) If there is no unobserved heterogeneity in the treatment effect $(\tau(Z,U) = \tau(Z))$ then in the continuous case one can estimate the individualized treatment effects. With tree structure, weights are still present as the conditional densities are not necessarily same within leaf j for each values of z.

ii) In case the tree specification is correct in the sense that $\mathbb{E}\left[\tau(Z,U) \mid X = c, Z = z\right] = \mathbb{E}\left[\tau(Z,U) \mid X = c, Z \in \ell_j\right]$, then the interpretation is the same as if $\tau(Z,U)$ would be continuous in Z.

iii) If the tree is correctly specified and there is no unobserved heterogeneity in the treatment effect then the CATE via tree structure is the same as the individualized treatment effect.

1.2.4 Parametrization and estimation

The paper assumes q-th order polynomial functional form in X for each leaf to identify τ_j. Each conditional expectation function – $\mathbb{E}\left[Y(d)|X = x, Z \in \ell_j(\Pi)\right], d \in \{0,1\}$ – is given by a q-th order polynomial, which ensures a flexible functional form. [13] To formalize the parametrization of the conditional expectation function given by equation (1.5) first adjust X by c, and let $\boldsymbol{X}$ be the $(q+1) \times 1$ vector

$$\boldsymbol{X} = \left[1, (X-c), (X-c)^2, \ldots, (X-c)^q\right]'$$

[13] Nonparametric estimations such as local polynomial regression is not considered in this paper – mainly because of optimal criterion for growing a tree is more cumbersome in the presence of potentially multiple bandwidth – however in case of strong non-linearity in X, I recommend to use a restricted sample using a bandwidth (e.g., proposed by Imbens and Kalyanaraman (2012)), which is estimated on the whole sample.

CEU eTD Collection

For a given Π, one can then write

$$\mu_+(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\delta_j^+ , \qquad \mu_-(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\delta_j^-$$

where, $\delta_j^+ = \left[\alpha_j^+, \beta_1^+, \beta_2^+, \ldots, \beta_q^+\right]'$ and $\delta_j^- = \left[\alpha_j^-, \beta_1^-, \beta_2^-, \ldots, \beta_q^-\right]'$ are a $(q+1)\times 1$ parameter vectors[14] and depends on the partitioning. Note that this definition allows for each leaf (thus group) to have different functional forms in X.

To estimate $\tau(z;\Pi)$ consider a sample $\mathcal{S}$, consisting of independent and identically distributed observations $(Y_i, X_i, Z_i)\,; i = 1,\ldots,N$. The paper employs leaf-by-leaf estimation for the parameter vectors δ_j^+ and δ_j^-, using least squares.[15] The estimator for the parameters are given by

$$\hat{\delta}_j^+ = \arg\min_{\delta_j^+} \sum_{i\in\mathcal{S}} \left\{ \mathbb{1}_c(X_i)\mathbb{1}_{\ell_j}(Z_i;\Pi) \left(Y_i - \boldsymbol{X}_i'\delta_j^+\right)^2 \right\}$$

$$\hat{\delta}_j^- = \arg\min_{\delta_j^-} \sum_{i\in\mathcal{S}} \left\{ \left[1 - \mathbb{1}_c(X_i)\right] \mathbb{1}_{\ell_j}(Z_i;\Pi) \left(Y_i - \boldsymbol{X}_i'\delta_j^-\right)^2 \right\} , \qquad \forall j$$

Using these parameter vectors and the identification equation for CATE (equation 1.4), the least squares estimator for conditional average treatment effect for regression discontinuity tree is given by,

$$\hat{\tau}(z;\Pi,\mathcal{S}) = \hat{\mu}_+(c,z;\Pi,\mathcal{S}) - \hat{\mu}_-(c,z;\Pi,\mathcal{S}) = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\left(\hat{\alpha}_{+,j} - \hat{\alpha}_{-,j}\right)$$

Remark: The sample $\mathcal{S}$ is highlighted in this notation, due to later purposes to differentiate between estimates using different samples. Subscript i always refers to observations from sample $\mathcal{S}$, j index represents leaf j from tree Π and subscripts $+/-$ stands for above or below the threshold.

1.3 Discovering regression discontinuity tree

In this section, the assumption of a known tree is gradually relaxed. I approach this problem in three steps. Firstly, I introduce different distinct samples which are necessary to obtain an unbiased estimator of the CATE function, when using the regression tree algorithm. Here, I sketch some properties of the algorithm, which is detailed in

[14] For RD the main parameter of interest is $\alpha_j^{\pm}$. $\beta^{\pm}$ should also be $\beta_1, j^{\pm}$, but I neglect j subscript for convenience.

[15] This method has the advantage of relative fast estimation. Computationally it is much more compelling than the two other alternatives: 1) joint estimation of the whole tree and 2) also include in one regression the treated and non-treated units. Although, these methods use milder assumptions during the search for proper tree, but when estimating with these setups there is a need for inverting large sparse matrices (interactions of $\mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{X}'$), which can lead to computationally expensive methods and non-precise estimates.

CEU eTD Collection

the last step. Secondly, I analyse the criterion, which compares different trees. At this stage I assume that these different trees are exogeneously given. Finally, I show how the optimal tree is found by the regression tree algorithm, using the different samples and the proposed criterion.

1.3.1 Distinction of samples

An inherent problem of using only one sample for finding relevant sub-groups and estimating treatment effects is that it results in incorrect inference if there is no adjustment for multiple testing. (see, e.g., Romano and Shaikh, 2010 or Anderson, 2008)

Although regression tree algorithm controls for over-fitting in some way – as I will discuss in Section 1.3.3 – the estimate is biased in finite samples and disappears only slowly as the sample size grows. Athey and Imbens (2016) proposes '*honest regression tree*' approach to eliminate the bias from the estimated conditional average treatment effects in experimental settings or observational studies with the unconfoundedness assumption. By their definition, a regression tree is called 'honest' if it does not use the same information for growing the candidate trees as for estimating the parameters of that tree. This requires using two *independent* samples. The '*test sample*' ($\mathcal{S}^{te}$) is used for evaluating the candidate trees and the '*estimation sample*' ($\mathcal{S}^{est}$) for estimating the treatment effects. These samples are also used to derive and analyse the honest criterion for the regression discontinuity tree. In Section 1.3.3 I elaborate further on how the samples are used when growing a tree. Honesty has the implication that the asymptotic properties of treatment effect estimates within the partitions are the same as if the partition had been exogeneously given, thus biases are eliminated and one can conduct inference in the usual way. The cost of the honest approach is the loss in precision – less observation used – due to sample splitting (Athey and Imbens, 2016, p. 7353-7354).[16]

1.3.2 Criterion for RD tree

A natural – but in-feasible criterion – for evaluating the regression discontinuity tree would be minimizing the mean squared error of the estimated CATE on the test sample. Let a partition (Π) be exogeneously given. The CATE function ($\hat{\tau}(Z_i; \Pi, \mathcal{S}^{est})$) is estimated on $\mathcal{S}^{est}$ and evaluated on $\mathcal{S}^{te}$. The in-feasible MSE criterion is

$$MSE_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = \frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \left\{ \left[\tau(Z_i) - \hat{\tau}(Z_i; \Pi, \mathcal{S}^{est}) \right]^2 - \tau^2(Z_i) \right\} \tag{1.6}$$

where N^{te} is the number of observations in the test sample. Note, in this formulation, there is an extra adjustment term, $\tau^2(Z_i)$ – a scalar, independent of Π. Thus, it does not have any effect on the results, but it facilitates theoretical derivations. Furthermore, let

[16] With the honest approach one does not need to place any external restrictions on how the tree is constructed. In the literature, there are other papers, which use additional assumptions to get valid inference, which is also a possible - but in my opinion a more restrictive approach. An example is Imai et al. (2013), who use 'sparsity' condition: only few features affect the outcomes.

CEU eTD Collection

me emphasize that this in-feasible criterion utilizes both the estimation sample and the test sample in a way that observations are needed to be known for both samples.

Calculating this criterion for different exogeneously given trees, would allow one to find the tree, whose deviation from the true CATE function is the smallest in the test sample. The problem is that, $\tau(\cdot)$ is unknown, thus this criterion is in-feasible. Instead – following Athey and Imbens (2016) – I minimize the *expected* MSE over the test and estimation samples. This formulation has two advantages: i) it gives the best fitting tree for the *expected* test and estimation sample. This is favourable, because when the tree is grown, both of these samples are locked away from the algorithm (see Section 1.3.3). ii) using this formulation, an estimable criterion can be derived for comparing trees in practice. The expected MSE criterion is given by

$$EMSE_\tau(\Pi) = \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}}\left[MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi)\right] \tag{1.7}$$

This paper advocates trees (Π), which gives the smallest $EMSE_\tau$ value from all the candidate trees. Based on Athey and Imbens (2016), this EMSE criterion can be decomposed into two terms,[17] which helps to evaluate why this criterion offers a good choice for selecting a tree.

$$EMSE_\tau(\Pi) = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] \tag{1.8}$$

This formulation highlights the trade-off between finding new different treatment effects – hence larger trees – and minimizing the variance of the estimated treatment effects. The expected value[18] of the squared CATE $(\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right])$ prefers trees which are larger, as the expected squared treatment effects grows as there are more leaves (or groups). On the other hand any estimator for this term is increasing in the number of splits, which leads to select trees, that are too large, i.e. where the treatment effects are in fact the same in different leaves. This is called over-fitting the true tree. The first term, the expected value of the treatment effect variances, explicitly incorporates the fact that finer partitions generate greater variance in leaf estimates in finite samples. Therefore it prefers smaller trees, where the average variance of the estimated treatment effects are lower. Through this channel, this term offsets the over-fitting caused by the expected value of the squared treatments. Note that, the expected variance term may select larger trees if leaves (or groups) have the same treatment effect, but have lower expected variances.

A technical contribution of this paper is to provide estimators for the expected treatment variances and the expected squared treatment effects in the regression disconti-

[17] See the detailed derivations in Appendix A.1. To derive estimable EMSE criterion, the assumption of $\mathcal{S}^{est}$ and $\mathcal{S}^{te}$ being independent from each other is key.

[18] Z_i refers to features from $\mathcal{S}^{te}$.

CEU eTD Collection

nuity setup. Here I only present the results, refer to Appendix A.2 for the derivations.[19] In order to analyse the proposed estimator, let me first introduce the following expressions. Write the model as

$$Y_i = \mathbb{1}_c(X_i)\mu_+(X_i, Z_i; \Pi) + (1 - \mathbb{1}_c(X_i))\mu_-(X_i, Z_i; \Pi) + \epsilon_i$$

where ϵ_i is the idiosyncratic disturbance term. Furthermore, let

$$\hat{\sigma}^2_{+,j} = \frac{1}{N^{te}_{+,j} - q - 1} \sum_{i \in \mathcal{S}^{te}} \left[\mathbb{1}_c(X_i) \mathbb{1}_{\ell_j}(Z_i; \Pi) \hat{\epsilon}_i \right]^2 ,$$
$$\hat{\sigma}^2_{-,j} = \frac{1}{N^{te}_{-,j} - q - 1} \sum_{i \in \mathcal{S}^{te}} \left[\{1 - \mathbb{1}_c(X_i)\} \mathbb{1}_{\ell_j}(Z_i; \Pi) \hat{\epsilon}_i \right]^2$$

be the within leaf variance estimators for the disturbance terms in leaf j with $N^{te}_{+,j}, N^{te}_{-,j}$ number of observations within the same leaf for above and below the threshold respectively. $\hat{\epsilon}_i$ are the OLS residuals. For simplicity, I assume same finite variance within the leaves, when deriving these estimators.[20] (See Appendix A.4 for extensions, which relax the finite variance assumption.) Furthermore, let the cross-product of the running variable above and below the threshold for leaf j be

$$M_{+,j} = \frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \left(\boldsymbol{X}_i \boldsymbol{X}_i \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i) \right) ,$$
$$M_{-,j} = \frac{1}{N^{te}_{-,j}} \sum_{i \in \mathcal{S}^{te}} \left(\boldsymbol{X}_i \boldsymbol{X}_i \mathbb{1}_{\ell_j}(Z_i; \Pi) (1 - \mathbb{1}_c(X_i)) \right) .$$

Using these quantities, one can derive specifically scaled variance estimators for the parameter vectors in leaf j:

$$\mathbb{V}\left[\hat{\delta}^+_j\right] = \frac{\hat{\sigma}^2_{+,j} \hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} , \qquad \mathbb{V}\left[\hat{\delta}^-_j\right] = \frac{\hat{\sigma}^2_{-,j} \hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}$$

where $p^{est}_{+,j}$ and $p^{est}_{-,j}$ are the share of units above and below the threshold in the estimation sample within leaf j. (Specific scaling is explained in Remarks ii-iii), see below.)

Estimator for the expected variance of the treatment effects can be derived as an

[19] For the derivations I have used two further simplifying assumptions: i) the share of observations within each leaf – number of observations within the leaf compared to the number of observations in the sample – are the same for the estimation and test sample. ii) the share of units below and above the threshold within each leaf are the same for the estimation and test sample. Asymptotically both assumptions are true.

[20] Also called homoscedastic errors within each leaf – which refers to the variances of the errors *within* the leaves being the same. Note: that this only assumed for within leaves and not for the whole partition, thus disturbance terms for all leaves (ϵ_i) does not need to be homoscedastic.

CEU eTD Collection

average of these variance estimators,

$$\hat{\mathbb{E}}_{Z_i}\left\{\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} = \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\} \tag{1.9}$$

where $e_1 = [1, 0 \ldots, 0]$ is a $1 \times (q+1)$ selector-vector to choose the variances of the intercepts referring to the treatment effect.[21] N^{est} is the number of observations in the estimation sample, which is a result from the derivations (see Appendix A.2.1).

Remarks:

i) Although the variance of the treatment effects refers to the estimation sample, $\hat{\sigma}^2_{\pm,j}, M^{-1}_{\pm,j}, \forall j$ are calculated using only observations from the test sample. This is possible, as the estimation and the test samples are independent from each other, therefore the asymptotic estimators for these quantities are the same.

ii) To adjust the variance estimator in finite samples for the estimation sample, one only needs to use limited information from the estimation sample, namely the share of observations above and below the threshold $(p^{est}_{+,j}, p^{est}_{-,j})$.

iii) Using the leaf shares instead of the number of observations for above and below the threshold is possible, as the variance of the treatment effect estimators are the same for each observation within the leaf, therefore one can use summation over the leaves $(j = 1, \ldots, \#\Pi)$ instead of individual observations.

The estimator for the expected value of the squared true CATE (second part of equation 1.8), uses the squared of estimated CATE and corrects the resulting bias with the variance. The estimator uses only the test sample, apart from weights in the variance estimator.[22]

$$\hat{\mathbb{E}}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{te}) - \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\} \tag{1.10}$$

The averaged squared treatment estimator prefers trees with many leaves. It is the sample analogue for finding groups with different treatment effects. This term always increases as the number of leaves increases, while the average of the sum of squared treatment effects for two (or more) groups is always greater than the average of the sum of one averaged squared treatment effect. The second part is similar to the derived expected variance, but here the scaling for the average (N^{te}) comes from the test sample, as the estimator refers to the expected value over the test sample.[23] The

[21] In case of kink designs, the selector vector would choose the appropriate order of polynomial.

[22] See the derivations in Appendix A.2.2.

[23] An alternative estimator would be using the estimation sample only. However, my goal is to construct an EMSE estimator, which uses only the test sample's observation and only some additional information from the estimation sample, to ensure that during the tree building phase the estimation sample is locked away to get valid inference.

CEU eTD Collection

weights of $p^{est}_{+,j}$ and $p^{est}_{-,j}$ comes from the estimation sample and they help the algorithm to avoid sample specific splits – see the discussion in Section 1.3.3.
Putting together the two estimators one gets the following estimable EMSE criterion for regression discontinuity trees:

$$\widehat{EMSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = -\frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}^2(Z_i; \Pi, \mathcal{S}^{te}) + \left(\frac{1}{N^{te}} + \frac{1}{N^{est}}\right) \sum_{j=1}^{\#\Pi} \left\{ e_1' \left[\frac{\left(\hat{\sigma}^2_{+,j} \hat{M}^{-1}_{+,j}\right)}{p^{est}_{+,j}} + \frac{\left(\hat{\sigma}^2_{-,j} \hat{M}^{-1}_{-,j}\right)}{p^{est}_{-,j}} \right] e_1 \right\} \tag{1.11}$$

Minimizing this criterion leads to trees, where i) there is a strong evidence for heterogeneity in the treatment effects for different groups; and ii) penalize a partition that creates variance in leaf estimates. Furthermore, this criterion encourages partitions, where the variance of a treatment effect estimator is lower, even if the leaves have the same average treatment effect, thus finds features, which affect the mean outcome, but not the treatment effects itself.
Finally, let me compare the estimator for EMSE criterion and the initial in-feasible MSE criterion. As the in-feasible MSE criterion uses the estimation sample to get an estimator for the CATE function and then evaluates it on the test sample, the estimator for EMSE criterion uses the observations from the test sample and only scales it with the number of observations (N^{est}) and share of units below and above the threshold for each leaf $(p^{est}_{\pm,j})$ from the estimation sample. This means there is only a limited information needed from the estimation sample to calculate the EMSE criterion, but not individual observations. This property enables that the observation values from the estimation sample are locked away for the algorithm, when searching for an optimal tree.

1.3.3 Finding EMSE optimal RD tree

Unit now, I have compared different, already given partitions using the proposed criterion. In this sub-section, I introduce the basic notations and steps to grow the EMSE optimal regression discontinuity tree, following the literature on classification and regression trees (CART) and honest causal regression trees. For more detailed description see, Breiman et al. (1984), Ripley (1996) or Hastie et al. (2011) on CART algorithms and Athey and Imbens (2015, 2016) on honest causal tree algorithm.

Finding the EMSE optimal honest RD tree has four distinct stages:

1. Split the sample into two independent parts.
2. Grow a large tree on the first sample.
3. Prune this large tree to control for over-fitting. This is carried out by cross-validation and it results in an EMSE optimal tree.

CEU eTD Collection

4. Use this EMSE optimal tree to estimate the CATE function on the independent estimation sample.

In the first stage *'honest'* approach randomly assigns the initial sample into two samples to achieve an unbiased CATE estimator. The first sample is called the *'training sample'* ($\mathcal{S}^{tr}$) and its observations are used to grow trees. The second, *'estimation sample'* has a special role. In general, it is locked away from the algorithm, but information on the number of observations is utilized during the tree building phase to control for finding training sample specific patterns. Observation values from the estimation sample are not used until the last stage. This division ensures valid inference for the CATE function in the fourth step. Figure 1.2 shows these two samples, which are used to grow a large tree and providing valid inference.

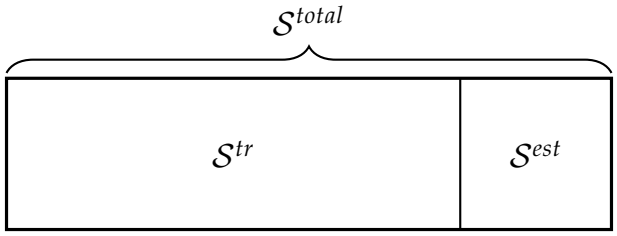


Figure 1.2: At stage 2, $\mathcal{S}^{tr}$ is used to grow large tree $\left(\hat{\Pi}^{large}\right)$

In the second stage, a large tree is grown using all the observations from the *training sample*. The algorithm recursively partitions the training sample along with the features. For each leaf the method evaluates all candidate features and their possible splits inducing alternative partitions, with the *'honest in-sample criterion'*: $\widehat{EMSE}_{\tau}(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi)$. This criterion uses additional information from the estimation sample. The treatment effects and the variances are estimated on the training sample only, but they are adjusted with the number of observations (N^{est}) and share of treated and non-treated units within each leaf from the estimation sample ($p^{est}_{\pm,j}$). N^{est} adjusts for the sample shares (how the initial sample is divided into two parts). This does not have a large impact on the in-sample criterion as the value is given by the first step and does not change during the partitioning. Using $p^{est}_{\pm,j}$ instead of $p^{tr}_{\pm,j}$ has larger implication in finite samples. It prevents the algorithm to choose such feature and splitting value, which is only specific to the training sample. After the split is done, the algorithm iterates the procedure on the newly created leaves. The process repeats itself and stops if the in-sample-criterion does not decrease any further or the magnitude of the reduction is smaller than a pre-set parameter.[24] With this method, one gets a large tree $\left(\hat{\Pi}^{large}\right)$.

[24]The algorithm accepts splits, where the in-sample-criterion decreases compare to a tree without the split. It is possible to specify a minimum amount of reduction in the in-sample-criterion, which by default is set to zero. Furthermore, the algorithm considers a split valid if the number of observations within the leaves is more than a pre-set value (typically 50 observations) for both the treated and control group. Finally, there are additional (optional) stopping rules implemented such as the maximum depth of the tree, the maximum number of leaves, the maximum number of nodes, or the maximum number of iteration.

CEU eTD Collection

The resulting large tree is prone to over-fitting as $\widehat{EMSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi)$ is not unbiased when one uses it repeatedly to evaluate splits on the training data. The bias comes from the fact that after the training sample has been divided once, the sample variance of observations in the training data within a given leaf is on average lower than the sample variance would be in a new, independent sample. This leads to finding features relevant, which are in fact irrelevant to the true CATE function. Thus using only $\widehat{EMSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi)$ is likely to overstate the goodness of fit as one grows deeper and deeper tree.

To solve for the over-fitting – in the third stage – cross-validation is used. The idea is to split the training sample into two further parts: a sample where the tree is independently grown $\mathcal{S}^{(tr,tr)}$ and to a test sample $\mathcal{S}^{(tr,te)}$ where the EMSE criterion can be safely evaluated. This ensures that the tree grown on $\mathcal{S}^{(tr,tr)}$ is exogenous for $\mathcal{S}^{(tr,te)}$, thus the estimated EMSE criterion is unbiased. [25] Figure 1.3, shows the splitting of the original training sample into $\mathcal{S}^{(tr,tr)}$ and $\mathcal{S}^{(tr,te)}$.

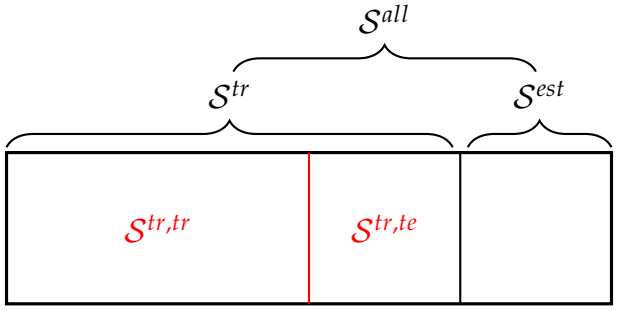


Figure 1.3: At stage 3, $\mathcal{S}^{tr,te}$ is used to evaluate the tree $\hat{\Pi}$ grown on $\mathcal{S}^{tr,tr}$

Note that, one needs to split the estimation sample as well for the accompanied information on the shares of treated and non-treated units to evaluate the EMSE criterion.

The EMSE optimal tree is found via cost-complexity pruning, which utilizes a complexity parameter (γ). The complexity parameter penalizes the number of leaves $(\#\Pi)$ grown on the tree. The *'honest cross-validation criterion'* adds this penalty term to the original EMSE criterion,

$$\widehat{EMSE}_{cv}(\gamma) = \widehat{EMSE}_\tau(\mathcal{S}^{(tr,val)}, \mathcal{S}^{(est,val)}, \hat{\Pi}) + \gamma \#\hat{\Pi} \tag{1.12}$$

where, $\hat{\Pi}$ is an estimator of the tree, grown on the samples of $\{\mathcal{S}^{(tr,tr)}, \mathcal{S}^{(est,tr)}\}$ and the EMSE criterion is evaluated on the independent sample pair of $\{\mathcal{S}^{(tr,te)}, \mathcal{S}^{(est,te)}\}$. To find the optimal complexity parameter (γ^*) – hence the EMSE optimal tree – one calculates the honest cross-validation criterion R times on the alternating test samples,

[25] The size of the samples are given by the number of folds (R) used in the cross-validation. $\mathcal{S}^{(tr,te)}$ has the smaller fraction: $N^{(tr,te)} = N^{tr}/R$, while the sample $\mathcal{S}^{(tr,tr)}$, which is used to grow the tree, contains the larger fraction of observations $N^{(tr,tr)} = (R-1)N^{tr}/R$. The estimation sample is split in the same way.

CEU eTD Collection

which results in R criteria for each different candidate of γ.[26] Taking the average over the cross-validation samples one can choose γ, which gives the minimum criterion value.[27]

$$\gamma^* = \arg\min_{\gamma} R^{-1} \sum_{cv=1}^{R} \widehat{EMSE}_{cv}(\gamma) \tag{1.13}$$

The final step of the third stage is to prune back the original large tree $\left(\hat{\Pi}^{large}\right)$ grown on the whole training sample with γ^* to get the optimal tree $\widehat{\Pi}^*$.
In the fourth stage, one uses the locked away estimation sample and the found tree structure $\widehat{\Pi}^*$ to estimate the CATE function for the regression discontinuity tree.

1.3.4 Refining honest tree algorithm for RD

In this subsection, I discuss the refinements of honest tree algorithms, which are needed for the estimation of the CATE function in regression discontinuity designs.
The main challenge for RD algorithm takes place during the tree-building phase to find the optimal splitting values for each candidate feature. As the algorithm employs many regressions when considering each possible splits, the inversion of $M_{\pm,j}$ is computationally challenging. Instead of calculating the inverse each time, I use the Sherman-Morison formula to estimate $M_{\pm,j}^{-1}$. This iterative estimation enables to calculate the inverse only once per splitting candidate feature.[28]

Another important detail of the honest algorithm is 'bucketing'. Following Athey and Imbens (2016), bucketing ensures that each candidate split has enough treated and non-treated units, thus there is no 'better' split value only due to adding treated or non-treated units, without the other. One should see bucketing as a smoother of the splitting criterion, as it groups the treated and non-treated units and prevents the splitting value to be a result of this unbalanced grouping of treated and non-treated units. I refine the classical causal tree algorithm[29], by carrying out the bucketing after the criterion is calculated and using the last valid split value instead of taking the average. This is an important nuisance as the criterion may vary too much without this modification for regression discontinuity trees.

Finally, there are two important issues specific to RD literature: selection of observations that are close to the threshold parameter to get 'local-randomization' and to get a precise conditional expectation estimate at the threshold from above and below.

[26]The candidates of γ, coming from weakest-link pruning: using the large tree built on the whole training sample, γ values represent those penalty parameters which would result in a smaller tree for this large partition. During cross-validation, these scaled candidate γ values are used to prune back the trees. Scaling adjusts to the 'typical value' for the accompanied sub-tree. 'Long introduction for rpart package' gives an excellent overview on the technicalities of the cross-validation as well, which is available at `https://rdrr.io/cran/rpart/f/inst/doc/longintro.pdf` .

[27]In the case of flat cross-validation criterion function it is well accepted to use 'one standard error rule': taking not the smallest value as the optimal, but the largest γ value which is within the one standard error range of the smallest value. This results in a smaller tree, which is easier to interpret and it filters out possible noise features, which would be relevant with the smallest cross-validation value.

[28]Note that there is a trade-off: if there are multiplicities in the value of the feature and it is not truly continuous, it may be faster to calculate the inverse for each candidate splitting value.

[29]Published at `https://github.com/susanathey/causalTree`

CEU eTD Collection

These issues, imply bandwidth selection procedure and choosing the order of polynomial during the estimation.
Let us start with bandwidth selection. This paper does not offer a non-parametric method to estimate the conditional expectation function in the running variable, only parametric polynomials.[30] However, in practice a properly working solution is to use an under-smoothing bandwidth on the full sample, then restrict the used sample and employ the algorithm in this restricted sample. There is a recent discussion on selecting the order of polynomials used during the estimation (see, e.g., Gelman and Imbens (2019) or Pei et al. (2020)). [31] This paper offers a natural approach to select the order of polynomials: use the cross-validation procedure jointly with the complexity parameter to select q. As the estimated EMSE value is an unbiased estimator, it will lead to EMSE optimal order of polynomial selection as well.

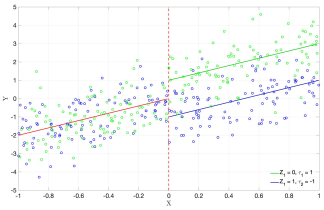
1.4 Monte-Carlo simulations

For Monte Carlo simulation, I created five different designs investigating different forms of heterogeneous treatment effects in RD. The first data generating process (DGP) is a simple example to demonstrate how the algorithm finds a simple tree structured DGP. Its simplicity comes from employing only two treatment effects which are defined by one dummy variable. The conditional expectation function (CEF) is linear and homogeneous across the leaves. DGP-2 imitates the step-function approximation nature of the algorithm: it has a continuous treatment effect function dependent on a single continuous variable, while the conditional expectation function is a linear function of another pre-treatment variable. DGP-3 to 5 revisit the simulation designs of Calonico et al. (2014) with non-linear conditional expectation function. I add heterogeneity to the treatment effects for DGP-3 and DGP-4, parallel to DGP-1 and DGP-2: two treatment effects defined by a dummy variable for DGP-3 and a continuous CATE for DGP-4. DGP-5 shows how the algorithm performs when there is no heterogeneity in the treatment effect. Figure 1.4 shows the different sharp RD designs.

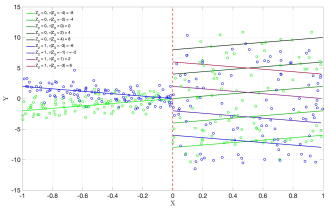
[30] It would be an interesting research avenue to extend the EMSE criterion to non-parametric estimators as well. The algorithm could handle this extension naturally by including splitting the running variable as well, but the bias-variance trade-off would alter the behavior of the criterion.

[31] The main recommendation of Gelman and Imbens (2019) is to use low order (local) polynomials to avoid noisy estimates. Pei et al. (2020) proposes a measure that incorporates the most frequently used non-parametric tools to select the order of polynomial.

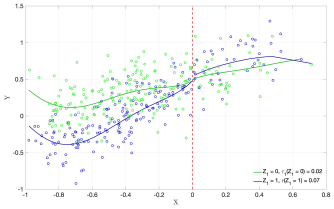
CEU eTD Collection



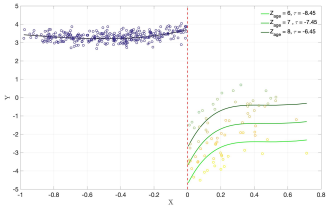
DGP-1: Two treatment effects with homogeneous CEFs



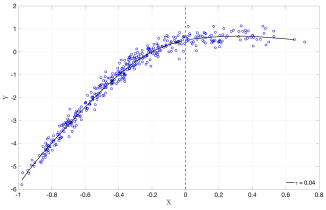
DGP-2: Heterogeneous CEF with continuous CATE



DGP-3: Two treatment effects with non-linear CEF



DGP-4: Continuous CATE with non-linear CEF



DGP-5: Homogeneous treatment effect with non-linear CEF

Figure 1.4: Monte Carlo simulation designs

During the simulations, I use three different sample sizes: $N = 1,000$; $5,000$ and $10,000$ to investigate the effect of the sample size on the algorithm. As the method splits the initial sample, I use half of the observations for the training and the other half for the estimation sample. I use $MC = 1,000$ Monte Carlo repetition and the variation comes from a normally distributed disturbance term, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, whereas the features are uncorrelated[32]. For DGP 1 and 2, I use $\sigma_\epsilon^2 = 1$ and for DGP 3-5, $\sigma_\epsilon^2 = 0.05$.

Next, I discuss the main features of the proposed DGPs. For the complete specification refer to Appendix A.5.

DGP-1: Imitates a simple tree structure: there is two distinct treatment effects, conditioning on one binary variable. There is also an additional irrelevant binary variable. Both of them generated by using the probabilities of $P(Z_k = 1) = 0.5\,, k = \{1, 2\}$.

- $\tau(Z_1 = 1) = 1$, $\tau(Z_1 = 0) = -1$, number of features: 2

DGP-2: The second design follows Athey and Imbens (2016), who uses heterogeneous conditional expectation function along with continuous treatment effect. DGP-2 is modified for sharp RD and uses four different features: two binary (Z_1, Z_2 with $P(Z_1 = 1) = P(Z_2 = 1) = 0.5$) and two continuous ($Z_3, Z_4 \sim U(-5, 5)$) variables. The conditional expectation is a function of Z_2 along with the running variable, but has no effect on the magnitude of the

[32]This means, during the simulations there are no issues with (highly) correlated features, which would alter the stability of the resulting trees.

CEU eTD Collection

treatment. CATE is a linear function of Z_3. This design shows a clean behavior for the step-function approximation, while allowing heterogeneity in the conditional expectation function.

- $\tau(Z_3) = 2\,Z_3$, number of features: 4

The last three designs investigate the performance of the algorithm when the conditional expectation function is non-linear. These setups use the functional forms proposed by Calonico et al. (2014) and imitate different RD applications. This exercise exhibits how the performance of the algorithm alters compared to the linear cases. To compare the behavior of the method I induce heterogeneity in the treatment effects similarly as in DGP 1 and 2, but add more potential pre-treatment variables.

DGP-3: Imitates Lee (2008) vote-shares application. I assume two treatment effects with different conditional expectation functions for the leaves. I use 52 dummy variables representing political parties and different states. The artificial political party dummy (Z_1) is relevant and has an effect on both treatment and the functional form. Artificial state variables are irrelevant.

- $\tau(Z_1 = 1) = 0.02$, $\tau(Z_1 = 0) = 0.07$, number of features: 52

DGP-4: Follows Ludwig and Miller (2007), who studied the effect of Head Start funding to identify the program's effects on health and schooling. I assume a continuous treatment effect based on the age of participants (Z_1), while adding (irrelevant) dummies representing different continents.

- $\tau(Z_1) = -0.45 - Z_1$, number of features: 7

DGP-5: An alternative DGP by Calonico et al. (2014), which adds extra curvature to the functional form. This design is the same as in Calonico et al. (2014), thus there is only one homogeneous treatment effect.

- $\tau = 0.04$, number of features: 52

To evaluate the performance of the algorithm I am using three different measures. The first measure investigates, whether the proposed estimable EMSE criterion is a good proxy to minimize the ideal in-feasible criterion (equation 1.6). For transparent comparison, I calculate this in-feasible criterion on a third independent evaluation sample, containing $N^{eval} = 10,000$ observations. The criterion is calculated on this evaluation sample, and the CATE estimator comes from the tree, which is grown on the training sample and estimated on the estimation sample. The Monte Carlo average of this estimate is reported as "*inf. MSE*". The second measure is the average number of leaves on the discovered tree $(\#\hat{\mathrm{I}})$. DGP-2 and 4 with continuous CATE function should have an increasing number of leaves as one increases the number of observations. This would imply proper step-function approximation nature of the algorithm as more observations allow the algorithm to split more along with the relevant feature. For DGP-1, 3, and 5 the number of leaves should be the same as the number of distinct treatment effects in the true DGP. This measure may be misleading in cases when the algorithm

CEU eTD Collection

finds different treatment effects, but the conditioning variables are not the same as in the true DGP (e.g., in DGP-1 the algorithm splits with Z_2 instead of Z_1, which would result in the same number of leaves, but not finding the true DGP). Therefore I also calculate the percent how many times the true DGP is found, when the DGP has a tree structure. (For DGPs with continuous CATE this is measure is not reported, as the algorithm only provides a step-function approximation of the true CATE.). Table 1.1. reports the results on the algorithm performance.

DGP	N	inf. MSE	$\#\hat{\Pi}$	DGP found (%)
DGP-1	$N = 1,000$	0.0620	2.00	100%
	$N = 5,000$	0.0135	2.04	96%
	$N = 10,000$	0.0065	2.04	96%
DGP-2	$N = 1,000$	9.3103	2.00	-
	$N = 5,000$	1.3852	7.72	-
	$N = 10,000$	0.9233	11.68	-
DGP-3	$N = 1,000$	0.0013	1.00	0%
	$N = 5,000$	0.0003	2.00	100%
	$N = 10,000$	0.0001	2.00	100%
DGP-4	$N = 1,000$	1.3904	1.00	-
	$N = 5,000$	0.4160	3.00	-
	$N = 10,000$	0.2013	4.92	-
DGP-5	$N = 1,000$	0.0007	1.00	100%
	$N = 5,000$	0.0002	1.03	97%
	$N = 10,000$	0.0001	1.02	98%

Table 1.1: Monte Carlo averages for performance measures
Number of true leaves: $\#\Pi_{DGP-1} = 2$, $\#\Pi_{DGP-3} = 2$, $\#\Pi_{DGP-5} = 1$
Algorithm setup: using the smallest cross-validation value to select γ^*,
$q = 1$ for DGP 1 and 2 and $q = 5$ for DGP 3,4 and 5.

From Table 1.1 one can see that the algorithm works considerably well. The infeasible MSE is decreasing in N for each setup. This supports the theoretical claim that the estimable EMSE criterion is a proper proxy for the infeasible MSE, thus the resulted tree is MSE optimal in this sense. The average number of leaves on the discovered trees reflects the expectations. For DGP-2 and 4, where the CATE is continuous the average number of leaves is increasing in N. Note that the algorithm performs better when the CEF is linear compared to the non-linear case. For DGP-1, 3, and 5 the average number of leaves reflects the true number of leaves for the DGPs with one exception: for DGP-3 with $N = 1,000$. In this case, the algorithm does not split but gives a homogeneous treatment effect instead of the two distinct treatment effects. The measure of DGP found (%) reflects that the algorithm does not split along irrelevant variables but along relevant variables. Finally, results in Table 1.1 shows that the algorithm is rather conservative in discovering different treatment effects and a data intensive method.

CEU eTD Collection

In the case of DGP-1 and DGP-5, the signal-to-noise ratio is relatively high and with $N = 1,000$ it does not discover any irrelevant features (only the true DGP) – however due to randomness it should in some cases. DGP-3 on the other hand has a relatively low signal-to-noise ratio, and with $N = 1,000$ it never discovers the true DGP. Increasing the number of observations solves this problem. $N = 5,000$ observations are enough for DGP-1 and 5, but it takes $N > 10,000$ for DGP-3, showing the data intensity of the method.

Another important result for the regression discontinuity tree is Monte Carlo evidence on providing valid inference. I calculate the average bias and the actual 95% confident interval (CI) coverage for each leaf. Table 1.2 reports the Monte Carlo average of the bias for each leaf $\left(MC^{-1}\sum_{mc=1}^{MC}(\tau_j - \hat{\tau}_{j,mc})\right)$ and the actual 95% CI coverage for the different leaves conditionally whether the algorithm found the true DGP. I report only DGPs, which has a tree structure, as in cases of continuous CATEs, the leaves are varying due to different splitting values, making the reporting and aggregation over the Monte Carlo sample non-trivial.[33]

DGP 1	Leaf	$\ell_1 : \tau_1(Z_1 = 1) = 1$		$\ell_2 : \tau_1(Z_1 = 0) = -1$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-0.0121	0.95	-0.0155	0.95
	$N = 5,000$	-0.0015	0.95	-0.0022	0.94
	$N = 10,000$	0.0009	0.96	0.0003	0.95
DGP 3	Leaf	$\ell_1 : \tau_1(Z_1 = 0) = 0.07$		$\ell_2 : \tau_1(Z_1 = 1) = 0.02$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-	-	-	-
	$N = 5,000$	0.0002	0.94	0.0000	0.95
	$N = 10,000$	-0.0000	0.95	0.0004	0.96
DGP 5	Leaf	Homogeneous Treatment, $\tau = 0.04$			
	Estimates	avgerage bias		actual 95% CI coverage	
	$N = 1,000$	-0.0001		0.95	
	$N = 5,000$	0.0001		0.96	
	$N = 10,000$	0.0004		0.95	

Table 1.2: Estimated Monte Carlo average for bias and actual 95% confidence intervals coverage for each leaves for tree structured DGPs, conditional on DGP is found

Note: For DGP-3, with $N = 1,000$, there is no case when the true DGP is found, thus no values are reported.

Table 1.2 shows that the average bias is decreasing in N for each leaf individually

[33]Note that for continuous CATE the treatment effect conditional on the leaf is still an unbiased estimator for the given feature partition and has proper standard errors, however simple aggregation by the Monte-Carlo simulation distorts these properties.

CEU eTD Collection

(at least up to 3 digits), similarly to the infeasible MSE, which is averaged over these leaves. The actual 95 % CI coverage reflects properly the nominal value. These results provide evidence of valid inference for the estimated CATE function.

1.5 Heterogeneous effect of going to a better school

To show how the algorithm works in practice, I replicate and augment the heterogeneity analysis of Pop-Eleches and Urquiola (2013) on the effect of going to a better school. Furthermore, I relate my results to Hsu and Shen (2019).

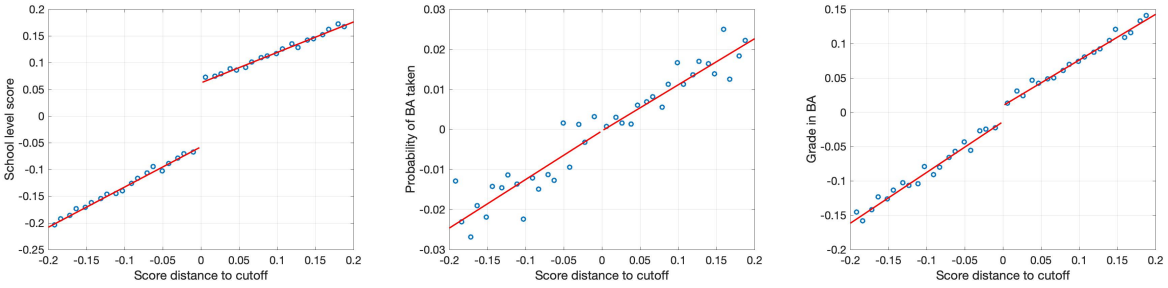
In Romania, a typical elementary school student takes a nationwide test in the last year of school (8th grade) and applies to a list of high schools and tracks. The admission decision is entirely dependent on the student's transition score, an average of the student's performance on the nationwide test, grade point average, and order of preference for schools.[34] A student with a transition score above a school's cutoff is admitted to the most selective school for which he or she qualifies. Pop-Eleches and Urquiola (2013) use a large administrative dataset (more than 1.5 million observations) and a survey dataset (more than 10,000 observations) from Romania to study the impact of attending a more selective high school during the period of 2003-2007. Based on the administrative dataset, they find that attending a better school significantly improves a student's performance on the Baccalaureate exam,[35] but does not affect the exam take-up rate.

Figure 1.5 summarizes the classic mean RD results from Pop-Eleches and Urquiola (2013). In all three graphs the horizontal axis represents the running variable, which is a student's standardized transition score subtracting the school admission cut-off. The vertical axis in Figure 1.5a) represents the *peer quality*, that each admitted student experiences, when going to school. Peer quality is defined as the average transition score for the admitted students in each school. This indicates that the higher the level of average transition score is (e.g., the admitted students performed great in the nationwide test), the better the peer quality. Figure 1.5b) shows the probability of a student taking the Baccalaureate exam, while Figure 1.5c) plots the Baccalaureate exam grade among exam-takers. In all outcomes, school fixed effects are used as in Pop-Eleches and Urquiola (2013), thus the vertical axis is centered around 0 for all plotted outcome. Both left and right graphs show a jump in the average outcome at the discontinuity point, but the jump in the exam-taking rate is quite noisy and seemingly insignificant.

[34]Grades on the nationwide test are from 1-10, where 5 is the passing score on each test. Grade point average is an average of the past years course grades for different disciplines. Order of preference for schools is a list submitted by the student before the nationwide test, showing their preferences for the schools that they apply.

[35]Marks in BA Exam vary from 1-10, where there are multiple disciplines, where in each, one needs to score above 5 and achieve a combined score of more than 6 to pass the BA Exam.

CEU eTD Collection



(a) Average transition score (b) Baccalaureate taken (c) Baccalaureate grade

Figure 1.5: Bin-scatter for main (pooled) RD results of Pop-Eleches and Urquiola (2013), using school fixed effects

1.5.1 Revisiting heterogeneity analysis of Pop-Eleches and Urquiola (2013)

First, I revisit Pop-Eleches and Urquiola (2013) heterogeneity analysis on the intent-to-treat effects using peer quality (level of school average transition score) and the number of schools in town as the sources of heterogeneity, using the administrative data between 2003 to 2005.[36] Similarly, I restrict the sample to observations which lies within the ± 0.1 interval of the admission cutoff for the running variable and I use the same linear specification. Pop-Eleches and Urquiola (2013) inspect heterogeneity in the treatment effect with pre-specified sub-samples. The first two sub-samples are differentiated by the level of peer quality effect. Pop-Eleches and Urquiola (2013) investigate treatment effects for students in the top and bottom tercile for the school level average transition score. The second analysis focuses on the numbers of schools in town and create groups defined by having i) four or more schools in towns, ii) three schools or iii) two schools only. Instead of using these pre-specified (ad-hoc) groups, I use the algorithm to identify the relevant groups and split values. I also use these two variables[37] to explore the heterogeneity, but use them simultaneously allowing for finding different non-linear patterns in the treatment effect. See more details about these variables in Appendix A.7.

Let us consider the peer quality effect as the outcome, which is measured by the average transition score at their respective school. Pop-Eleches and Urquiola (2013) find significant positive treatment effects in all five groups. The regression discontinuity tree algorithm finds a much more detailed tree, containing 24 leaves, which is an indication of a continuous CATE function. Instead of showing a large tree, Figure 1.6 shows the marginalized treatment effects along the two variables.[38] Figure 1.6a) shows

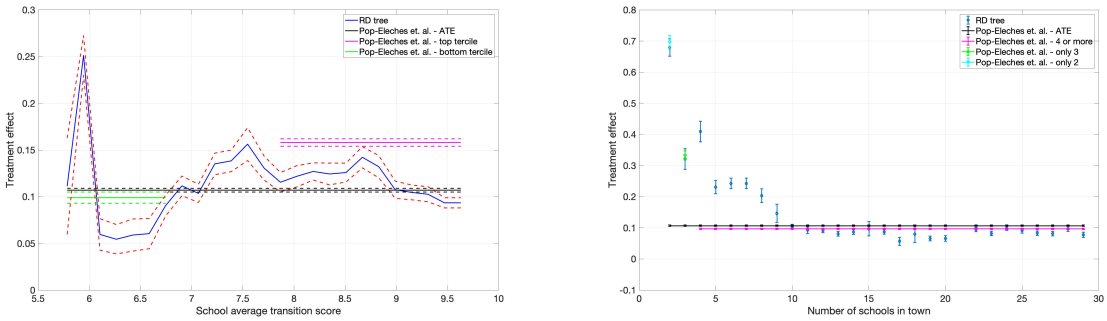
[36] Referring to Table 5 in Pop-Eleches and Urquiola (2013, p. 1310). See more details in Appendix A.7.

[37] I add dummy variables as well to search for a certain number of schools in the town.

[38] I have calculated the treatment effect for each observation then averaged them over the non-plotted variable. In case of number of schools, I take students with the same number of schools in town and average them along the level of school average transition score.

CEU eTD Collection

the treatment effects conditional on the level of school average transition score.[39] The blue line represents the CATE function found by the algorithm, the black line shows the overall average treatment effect, while the green and pink lines show the treatment effects reported by Pop-Eleches and Urquiola (2013) for the bottom and top tercile. Figure 1.6b) shows the heterogeneity in the treatment effects along the number of schools. Similarly to the previous plot, the different coloured error-bars show the treatment effects for the different models.



(a) Level of school avg. transition score (b) number of schools

Figure 1.6: CATE for peer quality, intent-to-treat effects, using school fixed effects, standard errors are clustered at student level

It is interesting to compare the algorithm's result (blue line) to Pop-Eleches and Urquiola (2013) results (green and pink lines). Figure 1.6a) shows that the 'bottom tercile' (green line) effect should be decomposed into two further parts: students with the lowest scores have high treatment effect, but students above score 6, but below 6.8 face the lowest treatment effects. This indicates a different mechanism of the treatment for these groups and aggregating them to bottom tercile may lead to misleading suggestions. Conditioning the treatment effects on the number of schools results in the same conclusion for two and three schools,[40] but as Figure 1.6b) shows the treatment effect suggested by the algorithm is still higher than the average for towns with 4-9 schools and it is significantly lower for towns with 18-20 schools.

Investigating the treatment effect on the probability of taking the Baccalaureate exam – in contrast with Pop-Eleches and Urquiola (2013), who do not find significant treatment effects – the algorithm discovers a group where there is a significant negative effect on the exam taking rate.

[39] I used 50 equal sized bins to group school average values.

[40] The treatment effects are not the same as the algorithm uses only half the sample to estimate the CATE, thus the blue line is not varying exactly around the pink line (or the black/green lines).

CEU eTD Collection

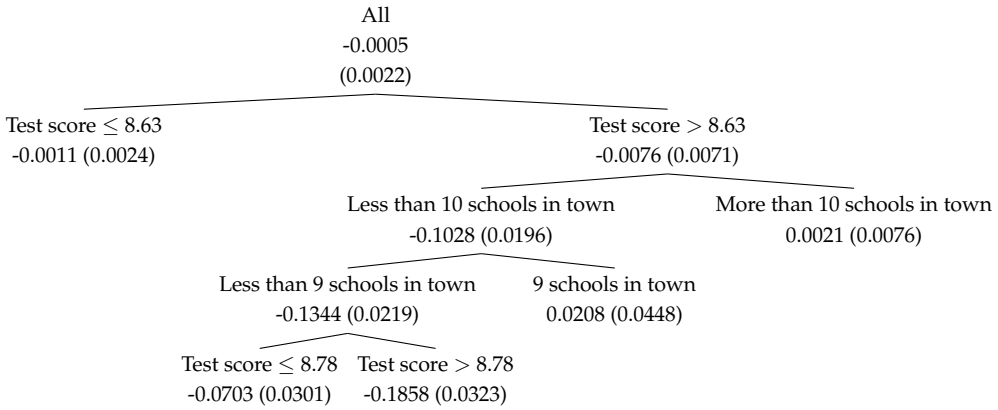


Figure 1.7: Conditional treatment effects for probability of taking Baccalaureate exam, intent-to-treat effects, using school fixed effects, standard errors are clustered at student level

Although the majority of the discovered groups have non-significant treatment effect, all these splits are needed to find the group which has a significant negative 19% treatment effect on the probability of taking the BA exam. As this value is surprisingly high, a researcher or policy maker may want to understand the background of this (sub-)population. The group is defined as students whose level of school average transition score is above 8.78 (top 10%) and in their town there are less than 9 schools. Thus these students are admitted into an extremely competitive school, but there are few (or no) outside option to change school within the town. Overall, there are more than 20,000 cases that fall into this category. This result is aligned with the negative peer effect that Pop-Eleches and Urquiola (2013) report. Namely, on a distinct survey data set they find evidence that comparatively less talented students in competitive schools are less likely to go and take the Baccalaureate exam.

Last, the heterogeneity found by the algorithm in the value of Baccalaureate grade is the simplest as there are only two relevant groups. One group contains students whose school average transition score is above the median (to be exact, 7.4, which is the 44-th percentile in the sample). These students can expect a 0.0282 (0.0054) higher exam grade on the Baccalaureate exam if going to a better school, while students below this splitting value can expect only 0.0152 (0.0061) higher Baccalaureate exam grade. The algorithm does not split further, thus providing no further evidence on heterogeneity across the number of schools in town within these groups.[41]

Finally, let me relate these results to Hsu and Shen (2019). They search for heterogeneity using peer quality as the potential source and find strong evidence for the exam-taking rate (under 1% p-values) and weak evidence for BA grade (around 10% p-values) among schools. Although they restrict their sample to towns with two or three schools and estimate the local average treatment effect, the conclusion is the same, values of school level average test score have an impact on the level of treatment effect.

[41] If one only uses the number of schools to find heterogeneity, the algorithm finds different treatment effects, but jointly it is non-relevant. See more details in Appendix A.7, Table A.5.

CEU eTD Collection

1.5.2 Exploring heterogeneity in survey-based dataset

To explore treatment effect heterogeneity and show how the algorithm performs, when there are many covariates with potential non-linearities, I use the survey dataset from 2005-2007. This sample contains less observations, but a rich variety of socio-economic factors (e.g., gender, ethnicity, education, accessibility of internet or phone), school characteristics (e.g., novice teacher among teachers, highly certified teachers in schools) and study behavior specific questions (e.g., parents pay for tutoring, parents help students, child does homework every day, peer ranking, teacher characteristics). In the survey, there are only 135 schools located in 59 towns with 2 to 4 schools and a questionnaire was administered between 2005 to 2007. Overall, I use 29 different features to search for heterogeneity. As the survey corresponds to later years, the data includes only observations on level of school average transition scores, but not on the other two outcomes. See more detailed description in Appendix A.7.

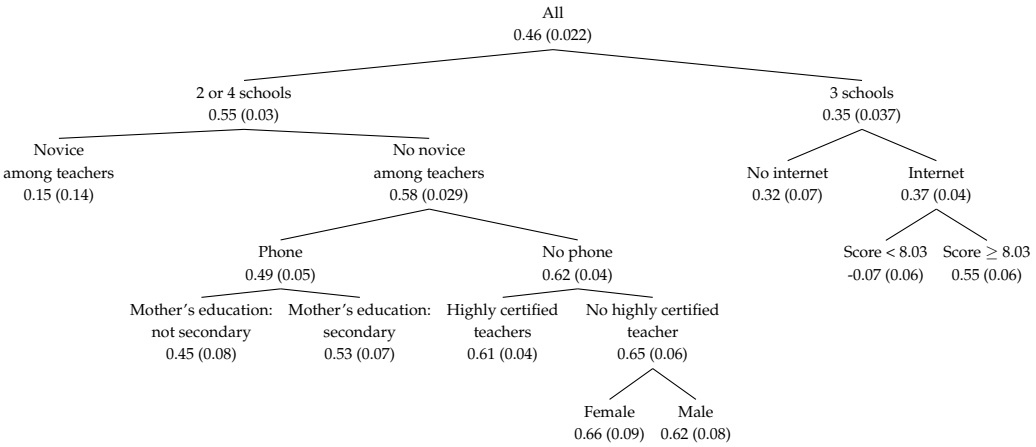


Figure 1.8: Exploring heterogeneous groups for peer quality - intent-to-treat effects, standard errors are in parenthesis and clustered on student level.

The fitted tree, shown in Figure 1.8, suggests an informative result: for towns with 2 or 4 schools, admitted students (on average) have 0.55 higher scores. If one goes further, the tree suggests that when having a novice among teachers (less than 2 years of experience), the treatment effect may disappear (although only 319 cases fall into this category). It is interesting that having a phone would somewhat reduce the peer quality effect in 2005-07 but it should also be noted that during the studied time-period it was not common for students aged 12-14 to have phones. I also find interesting splits with respect to i) education of the mothers, ii) if there are teachers with highest state certification in the school, and iii) if the student gender is male or female. These are indeed interesting splits, but statistically non-distinguishable.[42] Heterogeneity among groups in the other branch is more informative and more robust. If there are three schools and accessibility of the internet, the level of school average transition scores is an important split to identify a group. For schools, which have student scores above the bottom

[42] As the cross-validation criterion is quite flat with the one standard error rule, these splits are pruned back.

CEU eTD Collection

tercile (8.03 is the 35% percentile in this sample), the peer quality effect is similar to students with 2 or 4 schools in town (0.55 higher scores). However, for students in these schools with internet access at home, but below the bottom tercile, the peer quality effect is insignificant. This suggest potential segregation for this discovered group and encourages the researcher or policy maker to make further investigation on this specific group. Finally, let me note that the results are quite robust to randomization of the observations in the training/estimation sample. Some of the splits may vary but the main conclusion is similar in most of the cases.

1.6 Extension to fuzzy designs

The method can be extended to fuzzy designs as well, where the probability of treatment needs not change from 0 to 1 at the threshold, and can allow for a smaller jump in the probability of assignment.

Let me use a distinct variable T for getting the treatment in case of fuzzy design. As the probability does not change from 0 to 1 at the threshold, there are different types of participants, depending on whether they are subject of the treatment or not. Compliers are units that get the treatment if they are above the threshold but do not get the treatment if they are below: $T(1) - T(0) = 1$. Always takers get the treatment regardless of whether they are below or above the threshold, while never takers never take the treatment regardless of the threshold value. For both behavior, the following applies: $T(1) - T(0) = 0$. As in classical fuzzy RD, I eliminate by assumption defiers, who does not take the treatment if above the threshold and takes the treatment if below the threshold.

Fuzzy RD identifies treatment effect for compliers, thus extending the algorithm to fuzzy designs result in conditional local average treatment effects (CLATE). To identify CLATE, the following assumptions are needed:

Identifying assumptions of CATE in fuzzy RD

i) $\lim_{x \downarrow c} \mathbb{P}\left[T = 1 | X = x\right] \geq \lim_{x \uparrow c} \mathbb{P}\left[T = 1 | X = x\right]$

ii) $\mathbb{E}\left[Y(d) \,|\, T(1) - T(0) = d', X = x, Z \in \ell_j(\Pi)\right]$ exists and continuous at $x = c$ for all pairs of $d, d' \in \{0, 1\}$ and for all leaves j in the tree.

iii) $\mathbb{P}\left[T(1) - T(0) = d \,|\, X = x, Z \in \ell_j(\Pi)\right]$ exists and continuous at $x = c$ for $d \in \{0, 1\}$, $\forall j$ and for all leaves j in the tree.

iv) Let, f_j denotes the conditional density of x in leaf j. In each leaf j, c must be an interior point of the support of $f_j(x)$.

Identification assumptions are similar to classical fuzzy RD, but it needs to be valid within each leaf. Assumption i) rules out defiers as it requires a non-negative discontinuity in the probability of taking the treatment around the threshold. This is not only an assumption, but a built-in restriction for the algorithm. If this condition's sample

CEU eTD Collection

analogue is not satisfied, it is not considered as a valid split. Assumptions ii) and iii) ensure the existence and continuity of the expected potential outcomes at the threshold value for always-takers, compliers and never-takers with respect to the running variable within each leaf, while assumption iv) ensures that the conditional density of x for each leaf is well behaving, similarly to sharp RD.
Under these assumptions, the CLATE for RD tree is identified as

$$\begin{aligned}\tau_{FRD}(z;\Pi) &= \frac{\lim_{x\downarrow c}\mu_+^y(x,z;\Pi) - \lim_{x\uparrow c}\mu_-^y(x,z;\Pi)}{\lim_{x\downarrow c}\mu_+^t(x,z;\Pi) - \lim_{x\uparrow c}\mu_-^t(x,z;\Pi)} \\ &= \frac{\mu_+^y(c,z;\Pi) - \mu_-^y(c,z;\Pi)}{\mu_+^t(c,z;\Pi) - \mu_-^t(c,z;\Pi)} \\ &= \frac{\tau^y(z;\Pi)}{\tau^t(z;\Pi)} \\ &= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\ \frac{\alpha_{+,j}^y - \alpha_{-,j}^y}{\alpha_{+,j}^t - \alpha_{-,j}^t}\end{aligned}$$

where, similarly to sharp RD, I use a parametric functional forms for approximating the conditional expectation functions for both the participation and outcome equations below and above the threshold,

$$\begin{aligned}\mu_+^t(x,z;\Pi) &= \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\delta_j^{-,t}\ , \qquad \mu_+^y(x,z;\Pi) = \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\delta_j^{-,y}\ , \\ \mu_-^t(x,z;\Pi) &= \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\delta_j^{+,t}\ , \qquad \mu_-^y(x,z;\Pi) = \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\delta_j^{+,y}\ , \\ \delta_{j,\pm}^t &= \left[\alpha_{j,\pm}^t, \beta_{j,1,\pm}^t, \ldots, \beta_{j,p,\pm}^t\right]'\ , \qquad \delta_{j,\pm}^y = \left[\alpha_{j,\pm}^y, \beta_{j,1,\pm}^y, \ldots, \beta_{j,p,\pm}^y\right]'\end{aligned}$$

The sample estimates for fuzzy design is provided in Appendix A.3.

Using the same logic to find the optimal EMSE tree, I minimize the expected mean squared error function over the estimation and test sample. In case of homoscedastic disturbance terms within each leaf, the estimable EMSE criterion for fuzzy designs is given by

$$\begin{aligned}\widehat{EMSE}_{FRD}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}_{FRD}^2(Z_i;\Pi,\mathcal{S}^{te}) \\ &+\left(\frac{1}{N^{te}}+\frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi}e_1'\left(\frac{\mathcal{V}_{+,j}}{p_{+,j}^{est}}+\frac{\mathcal{V}_{-,j}}{p_{-,j}^{est}}\right)e_1\end{aligned}$$

where

$$\mathcal{V}_{\pm,j} = \frac{\hat{M}_{\pm,j}^{-1}}{\hat{\tau}_j^t(Z_i;\Pi,\mathcal{S}^{te})^2}\left(\hat{\sigma}_{\pm,j}^{2,y} + \frac{\hat{\tau}_j^y(Z_i;\Pi,\mathcal{S}^{te})^2}{\hat{\tau}_j^t(Z_i;\Pi,\mathcal{S}^{te})}\hat{\sigma}_{\pm,j}^{2,t} + \frac{\hat{\tau}_j^y(Z_i;\Pi,\mathcal{S}^{te})}{\hat{\tau}_j^t(Z_i;\Pi,\mathcal{S}^{te})}\hat{C}_{\pm,j}^{y,t}\right)$$

CEU eTD Collection

is the within leaf variance of the outcome equation at the threshold, estimated from above (+) or below (-) and $\hat{\tau}_j^t(\cdot)$, $\hat{\tau}_j^y(\cdot)$ are the j'th leaf treatment effect estimated on the participation equation ($\hat{\tau}_j^t(\cdot)$) and on the outcome equation ($\hat{\tau}_j^y(\cdot)$). $\sigma_{\pm,j}^{2,t}$, $\sigma_{\pm,j}^{2,y}$ and $C_{\pm,j}^{y,t}$ are estimators for the variances and co-variance for the leaf-by-leaf disturbance terms. See the derivations in the Appendix, Section A.3.

The EMSE criterion for the fuzzy design combines the jumps in the outcome and in the participation equation along with the variance. This means that if there is a difference in two groups in the participation probabilities at the threshold or in the outcome equation, then the EMSE criterion results in a lower value and finds this difference. Similarly, if the variance of $\hat{\tau}^y$, $\hat{\tau}^t$ or their co-variance gets lower by a split, the EMSE criterion will be lower, thus even if there is no big change in the treatment effect, but there is in its variance, the algorithm considers this split.

One feature of this criterion is that, if the changes in the jump in the outcome equation and in the participation equation are with the same magnitude – resulting in the same treatment effect – then the EMSE criterion does not changes. If one is interested in heterogeneity in the participation effect and in the intent-to-treat effect separately as well, then it is possible to use sharp design for both equation separately and then assemble the results from the two trees.

1.7 Conclusions

The paper proposes an algorithm, that uncovers treatment effect heterogeneity in classical regression discontinuity (RD) designs. Heterogeneity is identified through the values of pre-treatment covariates and it is the task of the algorithm to find the relevant groups. The introduced honest "*regression discontinuity tree*" algorithm ensures a fairly flexible functional form for the conditional treatment effect (CATE) with valid inference, while handling many potential pre-treatment covariates and their interactions.

The properties of the CATE function for sharp regression design is analysed in detail and the paper shows the properties of the algorithm. An estimable EMSE criterion is put forward, which uses the specifics of RD setup, such as distinct estimation of polynomial functions below and above the threshold. Furthermore the algorithm utilize two distinct samples to get valid inference.

Monte Carlo simulation results show that the proposed algorithm and criterion work well and discover the true tree in more than 95% of the cases. The estimated conditional treatment effects - if the true tree is found - are unbiased and the standard errors provide proper estimates for 95% confident interval coverage.

Finally, the paper shows how one can utilize the algorithm in practice. I use Pop-Eleches and Urquiola (2013) data on Romanian school system, and uncover heterogeneous treatment effects on the impact of going to a better school. The algorithm shows a more detailed picture, when revisiting the heterogeneity analysis done by Pop-Eleches and Urquiola (2013). Furthermore, results suggest that i) in the most competitive schools without an outside option, students are less likely to take the Baccalau-

CEU eTD Collection

reate exam when going to a better school indicating a negative peer effect; and ii) there is a no positive peer quality effect for students who scored in the lowest 35% with internet access. The discovery of these groups encourage further investigations and future research may help to better understand the different effects when students are admitted to a better school.

CEU eTD Collection

10.14754/CEU.2021.06

Chapter 2

Modelling with Discretized Continuous Covariate

joint with Felix Chan and László Mátyás

2.1 Introduction

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

CEU eTD Collection

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

$$x_i^* = \begin{cases} z_1 & \text{if } c_0 \leq x_i < c_1 \quad \text{or} \quad x_i \in C_1 = [c_0, c_1) \quad \text{1st choice,} \\ z_2 & \text{if } c_1 \leq x_i < c_2 \quad \text{or} \quad x_i \in C_2 = [c_1, c_2), \\ \vdots & \vdots \\ z_m & \text{if } c_{m-1} \leq x_i < c_m \quad \text{or} \quad x_i \in C_m = [c_{m-1}, c_m), \\ \vdots & \vdots \\ z_M & \text{if } c_{M-1} \leq x_i \leq c_M \quad \text{or} \quad x_i \in C_M = [c_{M-1}, c_M], \\ & \text{last choice.} \end{cases} \tag{2.1}$$

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

It is not uncommon among empirical researchers to estimate linear regression models using x_i^* instead of x_i as the latter is not available.[2] Manski and Tamer (2002) show that the parameters in those cases are not point-identifiable, even though they may be partially identifiable. That is, it is possible to identify a region where the true parameters reside. This paper echoes the results in Manski and Tamer (2002) and shows that the Least Squares (LS) estimator is inconsistent in general and can only be consistent in a few very specific and restrictive cases.

More importantly, this paper proposes a new data gathering technique, called *split sampling*,[3] which can map the underlying distribution of the unobserved random variables, and thus, lead to consistent estimation of the parameters in (linear) regression models. The basic idea is to allow each survey to have different class boundaries. This induces additional information on the distribution of the random variables when considering all the responses as a whole. The proposed techniques do not induce any disincentive for respondents since the number of choices of each question remains the same. It also does not create additional complexity in the design of the questions, since the adjustments focus on the responses rather than the questions. Perhaps more importantly, the proposed solution focuses on the data collection stage and is invariant to the relation between the variables.

[1] A complete list of the notations used in the paper is given in Appendix B.4.

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

[3] The term *split sampling* in this paper is not related to the technique occasionally used in chromatography (Schomburg et al., 1977, Schomburg et al., 1981) or methods in machine learning, which splits the initial sample into folds.

CEU eTD Collection

The organisation of the paper is as follows. Section 2.2 motivates the problem from both empirical and theoretical perspectives. It shows that LS is inconsistent in general, except in a few restricted cases, and provides support to the results in Manski and Tamer (2002) on the limit of identification when using discretized data that share the same boundary points. Section 2.3 introduces the two split sampling techniques namely, the *magnifying* and *shifting* methods, that allow consistent estimation of the underlying distribution as well as of the parameters in the linear regression model using discretized data. The finite sample performance of these techniques is analysed in Section 2.4. Section 2.5 discusses some possible extensions of the techniques and concluding remarks are made in Section 2.6. All technical proofs and additional Monte Carlo results can be found in the Appendix.

2.2 Motivation

Consider the following data generating process

$$y_i = w_i'\gamma + x_i'\beta + u_i, \tag{2.2}$$

and the following linear regression model

$$y_i = w_i'\gamma + x_i^{*'}\beta + \varepsilon_i, \tag{2.3}$$

where $i = 1, \ldots, N$, w is a $K_1 \times 1$ vector of explanatory variables that can be directly observed, x is a $K \times 1$ vector of continuous random variables that cannot be directly observed and x^* is the corresponding $K \times 1$ vector of discretized choice variables as defined in (2.1). u_i is the idiosyncratic disturbance term of model (2.2) and $\varepsilon_i = (x_i - x_i^*)'\beta + u_i$ denotes the disturbance term of model (2.3), while γ and β are unknown parameter vectors. We also maintain the assumption of independence between individuals. The two main questions are the identification and consistent estimation of β based on model (2.3). Equation (2.2) and model (2.3) represent a common problem in empirical research.

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

$$TE_i = w_i'\gamma + \beta\, UPT_i + \varepsilon_i\,, \tag{2.4}$$

where, TE_i is the transport expenditure for individual i, w_i are 'usual' controls and UPT_i is the daily average use of public transport in commuting measured in minutes. UPT_i is not observed directly, but we observe only the individual's choice from a preset list UPT_i^* via a questionnaire.

CEU eTD Collection

We ask the use of public transport in the following way

$a)$ between 1 and 2 hours,
$b)$ between 30 minutes and 1 hour,
$c)$ between 15 and 30 minutes,
$d)$ between 5 and 15 minutes,
$e)$ less than 5 minutes,

For simplicity, let us neglect the possible answer of 'more than 2 hours' travelling time, but we will come back to this issue at Section 2.3. These choice options can be converted to given intervals and let us set the choice values (z_m) as the mid-value for each class. The discretized variable UPT_i^* has the following form

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

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

To the best of our knowledge, with the exception of Hsiao (1983), Terza (1987) and Manski and Tamer (2002), there has been no study investigating the estimation of discretized continuous variable(s) when the categories/classes are not represented by the expected values of the underlying distribution(s). There is, however, some work on related issues. Taylor and Yu (2002) consider a regression model with three multivariate normal random variables. In their setting, the response variable is correlated with the first variable while the second variable does not affect the response variable conditional on the first. They show that if one dichotomizes the first variable, the least squares estimate of the coefficient for the second variable will be biased. However, they do not extend their results to the more general settings where the response variable may depend on more than two covariates. Lagakos (1988) analyses the correct cut values for the grouping of continuous explanatory variables. He derives a test on deviating from the expected group mean and the categorized value if the group mean is known. He refers to this solution as the optimization criterion for discretizing an explanatory variable, using the argument in Connor (1972).

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

The measurement error literature has not considered the problem in details either, as it has been assumed that the class choice values are taking the expected values of

CEU eTD Collection

the known underlying distribution (Wansbeek and Meijer, 2001), or the measurement error is on top of a categorized variable (Buonaccorsi, 2010).

Hsiao (1983) shows that LS is inconsistent in general when assigning z_m using the mid-point values.[4] In a seminal paper, Manski and Tamer (2002) extend this result and show that β in model (2.3) is not point-identifiable without any further assumption and can only be partially identifiable. That is, it is possible to identify the region in which β resides. However, this region cannot be estimated using the LS estimator on model (2.3) as it is inconsistent. In fact, it can be shown that with one regressor[5]

$$\operatorname*{plim}_{N\to\infty}\left(\hat{\beta}^*_{LS}-\beta\right)=\frac{\beta\sum_{m=1}^{M}z_m\left\{\mathbb{E}(x_i|x_i\in C_m)-z_m\right\}}{\sum_{m=1}^{M}z_m^2}. \tag{2.5}$$

Equation (2.5) is insightful for two reasons. First, the right-hand side is generally not zero which shows that LS is inconsistent in general. Second, the right-hand side can be zero when $\mathbb{E}(x_i|x_i \in C_m) = z_m$. That is, when the choice value equals the expectation of the explanatory variable given its value lies in the corresponding class.

The result here also justifies the Berkson model (see Berkson (1980) and Wansbeek and Meijer (2000) pp. 29-30). That is, if $f(\cdot)$, the probability density of x_i, is known with known boundaries, the expected value of each variable in x^* can be consistently estimated. As such, the LS estimator of model (2.3) is consistent.

Another implication is that assigning mid-point values of each class to the choice values would make sense if one could safely assume that the explanatory variable follows a uniform distribution. In that case, the mid-point value equals the conditional expectation in equation (2.5).

Together with the results from Manski and Tamer (2002), there are two immediate conclusions: (i) There is a limit on the identification of parameters. That is, β cannot be point-identified under equation (2.2) and model (2.3) and the procedure for partial identification is complicated. (ii) It follows from (i) and the analysis above, that simple techniques, such as the LS estimator, do not seem to be appropriate even when partial identification is possible.

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

2.3 Split sampling

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

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

[5] Detailed derivations and in-depth analysis on the consistency of the LS estimator for Equation (2.2) and Model (2.3) can be found in Appendix B.2

CEU eTD Collection

respondents to answer the questions truthfully. Second, it cannot create additional complications in the design of the survey questions.

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

The intuition behind the method is that this leads to a better mapping of the unknown distribution of x and thus reduces the estimation bias. By merging the different split samples into one data set, which we call the '*working sample*'. With the working sample, we get $b = 1, \ldots, B$ overall number of choice classes across the merged split samples, where B is much larger than M. In a given split sample each respondent (individual i) is given one questionnaire only.[6] The set of respondents who fill in the questionnaire with the same class boundaries defines a split sample. Each split sample has $N^{(s)}$, number of observations ($s = 1, \ldots, S$, $\sum_s N^{(s)} = N$).

In this setup, a split sample looks exactly as the problem introduced above in (2.1), with the only difference across the split sample that the class boundaries are different.[7] Note that the number of observations across split samples can be the same or, more likely, different. Now a split sample looks like,

$$x_i^{(s)} = \begin{cases} z_1^{(s)} & \text{if } x_i \in C_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ & \qquad \text{1st choice for split sample } s, \\ z_2^{(s)} & \text{if } x_i \in C_2^{(s)} = [c_1^{(s)}, c_2^{(s)}), \\ \vdots & \vdots \\ z_m^{(s)} & \text{if } x_i \in C_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ z_M^{(s)} & \text{if } x_i \in C_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last choice for split sample } s. \end{cases} \tag{2.6}$$

Let us take a very simple illustrative example. Assume that $M = 2$, $S = 2$, $N = 60$, $N^{(1)} = 30$ and $N^{(2)} = 30$. Let x be a continuously distributed variable in $[0, 4]$ and define the class boundaries in the first split sample as $[0, 2)$ and $[2, 4]$, while in the second split sample $[0, 1)$ and $[1, 4]$, with $10, 20, 5$, and 25 observations respectively in each class. Next, we merge the information obtained in the two split samples in one working sample in such a way that we are not introducing any selection bias. This working sample now has $B = 3$ classes (or bins): $[0, 1)$, $[1, 2)$ and $[2, 4]$ and number of

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

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

CEU eTD Collection

observations N^{WS} with the working sample's artificially created variable x_i^{WS}. Using the information from the 2nd split sample, we know that of 30 observations 5 are in the 1st bin. Similarly, we can deduce that in the 2nd bin there are 5 observations as well, while in the last 3rd bin 20 observations (see Figure 2.1 below). Piecing this information together, we can create x_i^{WS}. Clearly, this way the working sample maps the unknown distribution of x better than any one of the two split samples.

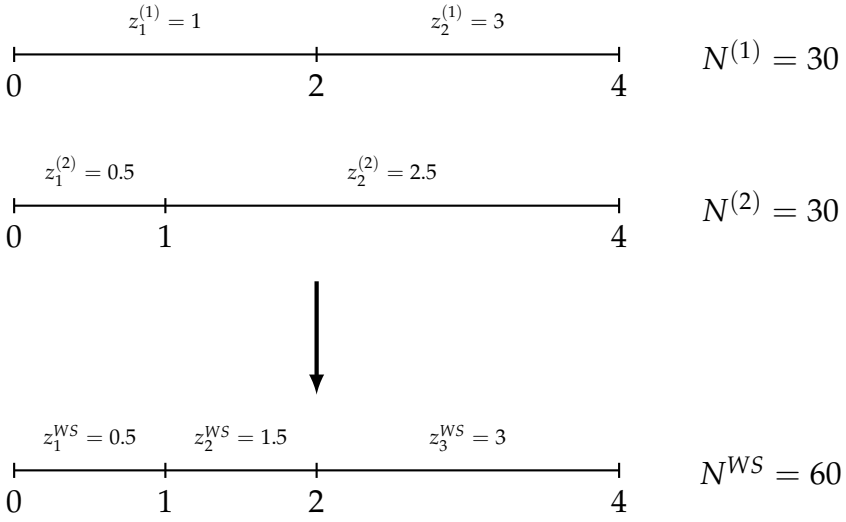


Figure 2.1: The basic idea of split sampling

2.3.1 Construction of the Working Sample

The construction of questionnaires for each split sample and the merger into the working sample can be done in many ways, depending on the assignments of boundary points $(c_m^{(s)})$ and on the choice values $(z_m^{(s)})$ for each split samples. We assume that the number of observations (N), their allocation among split samples $(N^{(s)})$ and the number of split samples (S) are given, and also that the number of choices (M) is fixed across the split samples. The class boundaries in the working sample are constructed by the union of the split samples' class boundaries, that is

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

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

CEU eTD Collection

With the creation of S split samples, we introduce

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

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

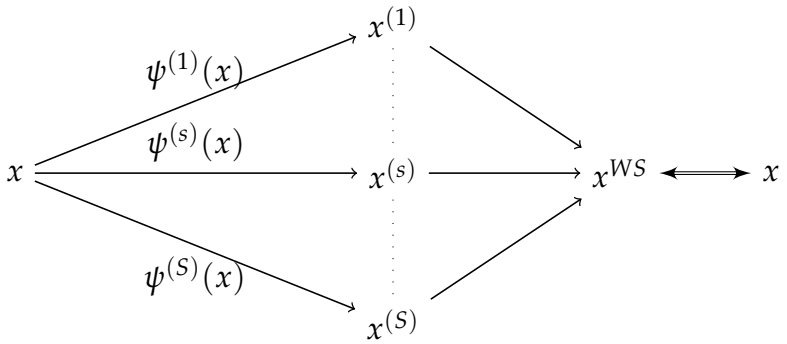


Figure 2.2: Creation of the working sample's random variable

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

$$\lim_{S \to \infty} \mathbb{E}_S \left[x^{WS} | y \right] = \mathbb{E} \left[x | y \right] , \tag{2.7}$$

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

2.3.2 Probabilities in the Working Sample

To show later on that Equation 2.7. holds for the introduced methods, we need to calculate the probability of an observation to fall into a working sample class. To derive this, we have to derive the probability of an observation falling into a given split sample's choice class and introduce an assigning mechanism taking an observation in a split sample to a working sample class. Based on these, we can get the unconditional probability for an observation to be in a given class in the working sample.

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

CEU eTD Collection

Now, we can define the probability for an observation to be in a given class in a given split sample,

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

As we observe a response in a given split sample, we would like to derive the probability of this observation falling between given boundary points in the working sample. We then assign these uniformly into the working sample's classes to avoid any systematic bias during the merging process. [8]

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

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

$$\Pr\left(x \in C_b^{WS}\right) = \sum_{s=1}^{S} \Pr(x \in \mathcal{S}_s) \sum_{m=1}^{M} \Pr\left(x \in C_b^{WS} \mid x \in C_m^{(s)}\right) \int_{c_{m-1}^{(s)}}^{c_m^{(s)}} f(x)\mathrm{d}x\,. \tag{2.8}$$

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

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

In some cases x may have infinite support, which implies classes not bounded from below and/or above. Usually, this is related to survey questions like *"... or less"* or *"... or more"*. Here we face censoring. As a consequence, the difference between the class's choice value (e.g., $z_1^{(s)}$ in Equation (2.1)) and the class's conditional mean for the underlying distribution can be potentially infinite, resulting in very large estimation biases. We will return to this issue later in the paper.

2.3.3 Magnifying Method

In the magnifying method, we magnify the domain of the answers within the original domain of the unknown distribution of x by one equally sized choice class. The size of each of the classes depends on the number of split samples (S) and the number of choice values (M). As the number of split samples increases class sizes decrease,

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

CEU eTD Collection

which is the main mechanism to uncover the unknown distribution. Figure 2.3 shows the main idea of the magnifying method: the last line shows the working sample, while above, we can see the individual questionnaires for the case of $M = 3, S = 4$.

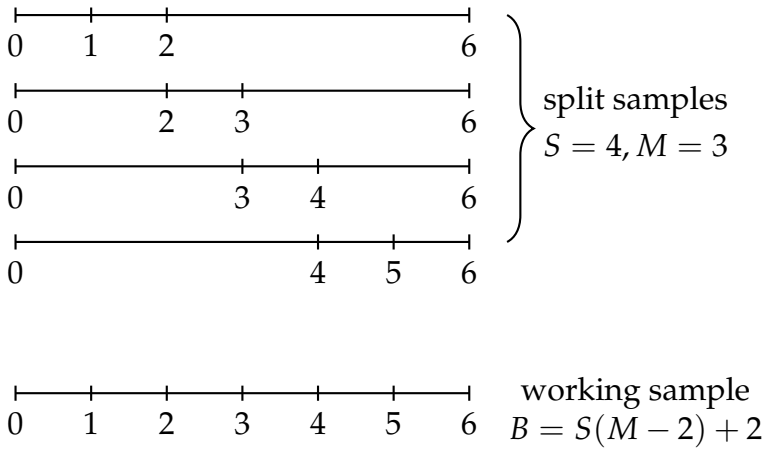


Figure 2.3: The magnifying method

The first and last split samples are slightly different from the split samples in between. They have one extra class with the same class width, while split samples in between have $M - 2$ classes with the same class width. To further explore the properties of the magnifying method, let us establish the connection between the number of magnified classes in the working sample (B), and the number of split samples (S) and choices (M)

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

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

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

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

CEU eTD Collection

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

1: For any given S and M. Set

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

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

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

else set

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

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

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

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

To show how the number of split samples affects the bias, we need the boundary points for each split sample, which can be derived as

$$c_m^{(s)} = \begin{cases} a_l \text{ or } -\infty & \text{if } m = 0, \\ a_l + mh & \text{if } 0 < m < M \text{ and } s = 1, \\ a_l + h\left[(s-2)(M-2) + M + m - 2\right] & \text{if } 0 < m < M \text{ and } s > 1, \\ a_u \text{ or } \infty & \text{if } m = M. \end{cases} \tag{2.9}$$

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

CEU eTD Collection

this process algebraically will result in the above expression.[9]

From Equation (2.9), it is clear that the class widths differ from each other within a split sample. Let $||C_m^{(s)}|| = c_m^{(s)} - c_{m-1}^{(s)}$ be the m-th class width, then for the split samples which are in-between the boundaries $(1 < s < S)$ and substituting for h, we can write

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

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

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

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

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

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

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

and for the first and last split sample

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

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

This formulation takes a_l as the starting point and expresses the boundary points

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

CEU eTD Collection

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

Based on the different magnitudes of measurement errors and depending on class widths, it is clear that there are two types of observations: The first type is $x_i^{(s)} \in \zeta$. Here, the error is the smallest and has the feature of $\lim_{S\to\infty} ||C_m^{(s)}|| = 0$. Moreover, these observations have the same class width as the working sample's classes and each of them can be directly linked to a certain working sample class by design. Formally, $\exists\, C_m^{(s)} \cong C_b^{WS}$ such that $c_m^{(s)} = c_b^{WS}$, $c_{m-1}^{(s)} = c_{b-1}^{WS}$. We call these values '*directly transferable observations*', as we can directly transfer and use them in the working sample. These observations are denoted by $x_{i,DTO}^{WS} := x_i^{(s)} \in \zeta$, $\forall s$, and the related random variable by x_{DTO}^{WS}.[10]

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

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

1: Set $m = 1, s = 1$ and $x_{i,DTO}^{WS}, y_{i,DTO}^{WS}, w_{i,DTO}^{WS} = \emptyset$.

2: If $C_m^{(s)} \in \zeta$, add observations from class $C_m^{(s)}$ to the working sample:

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

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

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

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

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

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

The probability that a *directly transferable observation* lies in a given class of the work-

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

CEU eTD Collection

ing sample can be written based on Equation (2.8) as follows

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

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

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

$$\begin{aligned}\mathbb{E}(N_b^{WS}) &= \mathbb{E}\left(\sum_{i=1}^{N} \mathbf{1}_{\{x_i \in C_b^{WS}\}}\right) \\ &= N \Pr(x \in \mathcal{S}_s) \int_{c_{b-1}^{WS}}^{c_b^{WS}} f(x)\mathrm{d}x.\end{aligned} \tag{2.10}$$

Following from Equation (2.10), consider the following assumptions,

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

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

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

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

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

Proposition 1. *Under Assumptions 1a - c,*

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

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

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

CEU eTD Collection

See the proof in Appendix B.3.1.

The proposition established convergence in distribution which allows consistent estimation of the underlying continuous distribution. This implies that the classical econometric results stand and the LS estimator is consistent for β.
Note that we can decrease c as close to 0 as we would like to. This means that there is an equal or higher number of observations than split samples. On the other hand, we exclude by assumption the case when c ≥ 1, which means that there is an equal or higher number of split samples than observations. In this case, we most certainly would not observe values for each working sample class.

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

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

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

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

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

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

Let us call $\hat{\pi}_\chi$ the 'replacement estimator' of the conditional expectation of the given class. Under the WLLN, it is straightforward to show that the 'replacement estimator' for the sample conditional averages converges to the conditional expectations, thus $\hat{\pi}_\chi \to \mathbb{E}(x | x \in C_\chi)$ as $N, S \to \infty$ and under the same assumptions as before. This also

CEU eTD Collection

implies $x^{WS}_{i,NDTO} \rightarrow \mathbb{E}(x|x \in C_\chi)$, which means using working sample $x^{WS}_{i,ALL}$ satisfies Equation 2.7.

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

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

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

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

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

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

We can obtain the asymptotic standard errors of this estimator, as if these are large, the replacement might not be favorable, as it may induce more uncertainty relative to the potential loss of efficiency by not including all the observations. To obtain the standard errors, one can think of $\hat{\pi}_\chi$ as an LS estimator, regressing $\mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}$ on $x^{WS}_{i,DTO}$. Here $\mathbf{1}_{\{x^{WS}_{i,DTO} \in C_\chi\}}$ is a vector of indicator variables, created by $2(S-1)$ indicator functions: It takes the value of one for the directly transferable observations, which are within C_χ.[11] We can now write the following:

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

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

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

and under the standard LS assumptions, we can write

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

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

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

CEU eTD Collection

The variance of the LS estimator is

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

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

As a last step we need to consider the censoring case for the magnifying method. A straightforward solution is to remove those observations which have infinite class boundary. In the magnifying method, this means to remove observations in the class/es C_1^{WS} if we have $a_l = -\infty$ or/and C_B^{WS} if $a_u = \infty$. This solution means we artificially truncate $y \to y^{tr}, x \to x^{tr}$ and $w \to w^{tr}$. For the truncated distribution, we can use all the derivations presented above, and we end up with convergence in distribution. That is, $f(x^{WS}_{DTO} \in \zeta^{tr}) \xrightarrow{d} f(x^{tr})$.[12] Furthermore, the parameter estimates $\beta^{tr} = \beta$ (under some reasonable assumptions), which implies that the LS estimator is consistent for the truncated observations. Note that truncation implies that we cannot replace the observations from the split samples with infinite boundaries, and also that the replacement estimator does not converge to the conditional expectation due to the truncation.

2.3.4 Shifting Method

The shifting method approaches the problem in a different way. It takes the original questionnaire as given, with fixed class widths, and shifts the boundaries of each choice with a given fixed value. This results in fixed class widths for the different split samples, except in the boundary classes where the widths are changing. Increasing the split sample size does not affect the boundary widths in between the support, only the size of the shift. We can approach this method in two ways. Logically we could consider the original questionnaire, and take the number of choices as fixed here, then as we shift the boundaries, add an extra class for each split sample at the boundary where, due to the shift, the class width has increased. For the mathematical derivations, however, it is more convenient to look at each split sample separately and take the number of classes in each split sample as given, with the exception of the first split sample, which we will regard as the starting benchmark. The first split sample in this case has one fewer class. The discussion below focuses on this approach and Figure 2.4 shows the split samples following this logic with $S = 4$ and with $M = 4$ classes.

[12] ζ^{tr} is the set of intervals, which do not contains C_1^{WS} and/or C_B^{WS} depending on the support. Furthermore, note that truncation implies that we cannot replace the observations from the split samples with infinite boundaries, and also that the replacement estimator does not converge to the conditional expectation due to the truncation.

CEU eTD Collection

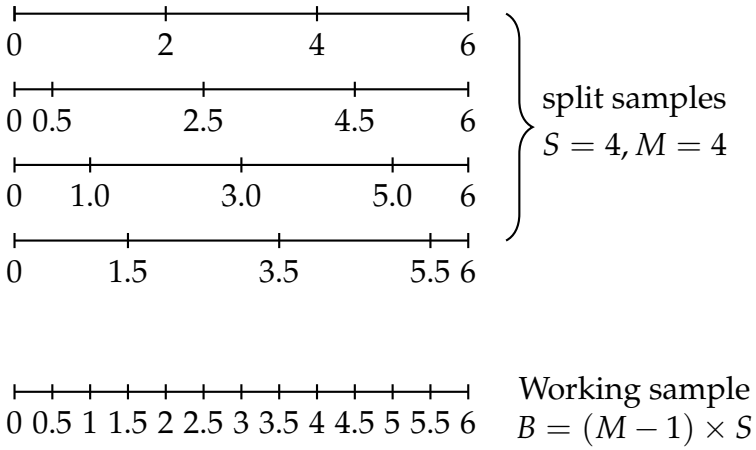


Figure 2.4: The shifting method

As Figure 2.4 shows there is one split sample (the benchmark $s = 1$) where there is one class fewer $(M-1)$, or if we prefer, we can look at the benchmark as where we shifted the boundaries with zero. To get the properties of the working sample, let us define the class widths for the first split sample as $\frac{a_u - a_l}{M-1}$. We want to split this into S part in order to be able to shift the boundaries S times in order to have S split samples. Thus, the size of the shift is $\frac{a_u - a_l}{S(M-1)}$. This way we can define the number of classes in the working sample as

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

Now, the boundary points for each split sample are

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

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

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

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

CEU eTD Collection

Some additional remarks on the boundary points:

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

Algorithm 4 The shifting method - creation of split samples $(\psi^{(s)}(\cdot))$

1: For any given S and M, set

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

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

3: If $s = 1$, set

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

else

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

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

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

The general idea is to reconstruct the underlying distribution $f(x)$, with creating a new random variable, which incorporates the information content of the boundary points.

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

CEU eTD Collection

the split samples

$$C_m^{(s)} = \begin{cases} \emptyset, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1} C_b^{WS}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-2)(M-1)}^{s-1+(m-1)(M-1)} C_b^{WS}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B} C_b^{WS}, & \text{if } m = M. \end{cases} \tag{2.11}$$

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

$$Z(s,m) = \begin{cases} \{\emptyset\}, & \text{if, } s = 1 \text{ and } m = 1, \\ \bigcup_{b=1}^{s-1} \{z_b^{WS}\}, & \text{if, } s \neq 1 \text{ and } m = 1, \\ \bigcup_{b=s-1+(m-2)(M-1)}^{s-1+(m-1)(M-1)} \{z_b^{WS}\}, & \text{if, } 1 < m < M, \\ \bigcup_{b=B-S+s-1}^{B} \{z_b^{WS}\}, & \text{if } m = M. \end{cases} \tag{2.12}$$

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

$$x_i^\dagger | x_i^{(s)} \in C_m^{(s)} = z \in Z(s,m), \text{ with } \begin{cases} \Pr(1), & \text{if } s = 1 \text{ and } m = 1, \\ \Pr(1/(s-1)), & \text{if } s \neq 1 \text{ and } m = 1, \\ \Pr(1/S), & \text{if } 1 < m < M, \text{ or} \\ \Pr(1/(S-s+1)), & \text{if } m = M. \end{cases} \tag{2.13}$$

While by the definition, there is a direct mapping between $z \in Z(s,m)$ and C_b^{WS}, we can write the probability of $x_i^\dagger \in C_b^{WS}$, using Equation (2.8) and assuming $\Pr(x \in \mathcal{S}_s) = 1/S$,

$$\Pr\left(x_i^\dagger \in C_b^{WS}\right) = \begin{cases} 0, & \text{if } s = 1 \text{ and } m = 1, \\ \frac{1}{S} \sum_{s=2}^{S} \frac{1}{s-1} \int_{C_1^{(s)} | C_b^{WS} \in C_1^{(s)}} f(x) \mathrm{d}x, & \text{if } s \neq 1 \text{ and } m = 1, \\ \frac{1}{S^2} \sum_{s=1}^{S} \int_{C_m^{(s)} | C_b^{WS} \in C_m^{(s)}} f(x) \mathrm{d}x, & \text{if } 1 < m < M, \\ \frac{1}{S} \sum_{s=1}^{S} \frac{1}{S-s+1} \int_{C_M^{(s)} | C_b^{WS} \in C_M^{(s)}} f(x) \mathrm{d}x, & \text{if } m = M. \end{cases} \tag{2.14}$$

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

CEU eTD Collection

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

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

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

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

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

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

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

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

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

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

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

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

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

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

It is possible to show that the distribution of this new variable converges to the distribution of the true underlying random variable (x) as we increase the number of split samples.

CEU eTD Collection

Proposition 2. *Under Assumptions 1a, 1c and* $\Pr(x \in \mathcal{S}_s) = 1/S$,

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

or

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

where

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

See the proof in the Appendix B.3.2.

In addition, we can investigate the speed of convergence, as we increase the number of split samples (S). The main result from the exercise is that on the boundaries of the support[13], the method converges slower, with $\frac{\log S}{S}$, while for the rest it converges with $1/S$. See the derivations in Appendix B.3.3.

Note that we cannot directly use $x_i^\dagger$ for estimation, while by design each individual observation only represents the conditional mean for the given split sample's class, and not the underlying variable's conditional expectation

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

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

$$\hat{\pi}_\tau := \left(\sum_{i=1}^{N} \mathbf{1}'_{\{x_i^{(s)} \in C_m^{(s)}\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}'_{\{x_i^{(s)} \in C_m^{(s)}\}} x_i^\dagger. \tag{2.15}$$

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

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

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

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

[13] Which is given by the maximum distance from the support given by the split samples. For the lower bound: $c_1^{(1)} + (c_2^{(S)} - c_1^{(1)})$ and for the higher bound: $c_M^{(1)} + (c_{M-1}^{(1)} - c_M^{(1)})$.

CEU eTD Collection

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

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

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

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

Under the standard LS assumption, we can write

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

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

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

where $\hat{\pi}_{\tau}$ represents the first moments of the underlying random variable, thus using $x_{i,Shifting}^{WS}$ for estimation will result in a consistent estimator for β.

Algorithm 6 Th shifting method – creation of working sample $(\Psi_{Shifting}(\cdot))$

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

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

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

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

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

$$y_i^{WS} := \left\{ y_i^{WS}, \bigcup_{j=1}^{N} y_j^{(s)} \mid \left(x_j \in C_m^{(s)} \right) \right\}$$

$$w_i^{WS} := \left\{ w_i^{WS}, \bigcup_{j=1}^{N} w_j^{(s)} \mid \left(x_j \in C_m^{(s)} \right) \right\}.$$

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

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

CEU eTD Collection

2.4 Monte Carlo Experiments

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

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

The explanatory variable, x_i, is generated as different distributions. Appendix B.1 contains detailed results with uniform, normal, exponential and weibull distributions with different parameter settings. Here we present only some demonstrative results with specifications shown in Table 2.1.

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

Table 2.1: Distributions used for the underlying random variable x.

Overall, the results are consistent with the theoretical findings. Tables 2.2, 2.3, 2.4 and 2.5 shows the Monte Carlo average of the biases $(\hat{\beta} - \beta)$, the Monte-Carlo mean absolute biases $(|\hat{\beta} - \beta|)$, the Monte Carlo standard deviation $\left(SD\left[\hat{\beta}\right]\right)$ and the average of the number of effective observations $\left(N^{eff}\right)$. The bias is in general decreasing as the number of observations and the number of split samples increase. The relative performance of the methods depends on two characteristics of the underlying distribution namely, curvature (or the classes' conditional expectations relative to the choice values, $\mathbb{E}[x \mid x \in C_m]$ and z_m), and the fraction of the probability mass covered by the surveys (or what is the probability that a certain part of the distribution is neglected by the surveys: $\Pr(x < a_l)$ or $\Pr(x > a_u)$).

The Monte Carlo setup allows us to disentangle these two effects as shown in Table 2.1. The exponential distribution with parameter $\lambda = 0.5$ provides a distribution with flat curvature. Hence, $\mathbb{E}[x \mid x \in C_m]$ and z_m are close to each other and the performance of the two methods are similar to each other. The normal distribution with $\mu_x = 0, \sigma_x^2 = 0.2$ has steeper curvature. Thus, $\mathbb{E}[x \mid x \in C_m]$ and z_m are far from each other. The magnifying method appears to be better than the shifting method in this case. In general, a large M appears to be critical if the distribution has a steep curvature. Furthermore, we have checked the truncated case, where the probability mass is completely covered by the surveys and the censored case, where there is a non-negligible part of the probability mass which cannot be utilized for the estimation.

CEU eTD Collection

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

*BM = Benchmark: simple mid-values are used. For complete specification see Table B.4 in Appendix B.1..

Table 2.2: Monte Carlo statistics for $x_i \sim Exp\left(0.5\right)$, M=3

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

*BM = Benchmark: simple mid-values are used. For complete specification see Table B.5 in Appendix B.1..

Table 2.3: Monte Carlo statistics for $x_i \sim \mathcal{N}\left(0, 0.2\right)$, M=3

CEU eTD Collection

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

*BM = Benchmark: simple mid-values are used. For complete specification see Table B.4 in Appendix B.1.

Table 2.4: Monte Carlo statistics for $x_i \sim Exp\left(0.5\right)$, M=5

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

*BM = Benchmark: simple mid-values are used. For complete specification see Table B.5 in Appendix B.1.

Table 2.5: Monte Carlo statistics for $x_i \sim \mathcal{N}\left(0, 0.2\right)$, M=5

CEU eTD Collection

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

- **Magnifying method – Truncated case**
  - $Exp(0.5; 0, 1)$: The bias decreases in S and N. The increase of M has no significant effect, because the conditional expected values and choice values are close to each other. The standard errors are decreasing in N, but slightly increasing in S. This is due to the fact that the share of directly transferable observations is decreasing in S. This implies more NDTOs, which increases the standard errors of the estimated coefficient. The absolute bias therefore first decreases, then starts to increase as the effect of standard errors starts to dominate. *Overall, with flat curvature and complete mapping of the probability mass, S/N should be above 0.01%, and M can be small.*
  - $\mathcal{N}(0, 0.2; -1, 1)$: The bias decreases in S and N. There is a significant decrease in the bias if we increase M, because the conditional expected values and choice values are not close to each other. All other results are the same as in the exponential case above. *Overall, with steep curvature and complete mapping of probability mass, S/N should be above 0.01%, and increasing M can significantly reduce the bias.*
- **Magnifying method – Censored case**
  - $Exp(0.5; 0, \infty)$ and $\mathcal{N}(0, 0.2; -\infty, \infty)$: The bias first decreases, but then it starts to increase again. This is due to the fact there are only a few observations to calculate the replacement estimator values for non-directly transferable observations. This lack of precision introduces bias during the estimation of β. The number of observations is radically decreasing as S increases and the standard errors are increasing in S. The absolute bias is mainly driven by the standard errors. *Overall, without complete mapping of the probability mass, the main driver of the bias is the number of observations in the working sample. With fewer split samples, we can decrease the absolute bias, but using too many split samples is counter-productive. $S/N > 0.01\%$ is a good rule of thumb here as well.*
- **Shifting method – Truncated case**
  - $Exp(0.5; 0, 1)$: The bias decreases in S and N. Using larger S will not help reduce the bias on the same scale as in the magnifying method due to the boundary classes' slow convergence. On the other hand, using more choices (M) will reduce the bias. It is interesting to note that the standard errors remain unchanged as S increases. The absolute bias decreases and gets smaller than in the benchmark case (with no split sampling) if we have a large amount of observations. *Overall, with complete mapping of the probability mass and flat curvature distribution, increasing M helps to reduce the bias, and*

CEU eTD Collection

increasing S also decreases it, but at a much slower rate. We need a large amount of observations in order to reduce the standard errors as well. As a rule of thumb we may use a smaller number of split samples.

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

- **Shifting method – Censored case**

  - $Exp(0.5; 0, \infty)$: The bias is decreasing in N and S, but it decreases more slowly in S, because the main drivers of the bias are the boundary classes. Increasing M will help to significantly reduce the bias. The standard errors and the absolute bias behave similarly as in the truncated case. Note that the number of observations used for the estimation is much larger than in the magnifying case! *Overall, without complete mapping of the probability mass, with flat curvature distribution, using few split samples will eliminate the main bias, and increasing M can help to reduce it even more.*

  - $\mathcal{N}(0, 0.2; -\infty, \infty)$: The bias is decreasing in N and S. Now, the boundary classes only take up a small fraction of the probability mass of the distribution, so these classes have a much smaller role in driving the bias, resulting in a much faster bias reduction. Furthermore, increasing the number of choices decreases the bias further. The standard errors, however, are slightly larger than in the benchmark case. The absolute bias is decreasing in N, M and S as well. *Overall, without complete mapping of the probability mass, with steep curvature distribution, increasing both S and M will significantly reduce the bias.*

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

  - $Exp(0.5; \cdot)$: In the truncated case the performances are very similar. In the censored case, the *bias* is smaller for the magnifying method when $S/N < 0.01\%$. In all other cases, the shifting method outperforms the magnifying one. This is due to the fact that the magnifying method drops many more observations by construction.

  - $\mathcal{N}(0, 0.2; \cdot)$: In the truncated case, the magnifying method decreases the bias much more efficiently than the shifting method. For the censored case, the results are very similar to the exponential distribution if M is small. However, the shifting method becomes better if we use larger M.

CEU eTD Collection

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

2.5 Extensions

2.5.1 Perception Effect

There is some evidence in the behavioural literature that the answers to a question may depend on the way the question is asked (see, e.g., Diamond and Hausman (1994), Haisley et al. (2008) and Fox and Rottenstreich (2003)). Let us call this the *perception effect*. The presence of this effect is independent of the implementation of the two split sampling methods. However, with split sampling, there is a way to tackle this issue, much akin to a familiar approach in the panel data literature.

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

$$x_i^{**} = \begin{cases} z_1 & \text{if} \quad c_0 < x_i + B_s < c_1 \\ \vdots & \\ z_m & \text{if} \quad c_{m-1} < x_i + B_s < c_M, \end{cases} \tag{2.16}$$

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

$$\tilde{x}_i^{**} = \tilde{x}_i^* + B_s \tag{2.17}$$

CEU eTD Collection

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

$$y_i = \beta \tilde{x}_i^{**} + u_i \tag{2.18}$$

is equivalent to

$$y_i = \beta \tilde{x}_i^{*} + B_s \beta + u_i. \tag{2.19}$$

Rewrite the above in matrix form using standard definitions gives

$$\mathbf{y} = \tilde{\mathbf{x}}^{*} \beta + \mathbf{D}\mathbf{B}\beta + \mathbf{u}, \tag{2.20}$$

where $\mathbf{B} = (B_1, \ldots, B_S)'$ and $\mathbf{D}$ is a $N \times S$ zero-one matrix that extracts the appropriate elements from $\mathbf{B}$. Thus, the estimation of β can be done in the spirit of a fixed effect estimator. Define the usual residual maker, $\mathbf{M_D} = \mathbf{I}_N - \mathbf{D}\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'$, then

$$\hat{\beta} = \left(\tilde{\mathbf{x}}^{*'}\mathbf{M_D}\tilde{\mathbf{x}}^{*}\right)^{-1}\tilde{\mathbf{x}}^{*'}\mathbf{M_D}\mathbf{y} \tag{2.21}$$

is a consistent estimator of β following the similar argument for the standard fixed effect estimator in the panel data literature.

We also need to slightly modify the replacement estimator in order for the above to hold. The main problem is to keep track of the perception effects. This means we need to keep track of which split sample each observation comes from when estimating the conditional averages. This means

$$\hat{\pi}_{\chi,s} = \left(\sum_{i=1}^{N} \mathbf{1}_{\{\tilde{x}_i^{**} \in C_\chi, \tilde{x}_i^{**} \in \mathcal{S}_s\}}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}_{\{\tilde{x}_i^{**} \in C_\chi, \tilde{x}_i^{**} \in \mathcal{S}_s\}} \tilde{x}_i^{**} \tag{2.22}$$

and as $N \to \infty$

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

This shows that equation (2.21) provides a valid replacement estimator in the presence of perception effects.

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

2.5.2 Non-linear Models

Another possible extension is to consider the application of the proposed methods in the context of non-linear models. So far the discussion has focused on the linear model as defined in equation (2.2). Given the presented methods focus on data collection,

CEU eTD Collection

they could also be applied for non-linear model. To see this, consider

$$y_i = g(x_i; \beta) + u_i \tag{2.23}$$

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

2.6 Conclusion

This paper has investigated the effects of using interval data as covariates in a linear regression model when the underlying discretized continuous variable is not observed. This situation often arises in survey data when such variables – like income – are not captured directly, but rather, are replaced by a set of m choices. Unlike other studies in the literature, our approach has considered the more realistic case when the underlying distribution of the unobserved explanatory variables is unknown and the values of each choice can be arbitrarily assigned. With fixed m, the results show that using the discretized ordered choices as explanatory variables in a linear regression will lead to biased and inconsistent parameter estimates. The well-known techniques to create consistent estimators require information from the distributions of the underlying explanatory variables, which are presumed to be unknown, and therefore cannot be applied here.

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

CEU eTD Collection

10.14754/CEU.2021.06

Chapter 3

Modelling with Discretized Continuous Depedent Variable

joint with Felix Chan and László Mátyás

3.1 Introduction

Recently, there has been increasing use of econometric models where the dependent variable is continuous but cannot be observed directly. Instead, it is observed through a discretization process. Paper or internet-based survey questions are common examples for such discretizations. These questions are usually asked in the following way e.g., *'Is your weekly personal income below* 100$*, between* 100 *and* 400$ *or above* 400$, where specific intervals are given for each option. This discretization leads to interval data, where respondents typically need to pick one option from a pre-set list, creating discrete ordinal observations from an underlying continuous variable. These responses are qualitative values, but the choices are ordered, and this order is the only quantitative information available in the resulting variable.

In the empirical literature, income is a typical example for interval data. Income is usually discretized in surveys because it dramatically improves response rate when the question is asked in the form of income categories rather than as an exact amount. Another related reason for discretization is data confidentiality: e.g., statistical offices are not allowed to give exact information on personal income. (For more details on these practices, see e.g., Duncan et al. (2001)) Just to give a few examples, Bhat (1994) shows the effects of age, employment and other socio-economic variables on income where income is observed through three different categories; Micklewright and Schnepf (2010) compare individual and household income distributions controlling for age, gender, and employment. The income is discretized and observed through single question surveys.

Modelling the conditional expectation for such a discretized dependent variable with interpretable parameters is challenging as the regression parameters are generally not point-identifiable. Here we depart from the classical econometric approach to identification. That is, we do not assume that the sample is given and the outcome

CEU eTD Collection

variable is observed as interval values, but we propose a new sampling method, which we called *split sampling*[1] that estimates the distribution of the continuous unobserved outcome rather than its discretized version. In other words, we investigate how to gather the data in order to point-identify and estimate the conditional expectation by using simple least squares regression techniques.

This paper deals with linear regression models where the dependent variable is observed through a discretization process, resulting in an interval variable. Here, let us note that there is a substantial difference between *interval* and *ordinal* variables. Interval variables have known lower and upper boundaries for each choice interval, and numeric intervals can be assigned for each observation. In principle, an interval variable has a (conditional) expectation, but it cannot be directly estimated using discretized data. The only information provided by this type of data is the lower and upper bounds of each of the categories and their frequencies. An ordinal variable embodies an even more severe information loss relative to the underlying continuous variable. The observed values have only a relational connection to each other (e.g., they are higher or lower), they are sorted into classes, and numerical values cannot be assigned to the observations. Data from these qualitative variables cannot be used to estimate any conditional moments of the underlying random variables and can only be used to obtain the frequencies of each class. In this paper we only deal with interval variables.

Manski and Tamer (2002) show that the parameters of a regression model cannot be point-identified without any further restrictive assumption, when an interval variable is used as the dependent variable. To circumvent this identification problem, there are two known solutions in the literature, both taking the discretization process as given. This paper adds a third possible solution by revisiting the discretization process itself.

The first and more popular solution in applied papers relies on the so-called ordered discrete choice models (Greene and Hensher, 2010), such as the ordered logit or probit. These models can handle the interval and ordinal variables and they aim for prediction or categorisation instead of (parameter) interpretation. The key to these models is that instead of modelling the conditional expectation, they focus on the conditional probabilities for each interval or class (e.g., the probability for an observation to fall into a certain class given a set of covariates). Ordinal choice models use a priori distributional assumptions to create the mapping between the outcome variable and the explanatory variables. This can be a strong assumption, especially in the absence of any probabilistic justification. These models tend to have their names based on the assumed distribution (e.g., ordered logit uses a (standardized) logistic, while ordered probit uses (standard) normal distribution). Under the assumed distribution along with some mild conditions, the parameters are generally point-identifiable up to scaling. The main disadvantage of this approach is that it does not directly model

[1]The term *split sampling* in this paper is not related to the technique occasionally used in chromatography (Schomburg et al., 1977, Schomburg et al., 1981) or methods in machine learning, which splits the initial sample into folds.

CEU eTD Collection

the conditional mean, but rather, it provides the conditional probabilities. Therefore, the interpretation of the estimated parameters is markedly different than in a (linear) regression model. Generally, *'neither the sign nor the magnitude of the coefficient is informative [...], so the direct interpretation of the coefficients is fundamentally ambiguous.'* (Greene and Hensher, 2010, p. 142). To get meaningful interpretations, one can calculate the partial effects on the probabilities with the use of the assumed distribution. Greene and Hensher (2010) give a thorough overview of ordered choice models estimated via maximum likelihood. We should also note the case when ordered choice models are augmented with the information on the observed interval boundaries. In practice, it is common to use this additional information and incorporate it into the model, but the rather strong distributional assumption remains for point-identification. The difficulties and drawbacks of this approach are nicely summarized by Greene and Hensher (2010, p. 133).

The second solution comes from the literature of partially identified parameters (see e.g., Manski and Tamer, 2002, Manski, 2003, Tamer, 2010). This approach focuses on interval data and assigns numerical *intervals* for each observation. The main advantage of this method is that it does not require any distributional assumptions and still allows valid statistical inference on the conditional expectation. The magnitudes and the signs of the estimated parameter vector can be interpreted in the same way as the classical regression coefficients. The drawback is that the estimated parameters are not point-identified, but rather, a set is identified in which the parameter vector may belong. In other words, it only obtains a lower and upper bound for each of the unidentified point estimates. Empirical applications are rare because the estimation method is complex and, in our experience, the estimated parameter intervals are too wide for them to be empirically useful.

This paper adopts the framework of Manski and Tamer (2002) and proposes a new solution for the point identification of the parameters of a linear regression model where the dependent variable is discretized into interval data. By revisiting the discretization process, a (survey) method is put forward that collects enough information for the point-identification without any additional (e.g., distributional) assumption. Intuitively, the parameters can be point-identified when the discretization process is designed in such a way that the lower and upper bounds for each interval converge. In this case, any linear regression models can be estimated in the usual way, e.g., by least squares (LS). The resulting point-estimates are then consistent and can be interpreted as in the classical regression framework.

The above discretization does not deviate substantially from the typical methods, but it allows to obtain additional information on the distribution of the dependent variable with the use of split sampling. In the context of surveys, this means using multiple questionnaires. These questionnaires have the same set of questions but the choices (possible answers) of each questions are different. The term *split sample* refers to the fact that the sample is *split* between these questionnaires. In general, the idea is to collect the data for the same set of questions with each question containing het-

CEU eTD Collection

erogenous sets of possible outcomes in different split samples.

Furthermore, the *perception effect* or survey heterogeneity – due to the use of multiple surveys – can also be estimated through fixed effects type estimator. Another useful property of the proposed approach is that it maintains the privacy considerations through the discretization process; therefore, the data provider can safely use this method as the individuals behind the answers cannot be identified.

The paper is organized as follows: Section 3.2 introduces the identification problem and justifies the proposed split sampling approach. Section 3.3 revisits two the split sampling approach, the *magnifying* and *shifting* methods, when the discretized variable is the outcome. Section 3.3.3 derives consistent estimates via least squares, while Section 3.3.4 presents some Monte Carlo simulations. Section 3.4 extends the simple framework in two ways. First, it proposes a method to estimate and test perception effects. Second, it looks at non-linear models. Section 3.5 concludes.

3.2 Identification Problem

This section discusses the identification problems associated with the discretization of the data and justifies the split sampling approach by using the results from Manski and Tamer (2002).

Consider $y_i \sim f(a_l, a_u)$ an i.i.d. random variable, where $f(a_l, a_u)$ denotes the parent probability density function (pdf) with support $[a_l, a_u]$, where $a_l, a_u \in \mathbb{R}$, $a_l < a_u$ and $i = 1, \ldots, N$. We assume that $f(\cdot)$ is unknown and can be continuous, discrete or mixed. Instead of observing the outcomes of y_i, we observe y_i^* through a discretization process:

$$y_i^* = \begin{cases} z_1 & \text{if } c_0 \leq y_i < c_1 \quad \text{or} \quad y_i \in C_1 = [c_0, c_1) \quad \text{1st choice} \\ z_2 & \text{if } c_1 \leq y_i < c_2 \quad \text{or} \quad y_i \in C_2 = [c_1, c_2) \\ \vdots & \vdots \\ z_m & \text{if } c_{m-1} \leq y_i < c_m \quad \text{or} \quad y_i \in C_m = [c_{m-1}, c_m) \\ \vdots & \vdots \\ z_M & \text{if } c_{M-1} \leq y_i < c_M \quad \text{or} \quad y_i \in C_M = [c_{M-1}, c_M) \\ & \qquad\qquad\qquad\qquad \text{last choice,} \end{cases} \tag{3.1}$$

where, $z_m \in C_m$, $m = 1, \ldots, M$ is the assigned value for each choice. It can be a measure of centrality (e.g., mid-point), or an arbitrarily assigned value within its interval. M denotes the number of choices, which is known. For simplicity, we refer to each choice or choice interval as a class.

We examine identification of $\mathbb{E}[y|x]$ when

$$\mathbb{E}[y|x] = g(x; \beta), \tag{3.2}$$

where $g(\cdot)$ is a known function, β is a parameter vector belonging to a subset of a

CEU eTD Collection

compact finite-dimensional space ($\mathcal{B}$), and x denotes the vector of covariates.

Observing y_i^* instead of y_i leads to an identification problem. Following Manski and Tamer (2002) and Lewbel (2019), we show the conditions for the partial or set identification of β, which then point to the cases when point-identification is possible. Let $\underline{y_i^*}$ and $\bar{y}_i^*$ denote the random variables that take the lower and upper bounds as the choice value in a given interval, respectively. In other words, $z_m = c_{m-1}, m = 1, \ldots, M$ for $\underline{y_i^*}$ and $z_m = c_m, m = 1, \ldots, M$ for $\bar{y}_i^*$. By the design of the discretization, the unobserved values lie between these lower and upper bounds, $\underline{y_i^*} \leq y_i \leq \bar{y}_i^*$, $\forall i$. Furthermore, as these refer to the same random variable, the *unknown* conditional probabilities are the same $\Pr\left[\underline{y}^* \in C_m|x\right] = \Pr\left[y \in C_m|x\right] = \Pr\left[\bar{y}^* \in C_m|x\right]$, $\forall m$. Using the law of total expectation, it is easy to show that

$$\mathbb{E}\left[y|x\right] = \sum_m \left[\int_{c_{m-1}}^{c_m} y \Pr\left[y|x\right] dy\right] \Pr\left[y \in C_m|x\right], \tag{3.3}$$

where $\mathbb{E}(y \in C_m|x) = \int_{c_{m-1}}^{c_m} y \Pr\left[y|x\right] dy$, and by design $c_{m-1} \leq \mathbb{E}(y \in C_m|x) \leq c_m$, $\forall m$. Now, for the conditional expectations we get,[2]

$$\mathbb{E}\left[\underline{y}^*|x\right] \leq \mathbb{E}\left[y|x\right] \leq \mathbb{E}\left[\bar{y}^*|x\right]. \tag{3.4}$$

This bound reduces to a point in the limit when y_i is measured (or observed) precisely. However if y_i is only observed through an interval, we have a set of conditional expectations, which leads to the set identification for β. That is, under Equation 3.2, any $\mathrm{b} \in \mathcal{B}$ that satisfies $\mathbb{E}\left[\underline{y}^*|x\right] \leq g(x;\mathrm{b}) \leq \mathbb{E}\left[\bar{y}^*|x\right]$ is said to be observationally equivalent to β.[3]

Remark 1: β cannot be point-identified when $\mathbb{E}\left[\underline{y}^*|x\right] < \mathbb{E}\left[\bar{y}^*|x\right]$ or $\Pr\left[y|x\right]$ is unknown.

Remark 2: If the density of the conditional probability $(\Pr\left[y|x\right])$ is known, β is point-identified, which leads to the special cases of ordered choice models.

Following Manski and Tamer (2002), point identification of β can be achieved by using the equality condition in Equation (3.4) and by reconstructing the conditional probability of y_i. We maintain the assumption that y_i cannot be directly observed only through a limited number of choices/classes and we are not making any further (distributional) assumptions. The key to our approach is the use of split sampling, which uses different thresholds for the choices in each split sample. As we increase the number of split samples we achieve point-identification of $\mathbb{E}\left[y|x\right]$, and thus β. This split sampling method can be viewed as a non-parametric estimator on $\Pr\left[y|x\right]$.
Split sampling method works in two ways:

[2]The same result can be found in Manski (1989) or Manski and Tamer (2002).

[3]See more on the terminology of observational equivalence in Chesher and Rosen (2017) or Lewbel (2019).

CEU eTD Collection

1. The bounds are made narrower as we increase the number of split samples ($c_m - c_{m-1}$) $\to 0$. This leads to $\mathbb{E}\left[\underline{y_i^*}|x\right] \to \mathbb{E}\left[y_i|x\right]$ and $\mathbb{E}\left[\bar{y}_i^*|x\right] \to \mathbb{E}\left[y_i|x\right]$, without the need of $\underline{y_i^*} \to y_i$ and $\bar{y}_i^* \to y_i$.

2. It gives a better mapping of $\Pr\left[y|x\right]$ as we increase the number of different questions, therefore provides additional knowledge on $\mathbb{E}\left[y|x\right]$

Let us mention the implementations of the two other possible solutions. In set identification, Manski and Tamer (2002) propose a modified minimum-distance estimator, where the lower and upper bounds on the conditional expectation are also estimated along with the parameter set. Moment (in)equality models generalize Manski and Tamer (2002) for cases where there are multiple equations and/or inequalities (i.e., Chernozhukov et al. (2007) or Andrews and Soares (2010)). Beresteanu and Molinari (2008) shows asymptotic properties of such partially identified parameters. Imbens and Manski (2004), Chernozhukov et al. (2007) and Kaido et al. (2019), among others, derive confidence intervals for these set identified parameters. These methods are feasible ways to estimate parameter sets for a given conditional expectation function without any further assumption. However, these methods do not produce point-estimates for β, only estimated lower and upper bounds.
On the other hand, ordered choice models point-identifies β up to a scale, by a particular distributional assumption on $F(c_m - \beta' x) = \Pr\left[y_i^* \leq m \mid x\right]$. Here $F(\cdot)$ is the assumed cumulative distribution function, usually Gaussian or logit and m is the m'th interval value (mid point or arbitrarily chosen ordinal value). β is identified through the assumption on $F(\cdot)$, which creates the mapping between the conditional probabilities and x. Point-identification of β up to a scale means that sample estimator(s) for β is dependent on the distributional assumption, yielding different values for different distributions.
Ordered choice models aims to produce *conditional probabilities* rather than a proper interpretation for coefficients. Although there are many papers which tries to interpret the resulting parameters, which encourages us to emphasise the following properties of these models:

1. The interpretation of the parameters in an ordered choice model are different from the models where identification is based on the conditional expectation function.

2. This approach can be used to deal with interval and ordered variables as well, however these models (by default) handle the interval data as ordinal data.

3. If $F(\cdot)$ is wrongly assumed, parameter estimates will not provide consistent estimators of β.

4. In the case of the interval variable, it is possible to use the information on the intervals and fix the c_m parameters in the model. If one is interested in the conditional expectation, it is possible to use an EM algorithm to compute the maximum

CEU eTD Collection

likelihood estimator for β along with estimators for the expected values based on the assumed distribution. (Greene and Hensher, 2010, p. 133) We are going to refer to this method as 'interval regression' in our comparison study.

3.3 The Split Sampling Approach

The split sampling approach investigates useful discretization processes so that the data collected can be used to estimate the conditional expectation by using simple least squares regression techniques. The main idea is to create different questionnaires by using choices with different boundaries in each question, while fixing the number of choices (M). The term 'split sample' is referred to the fact that while the questions in each of these questionnaires are the same, the boundaries on their choices are different and therefore each questionnaire will have its own *split sample*. Due to human cognitive capacities, usually, a very limited number of choices is the only feasible way to construct such questionnaires.[4] The use of S split samples enables us to collect the answer of the same question in S different ways, which eliminates the discretization problem. We achieve this through changing the class boundaries (c_m) between each split sample.

The intuition behind the approach is that this leads to a better mapping of the unknown distribution of y and, in principle, to a complete mapping of the focus model. By merging the different split samples into one data set, calls *working sample*, we get $b = 1, \ldots, B$ overall number of choice classes across the merged split samples, where B is much larger than M. In a given split sample, each respondent (i) is given one questionnaire. The set of respondents who fill in a questionnaire with the same class boundaries defines a split sample. Each split sample has $N^{(s)}$ number of observations ($s = 1, \ldots, S$ and $\sum_s N^{(s)} = N$). In this setup, the discretization of a split sample looks exactly as the problem introduced above in Equation (3.1). The only difference across split samples is that the class boundaries are different. Note that the number of observations across split samples can be the same or, more likely, different. Now a split sample is as follows,

$$y_i^{(s)} = \begin{cases} z_1^{(s)} & \text{if } y_i \in C_1^{(s)} = [c_0^{(s)}, c_1^{(s)}), \\ & \qquad \text{1st choice for split sample } s, \\ z_2^{(s)} & \text{if } y_i \in C_2^{(s)} = [c_1^{(s)}, c_2^{(s)}), \\ \vdots & \vdots \\ z_m^{(s)} & \text{if } y_i \in C_m^{(s)} = [c_{m-1}^{(s)}, c_m^{(s)}), \\ \vdots & \vdots \\ z_M^{(s)} & \text{if } y_i \in C_M^{(s)} = [c_{M-1}^{(s)}, c_M^{(s)}], \\ & \qquad \text{last choice for split sample } s. \end{cases} \tag{3.5}$$

[4] Typically, the optimal number of choices for a survey is relatively small, $M = 3, 5, 7$ or at most $M = 10$. There is a large literature about the optimal number of choices (or 'scale points') in a survey, see e.g., Givon and Shapira (1984), Srinivasan and Basu (1989) or Alwin (1992).

CEU eTD Collection

The observed values $z_m^{(s)}$ are set to a numeric value between $c_{m-1}^{(s)}$ and $c_m^{(s)}$, typically to the mid-point. In the second step, we merge all the split samples and create a 'working-sample' used to estimate the parameter(s) of interest. The working-sample is an artificial construction created in such a way that the working class boundaries (c_b^{WS}) are the union of the class boundaries of split samples.[5]

$$\bigcup_{b=0}^{B} c_b^{WS} = \bigcup_{s=1}^{S} \bigcup_{m=0}^{M} c_m^{(s)} . \tag{3.6}$$

With proper re-distribution of the observations to the working sample, we can reconstruct the underlying unobserved continuous variable's distribution.

Apart from the specifics of discretized outcomes, here we only present the main ideas, assumptions and the main theoretical results for two split sampling methods: the magnifying and shifting.

3.3.1 The Magnifying Method

The magnifying method magnifies parts of the domain in each questionnaire by one equally sized choice class. The size of classes depends on the number of split samples (S) and number of choices (M). As the number of split samples increases, class sizes decrease, which uncovers the unknown underlying distribution. Figure 3.1 shows the main idea of the magnifying method with the individual questionnaires for the case $M = 3$ and $S = 4$. The last line shows the working sample.

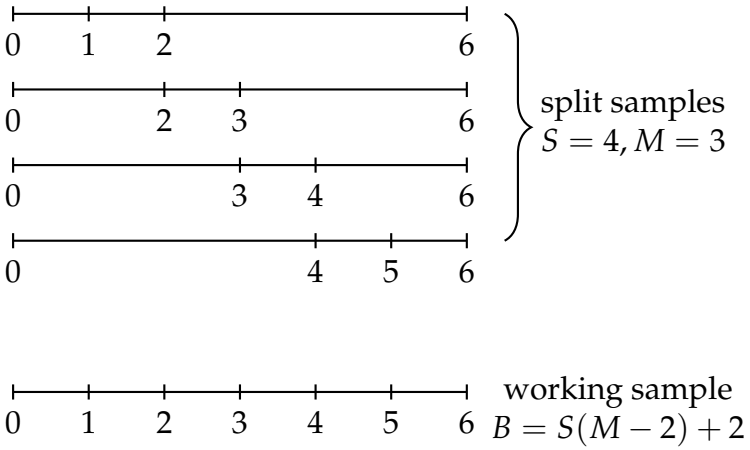


Figure 3.1: The magnifying method

The first and last split samples are slightly different from the split samples in-between. They have one extra class with the same class width, while split samples in-between

[5]Here, we discuss the cases where the domain (a_l, a_u) for y_i is known and the working sample's class boundaries maps the domain of y_i, $c_0^{WS} = a_l, c_B^{WS} = a_u$. Our proof holds for cases where (a_l, a_u) are unknown and $c_0^{WS} \neq a_l$ and/or $c_B^{WS} \neq a_u$. In these cases, one might drop observations which are outside the domain of the survey(s) e.g., $a_l < c_0^{WS}$, or there is a known censoring in the survey, $a_l = -\infty$ and/or $a_u = \infty$. Note that in these cases the sample properties (e.g., speed of convergence) can be different.

CEU eTD Collection

have $M-2$ classes with the same class width. Observations which fall into these classes are called *directly transferable observations* (DTOs). The connection between the number of magnified classes in the working sample (B), and the number of split samples (S) and choices (M) is given by

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

Given the fact that there are B classes in the working sample, we get the widths of these classes,

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

Fixing the upper and lower bounds on the domain[6] for the split samples $(a_l = c_0^{WS} = c_0^{(s)};\ a_u = c_B^{WS} = c_M^{(s)},\ \forall s)$, we can reduce the class size $h \to 0$ as $S \to \infty$, which enables us to ensure convergence in distribution. This can also be seen through the working sample's boundary points, which have the following simple form

$$c_b^{WS} = a_l + bh\,.$$

With the magnifying method we can separate two types of observations. The first is the already mentioned directly transferable observations. Formally, $y_i^{(s)} \in \zeta$, where ζ is the set of choice intervals of $\zeta = C_m^{(s)},\ \forall$ pair of $(1 < s < S, 1 < m < M)$, and $(s = 1, m = 1), (s = S, m = M)$. Here, $\lim_{S\to\infty} ||C_m^{(s)}|| = 0$, which means that at the limit we observe responses without any discretization. Moreover, these observations have the same class width as the working sample's classes and each can be directly linked to a certain working sample class, by design, hence the name '*directly transferable observations*'. These observations are denoted by $y_{i,DTO}^{WS}$, with $i = 1, \ldots, N_{DTO}^{WS}$.
To achieve point-identification of β, the first step is to have a consistent mapping of the unconditional distribution of the unknown y variable by using $y_{i,DTO}^{WS}$. To show $\lim_{S\to\infty} \Pr\left(y_{DTO}^{WS}\right) = \Pr(y)$, we consider the following assumptions.

Assumption 2. *Let y be a continuous random variable with probability density function $f(y)$ with S, N and $C_m^{(s)}$ follow the definitions above then*

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

b. All split samples will have non-zero respondents.

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

Assumption 2a. ensures that the number of respondents will always be higher than the number of split samples. Assumption 2b. provides utilisation of all split samples, i.e. each split sample will have non-zero respondents. Assumption 2c. imposes a mild

[6] In case of infinite support for the domain, one needs to set c_1^{WS} and/or c_{B-1}^{WS} into a reasonable value on the support. We discuss later the these truncated cases.

CEU eTD Collection

assumption on the underlying distribution. That is, the support of the random variable is not disjoint, which implies $\int_{c_{b-1}^{WS}}^{c_b^{WS}} f(y)\mathrm{d}y > 0$.
Remarks: we can decrease c as close to 0 as we would like to. This means that there is an equal or higher number of observations than split samples. On the other hand, we exclude by assumption the case when c $\geq$ 1, which means that there is an equal or higher number of split samples than observations. In this case, we most certainly would not observe values for each working sample class.
These assumptions allow us to claim the following proposition,

Proposition 3. *Under Assumptions 1.a - c,*

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

.

Proof of Proposition 3. is the same as the proof of Proposition 1. in Chapter 2. (See the complete proof in Appendix B.3.1.). The proposition establishes convergence in distribution which allows point-identification for the parameter of interest, discussed at Section 3.3.3.

Next, let us consider the second type of observations, which are all the other observations which fall into choice classes at the boundaries. We call them '*non–directly transferable observations*' (NDTOs) as for these observations $\lim_{S\to\infty} ||C_m^{(s)}|| = a_u - a_l$ for split sample and choice value pairs of $1 < s < S$, $m = \{1, M\}$ or $s = 1$, $m = M$ or $s = S$, $m = 1$. This means that there is no systematic reduction in the measurement error for these responses. One way to proceed is to drop them completely so they would not appear in the working sample (thus, only using $y_{i,DTO}^{WS}$ for estimation purposes). However, in practice it seems that too many could fall into this category, resulting in a large efficiency loss during the estimation.
Another approach is to use DTOs to proxy the measurement error for the NDTOs. We can utilise the information from y_{DTO}^{WS} to calculate specific interval means for the underlying distribution and use these to replace the non–directly transferable observations. The simplest way to get the estimators for these conditional means is to regress the directly transferable observations on a vector of indicator variables referring to the NDTOs' intervals. The resulting 'replacement estimators' are constructed similarly as in Chapter 2. Section 2.3.3 and has the same asymptotic properties.

The magnifying method can be seen as the simplest theoretical design for split sampling, which shows how the method works, but its use is limited in practice. It is mostly applicable when the survey design deal with small number of S, while with large number of split samples, the creation of the questions are infeasible. At last, let us remark this method does not preserve data confidentiality. It uses the fact that some individuals are correctly observed and y is i.i.d., therefore the generalization of those observations is correct.

CEU eTD Collection

3.3.2 The Shifting Method

The shifting method is an alternative to the magnifying method. It takes the original class width as given, with fixed class widths, and shifts the boundaries of each choice with a given fixed value. Increasing the split sample size does not affect the boundary widths in-between the domain, only the size of the shift. As we shift the boundaries, we add an extra class[7] for each split sample at the boundary where, due to the shift, the class width has increased. Figure 3.2 shows the split samples in this approach with $S = 4$ and with $M = 4$ classes.

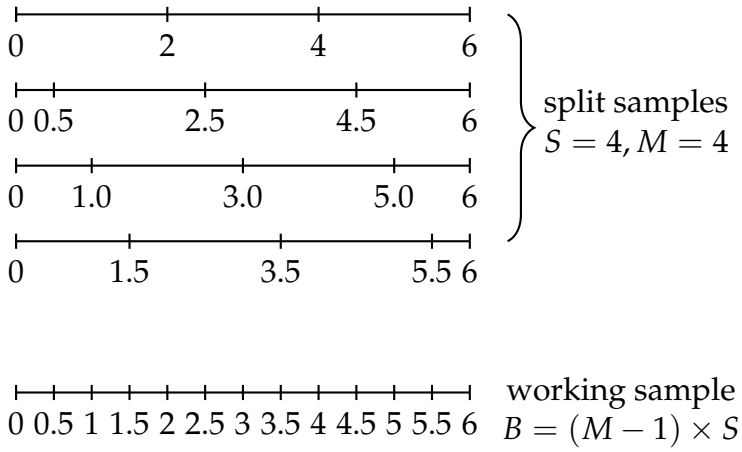


Figure 3.2: The shifting method

As Figure 3.2 shows, there is one split sample (the benchmark $s = 1$) where there is one class less, otherwise everywhere there is always M classes. The number of intervals in the working sample is

$$B = S \times (M - 1)$$

The boundary points for each split sample are

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

For the working sample, we get $c_b^{WS} = a_l + b\frac{a_u - a_l}{S(M-1)}$. Intuitively, we achieve complete mapping by reducing the shifting size to 0, thus we are be able to identify observations which lie in these small intervals, using the information available in all the other split samples.

Merging the split samples into the working sample is somewhat cumbersome, but works efficiently. The main idea is to uniformly assign each split sample's observations to the working sample's choice values, whose intervals are congruent with the split sample's class interval, creating an artificial variable $y_i^\dagger$, similarly as in Chapter 2.

[7] To be more specific, we in fact reveal a hidden class.

CEU eTD Collection

Section 2.3.4. The shifting method works similarly as the proposed artificial variable $y^\dagger$ converges in distribution to the underlying continuous variable with the following mild assumptions:

Proposition 4. *Under Assumptions 2a, 2c and that the split samples are equally distributed among respondents, i.e. the probability of an respondent responds to a particular split sample is* $1/S$ *then*

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

Proof of Proposition 4. is the same as the proof of Proposition 2. in Chapter 2.
In addition, with shifting method we can investigate the speed of convergence, as we increase the number of split samples (S). The main result from the exercise is that on the boundaries of the support[8], the method converges slower, with $\frac{\log S}{S}$, while for the rest it converges with $1/S$.

A caveat is we cannot directly use $y_i^\dagger$ for estimation, while by design each individual observation only represents the conditional mean for the given split sample's class, and not the underlying variable's conditional expectation. However, while $\lim_{S\to\infty} F_S(y^\dagger) = F(y)$, we can use these values to calculate the specific conditional means. Here, we departure from methods used in Chapter 2, and tailor the conditioning set specific to our problem.

Let $\hat{\pi}_\tau$ be the estimator vector for each conditional mean dependent on the partitions D_l with $l = 1, \ldots, L$ mutually exclusive partitions of the conditioning variable(s) x_i's.[9]
Let us define

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

By the weak law of large numbers, $\hat{\pi}_\tau$ is converging to the true underlying distribution's conditional expectations,

$$\hat{\pi}_\tau \to \mathbb{E}(y | x \in D_l)$$

as $N, S \to \infty$ under the same assumptions as before. Next, let us use a simple regression specification, which helps us to show the asymptotic properties of the estimator.

$$y_i^\dagger = \boldsymbol{\pi}_\tau \mathbf{1}_{\{x_i \in D_l\}} + \eta_i,$$

[8]Which is given by the maximum distance from the support given by the split samples. For the lower bound: $c_1^{(1)} + (c_2^{(S)} - c_1^{(1)})$ and for the higher bound: $c_M^{(1)} + (c_{M-1}^{(1)} - c_M^{(1)})$.

[9]Note that L is a fixed number by the researcher, asymptotically $\frac{L}{N} \to 0$ is always true, thus the number of observations in these partitions are asymptotically increasing as well. However in finite samples L should be chosen such that there are enough observations in each conditional set.

CEU eTD Collection

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

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

Under the standard LS assumptions, we can write

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

where $\mathbb{E}\left(\boldsymbol{\pi}_\tau\right) = \mathbb{E}(y|x \in D_l)\,\forall\tau$ and N^{WS} is the number of observations in the working sample. Furthermore, the variance of the LS estimator is given by

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

Algorithm 7 describes how to create the a working sample with the shifting method, from the artificial variable $y_i^\dagger$

Algorithm 7 The shifting method – creation of working sample

1: Set $s := 1, m := 1, y_i^{WS} = \emptyset$.
2: Calculate the sample conditional mean $\hat{\pi}_\tau$ using $y_i^\dagger$ conditioning on D_l class. D_l denotes a set containing L mutually exclusive partitions of the domain of x_i's.
3: Add the conditional mean $\hat{\pi}_\tau$ and the observed values to the working sample,

$$y_i^{WS} := \left\{ y_i^{WS}, \bigcup_{j=1}^{N} \hat{\pi}_\tau \right\}$$

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

We need to track the individual observations to be able to pair them with the right-hand side variables. Note that this pairing only applies to the conditional expected values not to the actual (un)observed value.
Some remarks: 1) The shifting method enables a more flexible survey design in practice, while the choice class widths are approximately the same. 2) The shifting method ensures data privacy considerations: creating artificial observations, and calculating their conditional averages given the covariates will make the individuals' real value intractable.[10]

[10]One needs to pay special attention to the responses near the support $(C_1^{(s)}, C_M^{(s)})$ of the questionnaire as those can reveal some information about the respondent. In the limit $S \to \infty$ $C_2^{(1)}$ and $C_M^{(S)}$ identifies the respondents, however this is rather a theoretical case. This problem does not emerges if the support of y is infinite or these observations are dropped.

CEU eTD Collection

3.3.3 OLS Estimation

The proposed split sampling methods lead to two possible ways to obtain a consistent estimate of β via the least squares estimator. The first approach uses only the DTOs, while the second relies on all observations.

LS Estimation Based on DTOs

Let N^D denote the number of DTOs and let

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim iid\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}\right), \tag{3.7}$$

$$= \hat{\mathbf{y}} + \boldsymbol{\varepsilon}, \tag{3.8}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{N^D})'$. We make the following assumptions:

Assumption 3. *Let the data generating process follows equation (3.7)*

a. $\mathbb{E}[|\mathbf{y}|]$ *exists and* $\mathbb{E}[\mathbf{X}\boldsymbol{\varepsilon}] = 0$.

b. There exists a bounded matrix $\mathbf{Q}$ *such that* $\frac{\mathbf{X}'\mathbf{X}}{N^D} - \mathbf{Q} = o_p(1)$.

c. $\forall \xi > 0\ \exists S$ *such that* $\Pr(|\varepsilon_i| > ||C_m||) = 1 - \xi$ *given* $\hat{y}_i \in C_m$.

The first two assumptions are common for LS estimator. Assumption 3c implies that ξ can be made arbitrarily small by choosing an appropriate S. This is due to the fact that $||C_m|| \to 0$ as $S \to \infty$.

Consider that the discretized version of $\mathbf{y}$ denotes $\mathbf{y}^*$ and write $\mathbf{y}^* = \mathbf{y} + \mathbf{u}$, where $\mathbf{u}$ represents the 'measurement errors' due to discretization. Consider the set

$$\mathbf{A} = \{i : |\varepsilon_i| > ||C_m|| \wedge \hat{y}_i \in C_m\}. \tag{3.9}$$

Set $\mathbf{A}$ allows us to distinguish between two sets of DTOs. Those belonging to $\mathbf{A}$ when the unobserved value, y_i, is in a different class than its corresponding $\hat{y}_i$. In this case, $|\varepsilon_i| > ||C_m||$ given $\hat{y}_i \in C_m$, which implies $x_i \perp u_i$. Those not belonging to $\mathbf{A}$ when both y_i and $\hat{y}_i$ belong to the same class, and therefore $Cov(x_i, u_i) \neq 0$. As we shall demonstrate below, these two sets of observations affect the properties of OLS differently. Partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ such that $\mathbf{X}_1$ contains those observations whose indexes belong to $\mathbf{A}$ and $\mathbf{X}_2$ contains observations whose indexes do not belong to $\mathbf{A}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^* \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_1'\mathbf{u} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_2'\mathbf{u} \\
&= \boldsymbol{\beta} + o_p(1) + \xi\left(\frac{\mathbf{X}'\mathbf{X}}{N^D}\right)^{-1}\frac{\mathbf{X}_2'\mathbf{u}}{\xi N^D} \qquad (3.10) \\
&= \boldsymbol{\beta} + o_p(1) + \xi O_p(1). \qquad (3.11)
\end{aligned}$$

CEU eTD Collection

The second last line follows from the fact that $X_1 \perp \mathbf{u}$. Under Assumption 3c, as $||C_m|| \to 0$ for all m, $\xi \to 0$ and $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1)$.
The argument above relies on the assumption that the number of DTOs approaches infinity. In order to show that this can be the case, at least theoretically, we need to derive the relation between S and N^D. Let $\mathbf{B}$ denote the set of DTOs. From the proof of Proposition 3 we have,

$$\begin{aligned}\Pr(y \in \mathbf{B}) =& \frac{1}{S}\left[\sum_{m=2}^{M}\Pr(y \in C_m^{(S)}) + \sum_{m=1}^{M-1}\Pr(y \in C_m^{(1)}) + \sum_{m=2}^{M-1}\sum_{s=2}^{S-1}\Pr(y \in C_m^{(s)})\right] \\ \leq& \frac{1}{S}.\end{aligned}$$

Since y can only belong to one and only one class, all the events on the right-hand side are mutually exclusive and their sum must be less than or equal to 1. Note that $N^D = N\Pr(y \in \mathbf{B})$, and hence

$$N^D = O_p(\frac{S}{N}). \tag{3.12}$$

$\hat{\boldsymbol{\beta}}$ from the Conditional Expectation

The method above relies only on the DTOs. However, it is also possible to use all observations for the purpose of estimation. Consider the following standard regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim iid\left(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}\right), \tag{3.13}$$

where $\mathbf{y} = (y_1, \ldots, y_N)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ with $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN})'$, $i = 1, \ldots, k$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$. Let $\mathbf{D} = \{\mathbf{D}_1, \ldots, \mathbf{D}_L\}$ denote a set containing L mutually exclusive partitions of the domain of $\mathbf{X}$. Then

$$\mathbb{E}(\mathbf{y}|\mathbf{X} \in \mathbf{D}_l) = \mathbb{E}(\mathbf{X}|\mathbf{X} \in \mathbf{D}_l)\boldsymbol{\beta} \qquad l = 1, \ldots, L. \tag{3.14}$$

Let $\tilde{\mathbf{y}}_l$ and $\tilde{\mathbf{X}}_l$ denote consistent estimates of $\mathbb{E}(\mathbf{y}|\mathbf{X} \in \mathbf{D}_l)$ and $\mathbb{E}(\mathbf{X}|\mathbf{X} \in \mathbf{D}_l)$, respectively, $l = 1, \ldots, L$. Following from equation (3.14), we get

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u}, \tag{3.15}$$

where $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1 \ldots, \tilde{\mathbf{y}}_L)$, $\tilde{\mathbf{X}} = \left(\tilde{\mathbf{X}}_1', \ldots, \tilde{\mathbf{X}}_L'\right)'$ and $\mathbf{u} = (u_1, \ldots, u_L)'$. Note that $\mathbb{E}(u_l) = 0$ for all l since $\tilde{y}_l$ and $\tilde{\mathbf{X}}$ are consistent estimates. Moreover, $\mathbb{E}(u_l u_g) = 0$ for $l \neq g$ due to D_l are mutually exclusive $\forall l$ and $\mathbb{E}(u_l|\tilde{\mathbf{X}}_l) = 0$ since the partition does not affect the sampling error. Furthermore, assume $\mathbb{E}[\tilde{\mathbf{y}}]$ exists and $\mathbb{E}\left[\tilde{\mathbf{X}}\mathbf{u}\right] = 0$. Let $\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$, then under the usual argument of the classical OLS, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(1)$.
The basic idea here is to obtain the averages of y_i conditional on different ranges of x_i. Such partitions preserve the correlation structure between y_i and x_i, as demonstrated above.

CEU eTD Collection

As we have shown that split sampling lead to $y_i^\dagger \xrightarrow{d} y_i$. Thus,

$$N_l^{-1} \sum_{\mathbf{x}_i \in \mathbf{D}_l} y_i^\dagger - \mathbb{E}\left(y_i | \mathbf{X} \in \mathbf{D}_l\right) = o_p(1). \tag{3.16}$$

Given this is a consistent estimator of $\mathbb{E}\left(y_i | \mathbf{X} \in \mathbf{D}_l\right)$, the above result holds and the consistency based on conditional expectation follows. It is worthwhile to point out that the above argument applies to any split sampling method that leads to convergence in distribution to the underlying dependent variable. Thus, its applicability goes beyond the shifting and magnifying methods.

3.3.4 Monte Carlo Evidence

For the simulation experiments, we consider the following data generating process

$$y_i = x_i'\beta + \epsilon_i$$

and set $\beta = 0.5$. We focus on the case where the support of y_i is known, thus any error (bias) may come only from the discretization. We set the lower bound as $a_l = -2$, and the upper bound as $a_u = 4$. (We have experimented with different boundaries; the details are in the Appendix C.1.) The exogenous variable x_i is generated by a normal distribution, and to ensure the support of y_i, it is truncated at -1 and 1, with a variance of 0.25.[11] Our main concern is the disturbance term (ϵ_i), therefore we have visited several common types of distributions. To ensure the support of y_i is being met, we truncated/set ϵ_i such that it lies between $-1 \leq \epsilon_i \leq 3$.[12] We experiment with the following distributions:

- Normal: Standard normal distribution truncated at -1 and 3.
- Logistic: Standard logistic distribution truncated at -1 and 3.
- Log-Normal: Standard log-normal distribution truncated at 4 and subtracted 1 (in order to adjust the mean).
- Uniform: Uniform distribution between -1 and 3.
- Exponential: Exponential distribution with rate parameter 0.5, truncated at 4 and subtracted 1.
- Weibull: Weibull distribution with shape parameter 1.5 and scale parameter 1, truncated at 4 and subtracted 1.

[11] This choice of variance ensures that even without truncation, 95% of the probability mass lies between -1 and 1.

[12] This creates an asymmetric distribution for ϵ_i in several cases, which favours distribution independent estimation methods rather than the maximum likelihood. In the online appendix, we show results with boundaries where ϵ_i was truncated in a symmetric way. The results and conclusions remain unchanged.

CEU eTD Collection

In the event that the distributions do not have a zero mean, we specified the conditional mean as $y_i = \alpha + x_i'\beta + \eta_i$, where $\epsilon_i = \alpha + \eta_i$ with $\mathbb{E}(\epsilon_i) = \alpha$.

For the discretization of y_i we use $M = 5$, $c_0 = a_l = -2$, $c_M = a_u = 4$ and equal distances for the thresholds between the boundaries.
To estimate β, we have used the following methods:

- *Set identification*: Estimates the lower and upper boundaries of the parameter set for β using y_i^* as interval data. Estimation is based on Beresteanu and Molinari (2008) and their published Stata package (Beresteanu et al., 2010)[13]. This method does not produces point-estimates for β, only lower and upper boundaries.

- *Ordered probit and logit*: Ordered choice models, where y_i^* values are ordinal data, and the model assumes a gaussian or logistic distribution (Greene and Hensher, 2010). The estimated maximum likelihood 'naive' parameters reported here are not designed to recover β and to be interpreted in the linear regression sense. Therefore, we call the difference from β *distortion* instead of bias. However, we find it important to report these values as (unfortunately) they are the most used and (mis-)interpreted estimates in applied work.

- *Interval regression*: A modification of the ordered choice model, where y_i^* values are interval data and the model assumes gaussian distribution in order to model the linear regression model. The maximum likelihood parameter estimates aim to recover β through the distributional assumption. For a detailed description, see Cameron and Trivedi (2010, p. 548-550) or Greene and Hensher (2010, p. 133).

- *Midpoint regression*: A simple linear regression using midpoints for y_i^* and OLS for estimation.

- *Magnifying*: The magnifying method with $S = 10$ split samples. We use only DTO observations.[14]

- *Shifting*: The shifting method with $S = 10$, all outcome observations are used and created as described in Algorithm 7.[14]

We have included an intercept wherever possible.[15] Finally, we used $N = 10,000$ observations and $1,000$ Monte Carlo repetitions. We report the Monte Carlo average bias or distortion from the true parameter along with the Monte Carlo standard deviation.

[13] https://molinari.economics.cornell.edu/programs.html

[14] We used mid-values as observations for the split samples' $(y_i^{(s)})$ and working sample's choice value. L is set to 50 equal distance partitions for x_i, where the conditioning was needed.

[15] Ordered choice models' implementation in Stata remove (restricts to zero) the intercept parameter to identify β.

CEU eTD Collection

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.1, 1.15]$ (0.02),(0.02)	$[-1.09, 1.15]$ (0.03),(0.03)	$[-1.09, 1.16]$ (0.02),(0.02)	$[-1.07, 1.17]$ (0.03),(0.03)	$[-1.06, 1.19]$ (0.03),(0.03)	$[-1.09, 1.15]$ (0.02),(0.02)
Ordered probit*	0.1971 (0.0256)	0.0688 (0.0253)	0.2085 (0.0262)	0.0158 (0.0234)	0.0986 (0.0241)	0.4461 (0.0295)
Ordered logit*	0.6509 (0.0464)	0.3814 (0.0455)	0.6862 (0.0499)	0.2379 (0.0422)	0.4338 (0.044)	1.2085 (0.0546)
Interval regression	0.0268 (0.0198)	0.0332 (0.0249)	0.0371 (0.0221)	0.0491 (0.0271)	0.0663 (0.0249)	0.0397 (0.0166)
Midpoint regression	0.0253 (0.0195)	0.0322 (0.0236)	0.0362 (0.0216)	0.0490 (0.0273)	0.2077 (0.0128)	0.0314 (0.0157)
Magnifying $(S = 10)$	-0.0060 (0.0515)	-0.0205 (0.0674)	-0.0072 (0.0616)	-0.0332 (0.0781)	0.0213 (0.0333)	0.0066 (0.0417)
Shifting $(S = 10)$	-0.0017 (0.0204)	0.0004 (0.0243)	-0.0012 (0.0215)	0.0010 (0.0269)	-0.0001 (0.0127)	-0.0001 (0.0149)

†: Set identification gives 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.
*: 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.

Table 3.1: Monte Carlo average bias and standard deviation

Table 3.1 shows the results. The shifting method consistently provides the smallest average bias, while the magnifying method also outperforms the other procedures in general. Set identification gives such large intervals for the parameter set that it is unlikely to be useful in practice. Distortions of ordered probit and logit models are rather large. Interval regression and midpoint regressions perform poorly in the sense that both methods result in large biases. The Monte Carlo standard deviation is similar for all cases except for the magnifying method. This is due to the fact that the magnifying 'only DTO' method uses fewer observations, for the estimation, in our case $N/S \approx 1,000$ observations.

We have run several other Monte Carlo experiments to investigate the finite sample properties of our methods. With moderate sample size $(N = 1,000)$, the results are similar: the shifting and magnifying methods outperform all alternatives. The magnifying method performs slightly more poorly in smaller samples, while the effective number of observations in this case is only around 100. Interestingly, in both the exponential and weibull setups, the magnifying method gives similar results as those from the interval and midpoint regressions. Naturally, if the distribution is well specified, methods with maximum likelihood estimation (ordered probit, logit or interval regression methods) produce even smaller biases. However, for the other (miss-specified) cases our split sampling methods work much better. For a smaller number of choices, $M = 3$, the biases are generally worse but the differences between the methods are similar. The shifting method still performs better, while the magnifying method still outperforms the alternatives in most cases. This suggests that our methods are robust to the underlying distributions.
Finally, we chose ϵ_i as a truncated standard normal and checked what happens if we

CEU eTD Collection

increase the number of observations and the number of split samples. The simulation results – see Appendix C.1 – give evidence on the consistency of the estimator based on our split sampling approach. By contrast, for all the other alternative methods the same magnitude of bias remained as we increased the number of observations. This suggests that alternative methods provide not only biased but also inconsistent estimates for β in N. For a detailed discussion of these results see the Appendix C.1.

3.4 Extensions

3.4.1 Perception Effect

There is some evidence in the behavioural literature that the answers to a question may depend on the way the question is asked (see, e.g., Diamond and Hausman (1994), Haisley et al. (2008) and Fox and Rottenstreich (2003)).[16] Let us call this the *perception effect*. This is present regardless 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.

Let S be the total number of split samples and define two sets of discretization of y_i namely,

$$y_i^* = \begin{cases} z_1 & \text{if} \quad c_0 < y_i < c_m \\ & \vdots \\ z_m & \text{if} \quad c_{m-1} < y_i < c_M \end{cases} \tag{3.17}$$

and

$$y^{**} = \begin{cases} z_1 & \text{if} \quad c_0 < y_i + B_s < c_1 \\ & \vdots \\ z_m & \text{if} \quad c_{m-1} < y_i + B_s < c_M, \end{cases} \tag{3.18}$$

where B_s denotes the perception effect for split sample s, $s = 1, \ldots, S$. Let $\tilde{y}_i^*$ and $\tilde{y}_i^{**}$ denote the observations in the working sample that derived from y_i^* and y_i^{**}, respectively. Following the construction of the working sample, it is straightforward to show that

$$\tilde{y}^{**} = \tilde{y}_i^* + B_s \tag{3.19}$$

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

$$\tilde{y}_i^{**} = \beta x_i + u_i \tag{3.20}$$

is equivalent to

$$\tilde{y}_i^* + B_s = \beta x_i + u_i. \tag{3.21}$$

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

CEU eTD Collection

Writing the above in matrix form using the normal definition gives

$$\tilde{\mathbf{y}}^* + \mathbf{DB} = \mathbf{x}\beta + \mathbf{u}, \tag{3.22}$$

where $\mathbf{B} = (B_1, \ldots, B_S)'$ and $\mathbf{D}$ is a $N \times S$ zero-one matrix that extracts the appropriate elements from $\mathbf{B}$. So the estimation of β can be done in the spirit of a fixed effect estimator. Define the usual residual maker, $\mathbf{M_D} = \mathbf{I}_N - \mathbf{D}\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'$, then

$$\hat{\beta} = \left(\mathbf{x}'\mathbf{M_D}\mathbf{x}\right)^{-1}\mathbf{x}'\mathbf{M_D}\tilde{\mathbf{y}}^{**} \tag{3.23}$$

is a consistent estimator of β given the results presented in this paper and the similar argument for the consistency of standard fixed effect estimator in the panel data literature.

We also need to modify the estimator for $\mathbb{E}\left(y_i|\mathbf{X} \in \mathbf{D}_l\right)$ slightly in order for the above to hold for the $\hat{\beta}$ based on conditional expectation. The main problem is to keep track of the perception effect. This means we need to keep track of which split sample each observation comes from when estimating the conditional averages. Specifically,

$$N_l^{-1} \sum_{\mathbf{x}_i \in D_l, \mathbf{x}_i \in s} \tilde{y}_i^{**} - \mathbb{E}\left(y_i|\mathbf{X} \in \mathbf{D}_l\right) + B_s = o_p(1). \tag{3.24}$$

While the above discussion focuses on one regressor, extension to K regressors is straightforward and requires no additional assumptions on B_s. This approach can also be extended to include interacting class and split sample effects, such as B_{sm} for $s = 1, \ldots, S$ and $m = 1, \ldots, M$, which hopefully would take care of all likely perception effects.

It is theoretically possible to test the impacts of the perception effects on the estimator. Since $\hat{\beta}$ as defined in equation (3.23) is consistent regardless of the presence of perception effects and

$$\tilde{\beta} = \left(\mathbf{x}'\mathbf{x}\right)^{-1}\mathbf{x}'\tilde{y}^{**} \tag{3.25}$$

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

$$\left(\hat{\beta} - \tilde{\beta}\right)'\left[\mathrm{Var}\left(\hat{\beta} - \tilde{\beta}\right)\right]^{-1}\left(\hat{\beta} - \tilde{\beta}\right) \overset{a}{\sim} \chi^2(K). \tag{3.26}$$

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 to the case where $\mathbf{B}$ is a random vector. It would also appear that some assumptions on $\mathbf{B}$ are required in order to compute the test statistics. This is another interesting avenue for future research.

CEU eTD Collection

3.4.2 Non-linear Models

Another possible extension is to consider the application of the proposed methods in the context of non-linear models. Given the presented methods focus on data collection, they could also be applied to non-linear models. To see this, consider

$$y_i = g(x_i; \beta) + u_i \tag{3.27}$$

where $g(\cdot)$ denotes a continuous function. Let $\mathbf{y}$, $\mathbf{y}^{WS}$ and $\mathbf{x}$ be the data matrix of y_i (if we do observe it), y_i^{WS} (observations from the working sample, and x_i, respectively). Let $\hat{\beta}(\mathbf{y}, \mathbf{x})$ denotes a consistent estimator of β with $\rho(\mathbf{x}) = \sqrt{N}\left[\hat{\beta}(\mathbf{y}, \mathbf{x}) - \beta\right]$ such that $\rho(\mathbf{y}, \mathbf{x}) \xrightarrow{d} f(0, \Omega)$. Under the assumptions made earlier, $y_i^{WS} \xrightarrow{d} y_i$, and therefore $\rho(\mathbf{y}^{WS}, \mathbf{x}) \xrightarrow{d} \rho(\mathbf{y}, \mathbf{x})$ by the continuous mapping theorem under appropriate regularity conditions. The technical details of these conditions, however, could be an interesting subject of future research.

3.5 Conclusion

This paper deals with econometric models where the dependent variable is continuous but observed through a discretization process that results in interval data. When such a variable is modelled in a (linear) regression framework, the regression parameter(s) cannot be point-identified.

Ordered choice models – which are most commonly used to treat such outcome variables – rely on distributional assumptions for point-identification. Alternatively, Manski and Tamer (2002) offer set identifying conditions, which results in large ranges of estimated parameter intervals.

Our proposed split sampling approach does not rely on any distributional assumption and does not restrict the validity to set-identification. Instead, we propose changes in the data gathering process (the way data is collected). We show that parameters can be point-identified and estimated consistently in a regression model. The split sampling approach put forward ensures that the least squares estimator is unbiased and consistent, and it also works well in moderate sample sizes.

The two split sampling methods put forward – magnifying and shifting methods – may guide survey designers and researchers who deal with questionnaires, as well as data providers. With the shifting method, data providers can also take into account data privacy considerations: individuals are not identifiable from the data, however, the data can still be used to simply estimate the parameters of interest.

CEU eTD Collection

Bibliography

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

Alwin, Duane F, "Information transmission in the survey interview: Number of response categories and the reliability of attitude measurement," *Sociological Methodology*, 1992, pp. 83–118.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484), 1481–1495.

Andrews, Donald WK and Gustavo Soares, "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 2010, *78* (1), 119–157.

Athey, Susan and Guido Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.

_ **and Guido W Imbens**, "Machine learning methods for estimating heterogeneous causal effects," *Stat*, 2015, *1050* (5).

_ **, Julie Tibshirani, Stefan Wager et al.**, "Generalized random forests," *Annals of Statistics*, 2019, *47* (2), 1148–1178.

Bargagli, Stoffi and Giorgio Gnecco, "Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms," *International Journal of Data Science and Analytics*, 2020, *9*, 315–337.

Becker, Sascha O., Peter H. Egger, and Maximilian von Ehrlich, "Absorptive Capacity and the Growth and Investment Effects of Regional Transfers: A Regression Discontinuity Design with Heterogeneous Treatment Effects," *American Economic Journal: Economic Policy*, 2013, *5* (4), 29–77.

Beresteanu, Arie and Francesca Molinari, "Asymptotic properties for a class of partially identified models," *Econometrica*, 2008, *76* (4), 763–814.

_ **,** _ **, and Morris Darcy**, *Asymptotics for partially identified models in Stata* 2010.

CEU eTD Collection

Berkson, J., "Minimum Chi-square, not Maximum Likelihood!," *The Annals of Statistics*, 1980, *8*, 457–487.

Bhat, Chandra R, "Imputing a continuous income variable from grouped and missing income observations," *Economics Letters*, 1994, *46* (4), 311–319.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen, *Classification and regression trees*, CRC press, 1984.

Buonaccorsi, John P, *Measurement error: models, methods, and applications*, CRC Press, 2010.

Caetano, Carolina, Gregorio Caetano, and Juan Carlos Escanciano, "Over-Identified Regression Discontinuity Design," *Unpublished, University of Rochester*, 2017.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik, "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 2014, *82* (6), 2295–2326.

_ , _ , Max H Farrell, and Rocio Titiunik, "Regression discontinuity designs using covariates," *Review of Economics and Statistics*, 2019, *101* (3), 442–451.

Cameron, Colin and Pravin Trivedi, *Microeconometrics Using Stata. College Stations*, United States: STATA Press, 2010.

Cattaneo, Matias D, Nicolás Idrobo, and Rocío Titiunik, *A practical introduction to regression discontinuity designs: Foundations*, Cambridge University Press, 2019.

_ , Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele, "Interpreting regression discontinuity designs with multiple cutoffs," *The Journal of Politics*, 2016, *78* (4), 1229–1248.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 2018, *21* (1), 1–68.

_ , Han Hong, and Elie Tamer, "Estimation and confidence regions for parameter sets in econometric models 1," *Econometrica*, 2007, *75* (5), 1243–1284.

Chesher, Andrew and Adam M Rosen, "Generalized instrumental variable models," *Econometrica*, 2017, *85* (3), 959–989.

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

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

CEU eTD Collection

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

Diamond, Peter A. and Jerry A. Hausman, "Contingent Valuation: Is Some Number Better than No Number?," *American Economic Review*, 1994, *8* (4), 45–64.

Duncan, George T, Stephen E Fienberg, Ramayya Krishnan, Rema Padman, and Stephen F Roehrig, "Disclosure Limitation Methods and Information Loss for Tabular Data," in Julia Ingrid Lane, Pat Doyle, Laura Zayatz, and Jules Theeuwes, eds., *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Washington DC: North-Holland, 2001, chapter 2.

Fan, Qingliang, Yu-Chin Hsu, Robert P. Lieli, and Yichong Zhang, "Estimation of Conditional Average Treatment Effects With High-Dimensional Data," *Journal of Business and Economic Statistics*, 2020. forthcoming.

Fox, Craig R. and Yuval Rottenstreich, "Partition priming in judgment under uncertainty," *Psychological Science*, 2003, *14* (3), 195–200.

Friedberg, Rina, Julie Tibshirani, Susan Athey, and Stefan Wager, "Local linear forests," *Journal of Computational and Graphical Statistics*, 2020, pp. 1–15.

Gelman, Andrew and Guido Imbens, "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business and Economic Statistics*, 2019, *37* (3), 447–456.

Givon, Moshe M and Zur Shapira, "Response to rating scales: a theoretical model and its application to the number of categories problem," *Journal of Marketing Research*, 1984, *21* (4), 410–419.

Greene, William H and David A Hensher, *Modeling ordered choices: A primer*, Cambridge University Press, 2010.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 2001, *69* (1), 201–209.

Haisley, Emily, Romel Mostafa, and George Loewenstein, "Subjective Relative Income and Lottery Ticket Purchases," *Journal of Behavioral Decision Making*, 2008, *21*, 283–295.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2011.

Hsiao, Cheng, "Regression analysis with a categorized explanatory variable," in Samuel Karlin, Takeshi Amemiya, and A. Leo Goodman, eds., *Studies in Econometrics, Time Series, and Multivariate Statistics*, Academic Press, 1983, chapter 5, pp. 93–129.

CEU eTD Collection

Hsu, Yu-Chin and Shu Shen, "Testing treatment effect heterogeneity in regression discontinuity designs," *Journal of Econometrics*, 2019, *208* (2), 468–486.

Imai, Kosuke, Marc Ratkovic et al., "Estimating treatment effect heterogeneity in randomized program evaluation," *The Annals of Applied Statistics*, 2013, *7* (1), 443–470.

Imbens, Guido and Karthik Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 2012, *79* (3), 933–959.

Imbens, Guido W and Charles F Manski, "Confidence intervals for partially identified parameters," *Econometrica*, 2004, *72* (6), 1845–1857.

_ **and Thomas Lemieux**, "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 2008, *142* (2), 615–635.

Johnson, David Richard and James C Creech, "Ordinal measures in multiple indicator models: A simulation study of categorization error," *American Sociological Review*, 1983, pp. 398–407.

Kaido, Hiroaki, Francesca Molinari, and Jörg Stoye, "Confidence intervals for projections of partially identified parameters," *Econometrica*, 2019, *87* (4), 1397–1432.

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

Knaus, Michael C., "Double Machine Learning based Program Evaluation under Unconfoundedness," Working Paper, `https://arxiv.org/abs/2003.03191` 2021.

Knaus, Michael, Michael Lechner, and Anthony Strittmatter, "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence," *Econometrics Journal*, 2021, *24* (1), 134–161.

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

_ , "Letter to the Editor: The Kullback-Liebler Distance," *The American Statistician*, 1987, *41*, 340–341.

_ **and R.A. Leibler**, "On information and sufficiency," *Annals of Mathematical Statistics*, 1951, *22*, 79–86.

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

CEU eTD Collection

Lee, David S, "Randomized experiments from non-random selection in US House elections," *Journal of Econometrics*, 2008, *142* (2), 675–697.

_ and Thomas Lemieux, "Regression discontinuity designs in economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

Lewbel, Arthur, "The identification zoo: Meanings of identification in econometrics," *Journal of Economic Literature*, 2019, *57* (4), 835–903.

Ludwig, Jens and Douglas L Miller, "Does Head Start improve children's life chances? Evidence from a regression discontinuity design," *The Quarterly Journal of Economics*, 2007, *122* (1), 159–208.

Manski, Charles F, "Anatomy of the selection problem," *Journal of Human resources*, 1989, pp. 343–360.

_ , *Partial identification of probability distributions*, Springer Science & Business Media, 2003.

_ and Elie Tamer, "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 2002, *70* (2), 519–546.

Mauro, Paolo, "Corruption and Growth," *The Quarterly Journal of Economics*, 1995, *110* (3), 681–712.

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

Micklewright, John and Sylke V Schnepf, "How reliable are income data collected with a single question?," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2010, *173* (2), 409–429.

Nekipelov, Denis, Paul Novosad, and Stephen P. Ryan, "Moment Forests," Working Paper, `https://cpb-us-w2.wpmucdn.com/sites.wustl.edu/dist/5/501/files/2016/10/momentTrees.pdf` 2019.

Pei, Zhuan, David S Lee, David Card, and Andrea Weber, "Local Polynomial Order in Regression Discontinuity Designs," Working Paper 27424, National Bureau of Economic Research June 2020.

Pop-Eleches, Cristian and Miguel Urquiola, "Going to a Better School: Effects and Behavioral Responses," *American Economic Review*, June 2013, *103* (4), 1289–1324.

Ripley, Brian D, *Pattern recognition and neural networks*, Cambridge university press, 1996.

Robson, Matthew, Tim Doran, Richard Cookson et al., "Estimating and decomposing conditional average treatment effects: the smoking ban in England," Technical Report, HEDG, Department of Economics, University of York 2019.

CEU eTD Collection

Romano, Joseph P. and Azeem M. Shaikh, "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 2010, *78* (1), 169–211.

Santos, Adella, Nancy McGuckin, Hikari Yukiko Nakamoto, Danielle Gray, and Susan Liss, "Summary of travel trends: 2009 National Household Travel Survey," Technical Report, United States Department of Transportation 2011.

Schomburg, G., H. Behlau, R. Dielmann, F. Weeke, and H. Husmann, "Sampling techniques in capillary gas chromatography," *Journal of Chromatography A*, 1977, *142*, 87 – 102.

Schomburg, G, H Husmann, and R Rittmann, ""Direct"(on-column) sampling into glass capillary columns: comparative investigations on split, splitless and on-column sampling," *Journal of Chromatography A*, 1981, *204*, 85–96.

Semenova, Vira and Victor Chernozhukov, "Debiased machine learning of conditional average treatment effects and other causal functions," *The Econometrics Journal*, 08 2020, *24* (2), 264–289.

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

Tamer, Elie, "Partial identification in econometrics," *Annual Review of Economics*, 2010, *2* (1), 167–195.

Taylor, Jeremy MG and Menggang Yu, "Bias and efficiency loss due to categorizing an explanatory variable," *Journal of Multivariate Analysis*, 2002, *83* (1), 248–263.

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

Toda, Takayuki, Ayako Wakano, and Takahiro Hoshino, "Regression Discontinuity Design with Multiple Groups for Heterogeneous Causal Effect Estimation," Technical Report, arXiv 2019.

Wager, Stefan and Susan Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

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

Wansbeek, Tom and Erik Meijer, "Measurement error and latent variables," in Badi H. Baltagi, ed., *A Companion to Theoretical Econometrics*, John Wiley & Sons, 2001, chapter 8, pp. 162–179.

Xu, Ke-Li, "Regression discontinuity with categorical outcomes," *Journal of Econometrics*, 2017, *201* (1), 1–18.

CEU eTD Collection

Appendix A

Appendix for Chapter 1

A.1 Decomposition of EMSE criterion

Here I provide the decomposition of $EMSE_\tau(\Pi)$ criterion.

$$
\begin{aligned}
EMSE_\tau(\Pi) &= \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}}\left[MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi)\right] \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2-\tau^2(Z_i)\right\} \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\tau(Z_i)\right\} \\
&\text{using law of iterated expectations} \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\mathbb{E}\left[\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\tau(Z_i)\mid X_i=c,\mathbb{1}_{\ell_1}(Z_i),\ldots,\mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]\right\} \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mathbb{E}\left[\tau(Z_i)\mid X_i=c,\mathbb{1}_{\ell_1}(Z_i),\ldots,\mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]\right\} \\
&\text{as } \mathbb{E}\left[\tau(Z_i)\mid X_i=c,\mathbb{1}_{\ell_1}(Z_i),\ldots,\mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]=\tau(Z_i;\Pi) \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\tau(Z_i;\Pi)\right\} \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i;\Pi)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2-\tau^2(Z_i;\Pi)\right\} \\
&\text{using } Z_i \perp\!\!\!\perp \mathcal{S}^{est} \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i;\Pi)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&\text{using law of iterated expectations and } Z_i \perp\!\!\!\perp \mathcal{S}^{est} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(Z_i;\Pi)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right)^2\mid Z_i\right]\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&\text{Note: } \mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]=\tau(z;\Pi) \text{ where } z \text{ is fixed, thus} \\
&\tau(Z_i;\Pi)=\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\big|_{z=Z_i} \overset{Z_i \perp\!\!\!\perp \mathcal{S}^{est}}{=} \mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mid Z_i\right] \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mid Z_i\right]-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right)^2\mid Z_i\right]\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mid Z_i\right]\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\big|_{z=Z_i}\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \qquad \square
\end{aligned}
$$

CEU eTD Collection

A.2 Derivation of honest sharp RDD criterion

In the following I derive the estimators for the EMSE function for regression discontinuity tree.

$$EMSE_\tau(\Pi) = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right]$$

Let me consider the two parts separately, starting with the expected variance, then the expected square term and finally I put them together.

A.2.1 Expected variance of CATE

Let start with the expected variance part and focus on the variance itself. Here z is fixed, thus

$$\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_-(c,z;\Pi,\mathcal{S}^{est})\right] \\
&= \mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,\hat{\delta}_j^{+,est}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,\hat{\delta}_j^{-,est}\right] \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,\mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\hat{\delta}_j^{+,est}\right] + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,\mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\hat{\delta}_j^{-,est}\right] \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,\left(e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{+,est}\right]e_1\right) + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,\left(e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{-,est}\right]e_1\right) \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\,e_1'\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{+,est}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{-,est}\right]\right)e_1
\end{aligned}$$

where $e_1 = [1,0\ldots,0]$ is still a $1\times(p+1)$ selector-vector. Because $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{+,est}\right]$ and $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{-,est}\right]$ can be estimated using the test sample and the additional knowledge for the number of observations in the estimation sample to adjust for sample size. In case of homoscedastic disturbance term within each leaf the estimator for the variances are

$$\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{+,est}\right] = \frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}}\,, \qquad \hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{-,est}\right] = \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}$$

where

$$\begin{aligned}
N^{est}_{+,j} &= \sum_{i\in\mathcal{S}^{est}}\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\,, \qquad N^{te}_{+,j} = \sum_{i\in\mathcal{S}^{te}}\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \\
\hat{M}_{+,j} &= \frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \\
\hat{\sigma}^2_{+,j} &= \frac{1}{N^{te}_{+,j}-q-1}\sum_{i\in\mathcal{S}^{te}}\left(Y_i - \boldsymbol{X}_i'\hat{\delta}_j^{+,te}\right)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\,, \\
j &= 1,2,\ldots,\#\Pi\,.
\end{aligned}$$

CEU eTD Collection

Same applies for the components of $\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{-,est}\right]$, but using observations, below the threshold, selected by $1-\mathbb{1}_c(X_i)$ instead of $\mathbb{1}_c(X_i)$.
Using these estimates, leads to the following expression for the variance, with scalar z,

$$\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right] = \sum_{j=1}^{\#\Pi}\left\{\mathbb{1}_{\ell_j}(z;\Pi)\; e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right]\right\} e_1$$

Now, I can express the expected value of this expression over the features in the test sample. A natural estimator is the mean of the variances, using Z_i values from the test sample,

$$\begin{aligned}
\hat{\mathbb{E}}_{Z_i}\left\{\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left\{\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(Z_i;\Pi)e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \sum_{j=1}^{\#\Pi}\left\{\left(\frac{\sum_{i\in\mathcal{S}^{te}}\mathbb{1}_{\ell_j}(Z_i;\Pi)}{N^{te}}\right) e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \sum_{j=1}^{\#\Pi}\left\{\frac{N^{te}_j}{N^{te}}e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{N^{est}\frac{N^{te}_j}{N^{te}}e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left(\frac{N^{te}_j}{N^{te}}N^{est}\frac{N^{est}_j}{N^{est}_j}\right) e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left(\frac{N^{te}_j}{N^{te}}\frac{N^{est}}{N^{est}_j}\right) e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}/N^{est}_j} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}/N^{est}_j}\right] e_1\right\} \\
&\quad \text{using assumption for same leaf shares: } \frac{N^{te}_j}{N^{te}} \approx \frac{N^{est}_j}{N^{est}} \\
&\approx \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}/N^{est}_j} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}/N^{est}_j}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left[e_1'\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right] e_1\right\}
\end{aligned}$$

where, N^{te}_j, N^{est}_j are the number of observations within leaf j for the test sample and estimation sample, respectively and $p^{est}_{\pm,j}$ is the share of units above (+) and below (-) the threshold. This derivation uses the fact that observations are randomly assigned to the test sample and to the estimation sample, thus the leaf shares in the test sample (N^{te}_j/N^{te}) is approximately the same as in the estimation sample, (N^{est}_j/N^{est}).

CEU eTD Collection

A.2.2 Expected square of CATE

The second part of the EMSE criterion is the estimator for the expected squared of the true CATE, $\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right]$ over the test sample's features. Using, $\tau(z;\Pi) = \mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]$, where z is fixed, therefore $\tau(Z_i;\Pi) = \mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i}$. Based on this fact, it follows:

$$\begin{aligned}
\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] &= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}^2(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i}\right\} \\
&\quad \text{using variance decomposition} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}^2_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i} - \mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i}\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}^2_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i}\right\}
\end{aligned}$$

The two parts can be estimated by two natural candidates. The expected square CATE is just the average of the squared CATE estimator given by the test sample. The expected variance term is similar to the previous, but note that the variance is estimated purely on the test sample. This means that the scaling factor for number of observations are coming only from the test sample.

$$\begin{aligned}
\hat{\mathbb{E}}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right] \Big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{te}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{te}_{-,j}}\right]e_1\right\} \\
&\quad \text{using assumption for same obs. shares within each leaf:} \\
&\quad p^{te}_{+,j} \approx p^{est}_{+,j},\ p^{te}_{-,j} \approx p^{est}_{-,j},\ \forall j \\
&= \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}
\end{aligned}$$

This expression is the same as the the expected variance using the test sample, the only difference is the scalar N^{te} is used instead of N^{est}. Assumption for same observation shares is used here in order to make the weights the same for the variance estimators. The estimator for expected value of the true squared CATE function over the test sample is given by,

$$\hat{\mathbb{E}}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{te}) - \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}$$

CEU eTD Collection

A.2.3 Estimator for EMSE

Plugging the two parts together yields an estimator for the EMSE criterion,

$$\widehat{EMSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = -\frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}^2(Z_i; \Pi, \mathcal{S}^{te}) + \left(\frac{1}{N^{te}} + \frac{1}{N^{est}}\right) \sum_{j=1}^{\#\Pi} \left\{ e_1' \left[\frac{\hat{\sigma}^2_{+,j} \hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j} \hat{M}^{-1}_{-,j}}{p^{est}_{-,j}} \right] e_1 \right\}$$

A.3 Derivation of honest fuzzy RDD leaf-by-leaf LS criterion

Let assume, that there is a sample $\mathcal{S},\ i = 1, \ldots, N$ with identically and independently distributed observations of (Y_i, X_i, T_i, Z_i). For leaf-by-leaf estimation, I use the fact, $\mathbb{1}_{\ell_j}(z; \Pi)$ creates disjoint sets, and one can estimate the parameters and their variances consistently in each leaf separately. The conditional mean estimator is given by

$$\hat{\mu}^t_+(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \delta_j^{+,t} \quad , \quad \hat{\mu}^t_-(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \delta_j^{-,t}$$

$$\hat{\mu}^y_+(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \delta_j^{+,y} \quad , \quad \hat{\mu}^y_-(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \delta_j^{-,y}$$

where $\delta_j^{+,t}, \delta_j^{-,t}, \delta_j^{+,y}$ and $\delta_j^{-,y}$ estimated by OLS:

$$\hat{\delta}_j^{+,t} = \arg\min_{\delta_j^{+,t}} \sum_{i \in \mathcal{S}} \mathbb{1}_c(x) \mathbb{1}_{\ell_j}(z; \Pi) \left(T_i - \boldsymbol{X}_i' \delta_j^{+,t}\right)^2$$

$$\hat{\delta}_j^{-,t} = \arg\min_{\delta_j^{-,t}} \sum_{i \in \mathcal{S}} (1 - \mathbb{1}_c(x)) \mathbb{1}_{\ell_j}(z; \Pi) \left(T_i - \boldsymbol{X}_i' \delta_j^{-,t}\right)^2$$

$$\hat{\delta}_j^{+,y} = \arg\min_{\delta_j^{+,y}} \sum_{i \in \mathcal{S}} \mathbb{1}_c(x) \mathbb{1}_{\ell_j}(z; \Pi) \left(Y_i - \boldsymbol{X}_i' \delta_j^{+,y}\right)^2$$

$$\hat{\delta}_j^{-,y} = \arg\min_{\delta_j^{-,y}} \sum_{i \in \mathcal{S}} (1 - \mathbb{1}_c(x)) \mathbb{1}_{\ell_j}(z; \Pi) \left(Y_i - \boldsymbol{X}_i' \delta_j^{-,y}\right)^2$$

Estimator for CLATE parameter based on these polynomial functions is given by

$$\hat{\tau}_{FRD}(z; \Pi, \mathcal{S}) = \frac{\hat{\mu}^y_+(c, z; \Pi, \mathcal{S}) - \hat{\mu}^y_-(c, z; \Pi, \mathcal{S})}{\hat{\mu}^t_+(c, z; \Pi, \mathcal{S}) - \hat{\mu}^t_-(c, z; \Pi, \mathcal{S})} = \frac{\hat{\tau}^y(z; \Pi, \mathcal{S})}{\hat{\tau}^t(z; \Pi, \mathcal{S})} = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \frac{\hat{\alpha}^y_{+,j} - \hat{\alpha}^y_{-,j}}{\hat{\alpha}^t_{+,j} - \hat{\alpha}^t_{-,j}}$$

CEU eTD Collection

and its variance:

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S})\right] = &\frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S})^2}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S})\right] \\ &+ \frac{\hat{\tau}^y(z;\Pi,\mathcal{S})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S})^4}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^t(z;\Pi,\mathcal{S})\right] \\ &- 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S})}{\hat{\tau}^t(z;\Pi,\mathcal{S})^3}\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}),\hat{\tau}^t(z;\Pi,\mathcal{S})\right]\end{aligned}$$

where $\mathbb{C}_{\mathcal{S}^{est}}[\cdot,\cdot]$ is the covariance of two random variable. Each part can be decomposed one step further,

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S})\right] \\ \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^t(z;\Pi,\mathcal{S})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S})\right] \\ \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}),\hat{\tau}^t(z;\Pi,\mathcal{S})\right] &= \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S})\right] \\ &\quad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S})\right]\end{aligned}$$

I use the same expected MSE criterion for fuzzy design as well. After the same manipulations as in Section A.1, one gets:

$$EMSE_\tau(\Pi) = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2_{FRD}(Z_i;\Pi)\right]$$

One can construct estimators for these two terms. The variance part from the expected variance is

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right] = &\frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\ &+ \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\ &- 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] \\ &\quad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right])\end{aligned}$$

Decomposing $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right]$:

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\delta}_j^{+,y,est}\right] \\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\hat{\delta}_j^{+,y,est}\right] \\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{+,y,est}\right]e_1\end{aligned}$$

CEU eTD Collection

and $\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right]$:

$$\begin{aligned}\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{C}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\delta}_j^{+,y,est}, e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\delta}_j^{+,t,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\mathbb{C}_{\mathcal{S}^{est}}\left[e_1'\hat{\delta}_j^{+,y,est}, e_1'\hat{\delta}_j^{+,t,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}_j^{+,y,est}, \hat{\delta}_j^{+,t,est}\right]e_1\end{aligned}$$

All the other variances/covariance have the same form with the appropriate parameter vector.

Because $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, one can estimate all the variances and covariances using the observations from the test sample and use only the additional knowledge on the number of observations in the estimation sample. In the simplest – finite variances of the error terms within each leaf – one can write the following sample analogues (below threshold units it is similar).

$$\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\delta}_j^{+,y,est}\right] = \frac{\hat{\sigma}_{+,j}^{2,y}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}, \quad \widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\delta}_j^{+,t,est}\right] = \frac{\hat{\sigma}_{+,j}^{2,t}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}, \quad \widehat{\mathbb{C}_{\mathcal{S}^{est}}}\left[\hat{\delta}_j^{+,y,est},\hat{\delta}_j^{+,t,est}\right] = \frac{\hat{C}_{+,j}^{y,t}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}$$

where

$$\begin{aligned}N_{+,j}^{est} &= \sum_{i\in\mathcal{S}^{est}}\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \quad , \quad N_{+,j}^{te} = \sum_{i\in\mathcal{S}^{te}}\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\\ \hat{M}_{+,j} &= \frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\\ \hat{\sigma}_{+,j}^{2,y} &= \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\left[\left(\epsilon_i^y\right)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right] \ , \qquad \epsilon_i^y = Y_i - \boldsymbol{X}_i'\hat{\delta}_j^{+,y,te}\\ \hat{\sigma}_{+,j}^{2,t} &= \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\left[(\epsilon_i^t)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right] \ , \qquad \epsilon_i^t = T_i - \boldsymbol{X}_i'\hat{\delta}_j^{+,t,te}\\ \hat{C}_{+,j}^{y,t} &= \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\left(\epsilon_i^y\epsilon_i^t\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right) \ , \qquad j=1,2,\ldots,\#\Pi\end{aligned}$$

Remark: the number of observations and the inverse of the running variable's product is the same for both treatment and outcome equation. It is also easy to use other variance estimators (e.g., heteroscedastic-robust versions or clustered), see Appendix A.4.

Putting together the variances, in the homoscedastic case I have the following expres-

CEU eTD Collection

sion,

$$\begin{aligned}
\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right] &= \frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\left[\frac{e_1'\left(\hat{\sigma}^{2,y}_{+,j}\hat{M}^{-1}_{+,j}\right)e_1}{N^{est}_{+,j}}+\frac{e_1'\left(\hat{\sigma}^{2,y}_{-,j}\hat{M}^{-1}_{-,j}\right)e_1}{N^{est}_{-,j}}\right]\\
&+\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\left[\frac{e_1'\left(\hat{\sigma}^{2,t}_{+,j}\hat{M}^{-1}_{+,j}\right)e_1}{N^{est}_{+,j}}+\frac{e_1'\left(\hat{\sigma}^{2,t}_{-,j}\hat{M}^{-1}_{-,j}\right)e_1}{N^{est}_{-,j}}\right]\\
&-2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\left[\frac{e_1'\left(\hat{C}^{y,t}_{+,j}\hat{M}^{-1}_{+,j}\right)e_1}{N^{est}_{+,j}}+\frac{e_1'\left(\hat{C}^{y,t}_{-,j}\hat{M}^{-1}_{-,j}\right)e_1}{N^{est}_{-,j}}\right]\\
&=\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\left(\frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left[\frac{\left(\hat{\sigma}^{2,y}_{+,j}\hat{M}^{-1}_{+,j}\right)}{N^{est}_{+,j}}+\frac{\left(\hat{\sigma}^{2,y}_{-,j}\hat{M}^{-1}_{-,j}\right)}{N^{est}_{-,j}}\right]\right.\\
&\quad+\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}\left[\frac{\left(\hat{\sigma}^{2,t}_{+,j}\hat{M}^{-1}_{+,j}\right)}{N^{est}_{+,j}}+\frac{\left(\hat{\sigma}^{2,t}_{-,j}\hat{M}^{-1}_{-,j}\right)}{N^{est}_{-,j}}\right]\\
&\quad\left.-2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}\left[\frac{\left(\hat{C}^{y,t}_{+,j}\hat{M}^{-1}_{+,j}\right)}{N^{est}_{+,j}}+\frac{\left(\hat{C}^{y,t}_{-,j}\hat{M}^{-1}_{-,j}\right)}{N^{est}_{-,j}}\right]\right)e_1\\
&=\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\left(\frac{\mathcal{V}_{+,j}}{N^{est}_{+,j}}+\frac{\mathcal{V}_{-,j}}{N^{est}_{-,j}}\right)e_1
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{V}_{+,j} &= \frac{\hat{M}^{-1}_{+,j}}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left(\hat{\sigma}^{2,y}_{+,j}+\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\hat{\sigma}^{2,t}_{+,j}+\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})}\hat{C}^{y,t}_{+,j}\right)\\
\mathcal{V}_{-,j} &= \frac{\hat{M}^{-1}_{-,j}}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left(\hat{\sigma}^{2,y}_{-,j}+\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\hat{\sigma}^{2,t}_{-,j}+\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})}\hat{C}^{y,t}_{-,j}\right)
\end{aligned}$$

The expected value of this variance over Z_i from the test sample, can be calculated similarly as in the sharp RDD case.

$$\begin{aligned}
\hat{\mathbb{E}}_{Z_i}\left\{\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left\{\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\left(\frac{\mathcal{V}_{+,j}}{N^{est}_{+,j}}+\frac{\mathcal{V}_{-,j}}{N^{est}_{-,j}}\right)e_1\right\}\\
&\approx\frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left(\frac{\mathcal{V}_{+,j}}{p^{est}_{+,j}}+\frac{\mathcal{V}_{-,j}}{p^{est}_{-,j}}\right)e_1\right\}
\end{aligned}$$

The second part of the EMSE criterion is the estimator for the expected squared

CEU eTD Collection

$\tau^2_{FRD}(Z_i;\Pi)$. Similarly to sharp RD, one can construct the following estimator,

$$\hat{\mathbb{E}}_{Z_i}\left[\tau^2_{FRD}(Z_i;\Pi)\right] = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2_{FRD}(Z_i;\Pi,\mathcal{S}^{te}) - \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi} e_1'\left(\frac{\mathcal{V}_{+,j}}{p^{est}_{+,j}} + \frac{\mathcal{V}_{-,j}}{p^{est}_{-,j}}\right)e_1$$

Note, here everything is estimated on the test sample and I used the assumption, that the number of unit shares for below and above the threshold – for all leaf – are approximately the same in the estimation and test sample ($p^{te}_{+,j} \approx p^{est}_{+,j}, p^{te}_{-,j} \approx p^{est}_{-,j}$). The feasible criteria for fuzzy design for EMSE:

$$\begin{aligned}\widehat{EMSE}_{\tau_{FRD}}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2_{FRD}(Z_i;\Pi,\mathcal{S}^{te}) \\ &+\left(\frac{1}{N^{te}}+\frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi} e_1'\left(\frac{\mathcal{V}_{+,j}}{p^{est}_{+,j}} + \frac{\mathcal{V}_{-,j}}{p^{est}_{-,j}}\right)e_1\end{aligned}$$

A.4 Derivation of variances for leaf-by-leaf LS criterion

Homoscedastic error assumption is rather a strong assumption in RD context, thus use of different heteroscedastic consistent estimators are favourable. First, I show derivation of $\widehat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j\right]$ – the other parts can be calculated similarly – then I put together with the other parts.
General case:

$$\begin{aligned}\widehat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j\right] &= \frac{1}{N^{est}_{+,j}}\left(\frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right)^{-1}\left[\frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\hat{\Omega}\boldsymbol{X}_i\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right] \\ &\qquad\left(\frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right)^{-1} \\ &= \frac{1}{N^{est}_{+,j}}\hat{M}^{-1}_{+,j}\left[\frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\hat{\Omega}\boldsymbol{X}_i\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right]\hat{M}^{-1}_{+,j} \\ &= \frac{1}{N^{est}_{+,j}}\hat{M}^{-1}_{+,j}\hat{\Sigma}_{+,j}\hat{M}^{-1}_{+,j}\end{aligned}$$

Estimators are different in how to calculate $\hat{\Sigma}_{+,j}$:
White's estimator ('HCE0'):

$$\hat{\Sigma}^{HCE0}_{+,j} = \frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\boldsymbol{X}_i(\epsilon^y_i)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

Adjusted 'HCE1':

$$\hat{\Sigma}^{HCE1}_{+,j} = \frac{1}{N^{te}_{+,j}-p-1}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\boldsymbol{X}_i(\epsilon^y_i)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

CEU eTD Collection

In case of clustered SE, with HC1

$$\hat{\Sigma}^{C}_{+,j} = \frac{N^{te}_{+,j} - 1}{(N^{te}_{+,j} - p - 1)^2} \frac{G^{te}_{+,j}}{G^{te}_{+,j} - 1} \sum_{i \in \mathcal{S}^{te}} \left(\sum_{c=1}^{G^{te}_{+,j}} \boldsymbol{X}'_{i,c} \boldsymbol{X}_{i,c} (\epsilon^{y}_{i,c})^2 \right) \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)$$

where $G^{te}_{+,j}$ is the number of clusters in leaf j above the threshold in the test sample. The variance estimators are similarly constructed for parameters below the threshold.

In sharp RD, one gets the variance estimator as,

$$\mathbb{V}_{\mathcal{S}^{est}} \left[\hat{\tau}_{SRD}(z; \Pi, \mathcal{S}^{est}) \right] = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(Z_i; \Pi) e'_1 \left\{ \frac{\hat{M}^{-1}_{+,j} \hat{\Sigma}_{+} \hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{M}^{-1}_{-,j} \hat{\Sigma}_{-} \hat{M}^{-1}_{-,j}}{N^{est}_{-,j}} \right\} e_1$$

In fuzzy RD, let $A_1 = \frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}$, $A_2 = \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}$ and $A_3 = \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}$. Putting

CEU eTD Collection

together the variance for CLATE parameters,

$$
\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_F(z;\Pi,\mathcal{S}^{est})\right] &= \left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\
&+ A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\
&- 2A_3(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] \\
&\qquad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]) \\
&= A_1\left(\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j\right]e_1 + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,y,est}_j\right]e_1\right) \\
&+ A_2\left(\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,t,est}_j\right]e_1 + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,t,est}_j\right]e_1\right) \\
&- 2A_3\left(\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j,\hat{\delta}^{+,t,est}_j\right]e_1\right. \\
&\qquad \left. + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,y,est}_j,\hat{\delta}^{-,t,est}_j\right]e_1\right) \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{A_1\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,y,est}_j\right]\right)\right. \\
&\qquad + A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,t,est}_j\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,t,est}_j\right]\right) \\
&\qquad \left. -2A_3\left(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j,\hat{\delta}^{+,t,est}_j\right] + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,y,est}_j,\hat{\delta}^{-,t,est}_j\right]\right)\right\}e_1 \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{A_1\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j\right] + A_2\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,t,est}_j\right] - 2A_3\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}^{+,y,est}_j,\hat{\delta}^{+,t,est}_j\right]\right. \\
&\qquad \left. + A_1\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,y,est}_j\right] + A_2 + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,t,est}_j\right] - 2A_3\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\delta}^{-,y,est}_j,\hat{\delta}^{-,t,est}_j\right]\right\}e_1 \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{\frac{1}{N^{est}_{+,j}}\hat{M}^{-1}_{+,j}\left(A_1\hat{\Sigma}^y_{+,j} + A_2\hat{\Sigma}^t_{+,j} - 2A_3\hat{C}_{+,j}\right)\hat{M}^{-1}_{+,j}\right. \\
&\qquad \left. + \frac{1}{N^{est}_{-,j}}\hat{M}^{-1}_{-,j}\left(A_1\hat{\Sigma}^y_{-,j} + A_2\hat{\Sigma}^t_{-,j} - 2A_3\hat{C}_{-,j}\right)\hat{M}^{-1}_{-,j}\right\}e_1 \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{\frac{1}{N^{est}_{+,j}}\hat{M}^{-1}_{+,j}\hat{\Sigma}^*_+\hat{M}^{-1}_{+,j} + \frac{1}{N^{est}_{-,j}}\hat{M}^{-1}_{-,j}\hat{\Sigma}^*_-\hat{M}^{-1}_{-,j}\right\}e_1
\end{aligned}
$$

This result is quite useful: there is no need to calculate and multiply with $\hat{M}^{-1}_{\pm,j}$ multiple times during calculating the variances, but they can be 'added up', using only the test sample.

A.5 Monte Carlo simulation setup

For Monte Carlo simulations, I use a general formulation for the DGPs and change the appropriate parts for each specific setup.

$$Y_i = \eta(X_i, Z_{i,k}) + \mathbb{1}_c(X_i) \times \kappa(Z_{i,k}) + \epsilon_i$$

CEU eTD Collection

where $\eta(X_i, Z_{i,k})$ is the conditional expectation function, which is depending on the running variable (X_i) and can be a function of the features $(Z_{i,k})$ as well. The disturbance term is generated from a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. I generate $k = 1, \ldots, K$ features such that $Z_{i,k}$ is independent across k and independent from ϵ_i, X_i. The source of variation comes from ϵ_i during the simulations, thus $X_i, Z_{i,k}$ are the same across the Monte Carlo samples. All the other terms are dependent on the setup.
I report three Monte Carlo average statistics to evaluate the performance of the algorithm:

1. Average of the infeasible MSE: $MSE = \frac{1}{N^{eval}} \sum_{i=1}^{N^{eval}} (\kappa(Z_{i,k}) - \hat{\tau}(Z_i; \hat{\Pi}(\mathcal{S}^{tr}), \mathcal{S}^{est}))^2$
2. Average number of leaves in the final tree.
3. DGP found: this is only feasible for DGPs, where the DGP itself has a tree structure. The DGP is said to be found if the used features for the final tree is exactly the same as for the DGP. [1]

For DGP 1 and 2, I use linear in X_i DGPs with $X_i \sim U[-1, 1]$ where the threshold value is $c = 0$. For the features, I use four variables, two binary $(Z_{i,1-2})$ with 0.5 probability of being 1. For DGP-2 I add two uniformly distributed continuous variables: $Z_{i,3-4} \sim U[-5, 5]$.

DGP 1: Two treatment effect and homogeneous $\eta(\cdot)$. $Z_i = [Z_{i,1}, Z_{i,2}]$, where $Z_{i,1}$ is relevant for CATE, the other is irrelevant.

$$\eta(X_i) = 2 \times X_i$$
$$\kappa(Z_{i,1}) = Z_{1,i} - (1 - Z_{1,i})$$

DGP 2: Continuous treatment effect and heterogeneous $\eta(\cdot)$. $Z_i = [Z_{i,1}, Z_{i,2}, Z_{i,3}, Z_{i,4}]$, $Z_{i,3}$ is relevant for CATE, $Z_{i,2}$ has an effect on $\eta(\cdot)$, the others are irrelevant.

$$\eta(X_i, Z_{i,2}) = 2 \times Z_{i,2} \times X_i - 2 \times (1 - Z_{i,2}) \times X_i$$
$$\kappa(Z_{i,3}) = 2 \times Z_{i,3}$$

DGP 3-5 uses nonlinear specification for X_i. I follow (Calonico et al., 2014) Monte Carlo setups, where $\eta(\cdot)$ is nonlinear in X_i and supplement with heterogeneous treatment effects. (Calonico et al., 2014) imitate two empirical applications and add one extra setup to investigate the effect of excess curvature. For all three designs the running variable is generated by $X_i \sim (2\mathcal{B}(2, 4) - 1)$, where $\mathcal{B}$ denotes a beta distribution and the disturbance term has the variance of $\sigma_\epsilon^2 = 0.05$. The threshold value is the same as in DGP-1 and 2.

[1] I allow the splitting value for each feature to be within 0.5 threshold to accept the split to be similar as the DGP's. Also note that growing smaller or larger trees has different types of errors.

CEU eTD Collection

DGP 3: Imitating (Lee, 2008) vote-shares. I assume two treatment effects and heterogeneous $\eta(\cdot)$. I use 52 dummy variables representing political parties and states. Political party dummy $(X_{i,1})$ is relevant and has an effect on both treatment and functional form. States are irrelevant. For $Z_{i,1} = 1$, I set the functional form as in (Calonico et al., 2014) first setup.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 0.48 + 1.27X_i + 7.18X_i^2 + 20.21X_i^3 + 21.54X_i^4 + 7.33X_i^5, & \text{if } X_i < 0, Z_{i,1} = 1 \\ 0.48 + 2.35X_i + 8.18X_i^2 + 22.21X_i^3 + 24.14X_i^4 + 8.33X_i^5, & \text{if } X_i < 0, Z_{i,1} = 0 \\ 0.48 + 0.84X_i - 3.00X_i^2 + 7.99X_i^3 - 9.01X_i^4 + 3.56X_i^5, & \text{if } X_i \geq 0, Z_{i,1} = 1 \\ 0.48 + 1.21X_i - 2.90X_i^2 + 6.99X_i^3 - 10.01X_i^4 + 4.56X_i^5, & \text{if } X_i \geq 0, Z_{i,1} = 0 \end{cases}$$

$$\kappa(Z_{i,1}) = 0.02 \times Z_{1,i} + 0.07 \times (1 - Z_{1,i})$$

DGP 4: (Ludwig and Miller, 2007) studied the effect of Head Start funding to identify the program's effects on health and schooling. I assume continuous treatment effect based on the age of participants. Age is assumed to be uniformly distributed: $Z_{i,1} \sim U[5, 9]$ and I add dummies representing different continents involved in the analysis.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 3.71 + 2.30X_i + 3.28X_i^2 + 1.45X_i^3 + 0.23X_i^4 + 0.03X_i^5, & \text{if } X_i < 0 \\ 3.71 + 18.49X_i - 54.81X_i^2 + 74.30X_i^3 - 45.02X_i^4 + 9.83X_i^5, & \text{if } X_i < 0 \end{cases}$$

$$\kappa(Z_{i,1}) = -5.45 - (Z_{1,i} - 5);$$

DGP 5: 'An Alternative DGP' by (Calonico et al., 2014) adds extra curvature to the functional form. This design is exactly the same as in (Calonico et al., 2014), thus it has homogeneous treatment effect. The features are the same as in DGP 4 and they are all set to be irrelevant. Treatment effect and $\eta(\cdot)$ is homogeneous.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 0.48 + 1.27X_i - 0.5 \times 7.18X_i^2 + 0.7 \times 20.21X_i^3 \\ \quad +1.1 \times 21.54X_i^4 + 1.5 \times 7.33X_i^5, & \text{if } X_i < 0 \\ 0.48 + 0.84X_i - 0.1 \times 3.00X_i^2 - 0.3 \times 7.99X_i^3 \\ \quad -0.1 \times 9.01X_i^4 + 3.56X_i^5, & \text{if } X_i < 0 \end{cases}$$

$$\kappa = 0.04$$

CEU eTD Collection

A.6 Monte Carlo simulation for fuzzy design

For fuzzy designs, I use the same functional forms and setups for the DGPs, but add a homogeneous first-stage for getting the treatment:

$$T_i = \begin{cases} \mathbb{1}\left(0.5 + 0.8X_i + \nu_i > 0\right) \ , & \text{if } X_i \geq 0 \\ 0 & \text{if } X_i < 0 \end{cases}$$

where $\nu_i \sim \mathcal{N}(0,1)$. For simplicity I use 'DGP-x-f' expression for referring these fuzzy setups. Table A.1 and A.2 show the same algorithm performance measures and the evidence on valid inference similarly to the sharp design. The results are aligned with the conclusion reported in Section 1.4, but the fuzzy design is even more data intensive.

DGP	N	inf. MSE	$\#\hat{\Pi}$	DGP found (%)
DGP-f-1	$N = 1,000$	1.1129	1.00	0%
	$N = 5,000$	0.0267	2.04	96%
	$N = 10,000$	0.0126	2.03	97%
DGP-f-2	$N = 1,000$	13.1595	2.00	-
	$N = 5,000$	4.6662	5.83	-
	$N = 10,000$	3.3652	8.99	-
DGP-f-3	$N = 1,000$	0.0012	1.00	0%
	$N = 5,000$	0.0003	1.99	99%
	$N = 10,000$	0.0001	2.00	100%
DGP-f-4	$N = 1,000$	1.6566	1.00	-
	$N = 5,000$	0.2255	3.00	-
	$N = 10,000$	0.1351	3.69	-
DGP-f-5	$N = 1,000$	0.0006	1.00	100%
	$N = 5,000$	0.0001	1.03	97%
	$N = 10,000$	0.0001	1.02	98%

Table A.1: Monte Carlo averages for performance measures in fuzzy designs

Number of true leaves: $\#\Pi_{DGP-1} = 2$, $\#\Pi_{DGP-3} = 2$, $\#\Pi_{DGP-5} = 1$
Algorithm setup: using the smallest cross-validation value to select γ^*,
$q = 1$ for DGP 1 and 2 and $q = 5$ for DGP 3,4 and 5.

CEU eTD Collection

DGP 1	Leaf	$\ell_1 : \tau_1(Z_1 = 1) = 1$		$\ell_2 : \tau_1(Z_1 = 0) = -1$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-	-	-	-
	$N = 5,000$	-0.0147	0.95	-0.0037	0.95
	$N = 10,000$	-0.0038	0.95	0.0020	0.96
DGP 3	Leaf	$\ell_1 : \tau_1(Z_1 = 0) = 0.07$		$\ell_2 : \tau_1(Z_1 = 1) = 0.02$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-	-	-	-
	$N = 5,000$	-0.0002	0.96	0.0004	0.96
	$N = 10,000$	-0.0003	0.95	-0.0003	0.94
DGP 5	Leaf	Homogeneous Treatment, $\tau = 0.04$			
	Estimates	avgerage bias		actual 95% CI coverage	
	$N = 1,000$	-0.0000		0.95	
	$N = 5,000$	0.0001		0.96	
	$N = 10,000$	-0.0003		0.95	

Table A.2: Estimated Monte Carlo average for bias and actual 95% confidence intervals coverage for each leaf for tree structured DGPs, conditional on DGP is found - fuzzy design

Note: For DGP-f-1 and DGP-f-3, with $N = 1,000$, there is no case when the true DGP is found, thus no values are reported.

CEU eTD Collection

A.7 Additional results on the empirical exercise

This part adds additional information on the empirical analysis. Table A.3 shows the descriptives for the used variables in the heterogeneity analysis. Here, I only present the variables used for revisiting the heterogeneity analysis by (Pop-Eleches and Urquiola, 2013).

	School level average transition score	Baccalaureate taken	Baccalaureate grade	Scaled School admission score	number of schools in town
Mean	7.65	0.74	8.12	0.10	17.50
Median	7.55	1.00	8.15	0.15	17.00
Std deviation	0.75	0.44	0.90	0.55	7.49
Min	5.78	0.00	5.19	-1.00	2.00
Max	9.63	1.00	10.00	1.00	29.00
N	1,857,376	1,857,376	1,256,038	1,857,376	1,857,376

Table A.3: Descriptive statistics of the variables used in heterogeneity analysis of (Pop-Eleches and Urquiola, 2013)

Table A.4 restates the main findings of (Pop-Eleches and Urquiola, 2013) on the heterogeneity exercise.

CEU eTD Collection

	School level average transition score	Baccalaureate taken	Baccalaureate grade
Full sample			
τ_0	0.107^{***}	0.000	0.018^{***}
$SE(\tau_0)$	(0.001)	(0.001)	(0.002)
N	1,857,376	1,857,376	1,256,038
Top tercile			
τ_1	0.158^{***}	0.003	0.048^{***}
$SE(\tau_1)$	(0.002)	(0.002)	(0.003)
N_1	756,141	756,141	579,566
Bottom tercile			
τ_2	0.099^{***}	-0.008^{*}	-0.005
$SE(\tau_2)$	(0.003)	(0.004)	(0.009)
N_2	392,475	392,475	212,282
Towns with four or more schools			
τ_1	0.097^{***}	0.000	0.016^{***}
$SE(\tau_1)$	(0.001)	(0.001)	(0.002)
N_1	1,806,411	1,806,411	1,223,341
Towns with three schools			
τ_2	0.333^{***}	-0.007	0.028^{*}
$SE(\tau_2)$	(0.007)	(0.009)	(0.016)
N_2	31,149	31,149	19,877
Towns with two schools			
τ_3	0.697^{***}	0.020	0.179^{***}
$SE(\tau_3)$	(0.010)	(0.013)	(0.023)
N_3	19,816	19,816	12,820

Notes: All regressions are clustered at the student level and include cutoff fixed effects. Standard errors are in parentheses. All estimates present reduced form specifications where the key independent variable is a dummy for whether a student's transition score is greater than or equal to the cutoff.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table A.4: Heterogeneity in Baccalaureate Effects - (Pop-Eleches and Urquiola, 2013), Table 5

Table A.5 summarize the different treatment effects estimated by (Pop-Eleches and Urquiola, 2013) and by the algorithm for Baccalaureate exam grade. Note that for RD tree: only number of schools, I only used number of schools only as features. RD tree: all variables are using both average transition score for the class and number of schools as features, but finds only average transition score variable as relevant.

CEU eTD Collection

	Avg. transition score for the class		Number of schools		
Pop-Eleches and Urquiola (2013)	Top tercile 0.048***	Bottom tercile −0.005	2 0.179***	3 0.028*	4-27 0.016***
RD tree: all variables	Below median† 0.015**	Above median† 0.028***	-	-	
RD tree: only no. schools	-	-	2 0.152***	3-24 and 26-27 0.021***	25 −0.013

Regressions are clustered at the student level and include cutoff FE.
***: significant at 1%, **: significant at 5%, *: significant at 10%.
†: the algorithm splits at 44th percentile.

Table A.5: Heterogeneity in treatment effects for Baccalaureate grade

CEU eTD Collection

	Mean	Median	Std. dev.	Min	Max	N
Outcome and running variables						
School level average transition score	8.20	8.29	0.60	6.53	9.41	11,931
Scaled Admission score	0.85	0.82	0.97	-2.07	3.91	11,931
Socioeconomic characteristics of households						
Female head of household (d)	0.89	1	0.32	0	1	11,931
Age of head of household	46.75	45	7.15	13	97	11,843
Romanian (d)	0.94	1	0.24	0	1	11,931
Hungarian (d)	0.05	0	0.22	0	1	11,931
Gypsy (d)	0.01	0	0.06	0	1	11,931
Other Ethnicity (d)	0.01	0	0.09	0	1	11,931
HH's Primary education (d)	0.66	1	0.47	0	1	11,840
HH's Secondary education (d)	0.20	0	0.40	0	1	11,840
HH's Tertiary education (d)	0.13	0	0.34	0	1	11,840
Socioeconomic characteristics of students						
Gender of student (d)	0.42	1	0.49	0	1	11,931
Age of student	18.08	18	0.94	14	23	11,866
Accessibility of households to goods						
Car (d)	0.57	1	0.49	0	1	11,820
Internet (d)	0.73	1	0.44	0	1	11,829
Phone (d)	0.47	0	0.50	0	1	11,807
Computer (d)	0.87	1	0.34	0	1	11,851
Parental and Child responses to survey questions						
Parent volunteered (d)	0.11	0	0.31	0	1	11,868
Parent paid tutoring (d)	0.24	0	0.42	0	1	11,931
Parent helps HW (d)	0.20	0	0.40	0	1	11,815
Child does HW every day - Parent (d)	0.75	1	0.43	0	1	11,779
Negative interactions with peers	0.12	0	0.37	0	5	11,838
Child does HW every day - Child (d)	0.63	1	0.48	0	1	11,908
HW percieved easy	5.45	5.60	1.02	1	7	9,628
Characteristics of schools						
No. schools	2.33	2	0.50	2	4	11,931
2 schools (d)	0.69	1	0.46	0	1	11,931
3 schools (d)	0.29	0	0.45	0	1	11,931
4 schools (d)	0.02	0	0.13	0	1	11,931
Highest certification teacher in school (d)	0.61	1	0.49	0	1	11,169
Novice teacher in school (d)	0.06	0	0.24	0	1	11,169

(d) indicates it is a dummy variable. 'HH' stands for household, 'HW' for homework.

Table A.6: Descriptive statistics of the used variables for exploring heterogeneity in a survey-based dataset

Table A.6 shows the descriptives for the candidate features used to find the tree shown by Figure 1.8.

CEU eTD Collection

Appendix B

Appendix for Chapter 2

The structure of the Appendix B is the following: Section B.1 contains detailed research of the Monte Carlo experiments. Section B.2 provides theoretical exposition of the simple Least Squares (LS) estimator on model with discretized data. The discussion covers cross section and panel data. Section B.3 contains technical proofs of all the Propositions in the paper. Section B.4 provides a list of notations used in the paper.

B.1 Monte Carlo Simulation Results on the Bias

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

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

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

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

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

CEU eTD Collection

B.1.1 Uniform Distribution

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

Table B.1: Uniform distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 5$

From Table B.1 the unbiasedness and consistency (in sample size) of the LS estimator can clearly be seen in the case of the uniform distribution, similarly to the, somewhat slower, convergence in M. We have also done simulations with different σ_ε^2 and β, where the same results hold. For smaller σ_ε^2, the bias is smaller, for different β the results are almost exactly the same. Next, let us turn our attention to some other distributions.

CEU eTD Collection

B.1.2 Other distributions

In this subsection we investigate the normal distribution alng with the exponential and Weibull distributions.

First let us note that the Kullback-Liebler index gives a good indication of the bias. The bias tends to be smaller where this index is small, and vice versa. Table B.2 shows simulation results with a normal distribution.

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

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

Similarly, with a normal distribution, Table B.3 shows that the LS estimator is biased and inconsistent, with a negative bias, as predicted by the theory, both in the case of truncation and censoring. Although the theory suggests that intercept picks up some of the bias, in practice the difference between with and without intercept – in this case – is small, approximately 3-5%.

CEU eTD Collection

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

Table B.3: Truncated and Censored Normal Distributions, estimated without intercept, $M = 5, \beta = 0.5, \sigma_\varepsilon^2 = 5, Supp = [-1, 1]$

CEU eTD Collection

We carried out a large number of simulations with different parametrisations for both distributions. In Table B.4 we report the bias from the exponential distribution, which highlights the effect of censoring. Although we do no observe large bias with truncation, when the choices are censored the bias increases dramatically. Table B.5 shows results on normal distribution, while Table B.6 uses a weibull distribution.

From Table B.4-B.6, the main takeaway is that, as expected, there is no convergence in the sample size, while the convergence speed in M is 'slow' and depends heavily on the shape of the distribution. Also, the results about the Kullback-Liebler index (not reported here) are very similar to those obtained for the normal distribution, i.e., a larger index implies systematically a larger bias.

We have also tried several different distributions and parameterisation, and the main take away is very similar.

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

Table B.4: Exponential distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 5, \lambda = 0.5$

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

Table B.5: Normal distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 1, \mu_x = 0, \sigma_x^2 = 0.2$

CEU eTD Collection

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

Table B.6: Weibull distribution: $\beta = 0.5, \sigma_\varepsilon^2 = 0.5, b = 1, c = 0.5$

B.2 Properties of LS using Discretized Data

Recall the data generating process is assumed to be

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

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

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

Let us assume for the sake of simplicity that there is only one explanatory variable in the model which is observed through discretized choices. It is also assumed, as said earlier, that it has a known support $[a_l, a_u]$ with known boundaries (C_m), and let z_m be the class midpoint.[1]

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

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

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

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

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

CEU eTD Collection

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

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

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

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

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

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

CEU eTD Collection

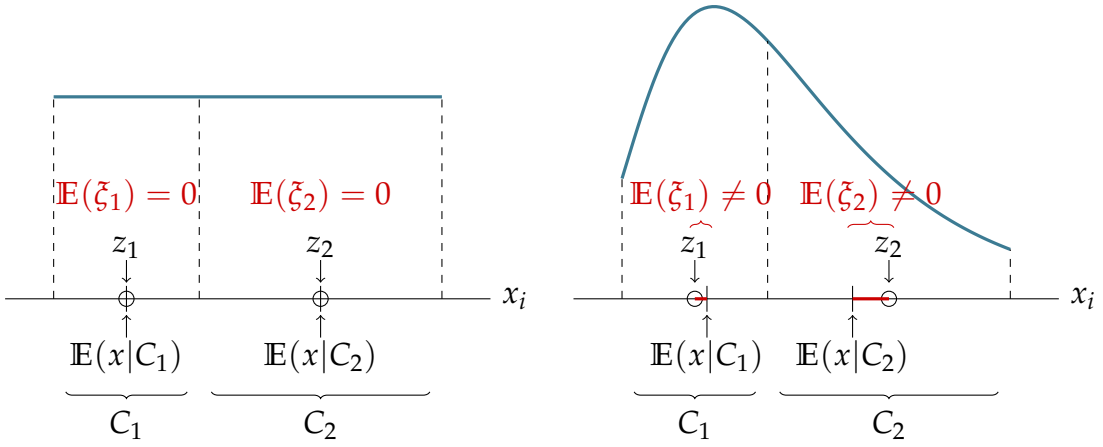


Figure B.1: The difference between uniform (left panel) and general distributions (right panel)

where a respondent makes an error $\xi_i = x_i - x_i^*$ for each observation by setting the possible answer values at x_i^*. The derivation above is based on the disturbance term u_i being independent of regressor x_i and $\mathbb{E}(u_i) = 0$ for all $i = 1, \ldots, N$. The last inference uses the fact that the errors ξ_i have the same conditional distribution over the class C_m, $v^m \overset{\mathrm{d}}{=} \xi_i | C_m$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. Importantly, the second term in Equation (B.4) does not vanish in general, since $v^m | C_m$ is not independent of $N_m | C_m$, $v^m | C_m \not\perp\!\!\!\perp N_m | C_m$ nor $\mathbb{E}(\xi_i | C_m) = \mathbb{E}(v^m) = 0$ (see Figure B.1, right panel). These would be sufficient assumptions for the LS to be unbiased. The former issue can be eliminated by conditioning on the underlying distribution of x_i. Conditional on the distribution x_i and the class C_m, the number of observations in the class and assuming that the errors are independent of each other, $N_m | x_i, C_m \perp\!\!\!\perp v^m | x_i, C_m$, but knowing the underlying distribution makes the problem trivial. Nonetheless, because of both issues, the 'naive' LS estimator is biased.

The uniform distribution, however, turns out to be a special case. Let us assume that $x_i \sim U(a_l, a_u)$ for all $i = 1, \ldots, N$, then both of the above disappear (see the left panel in Figure B.1) if we are using the class mid points. The first problem is resolved, because in the case of the uniform distribution, both the number of observations N_m in each class C_m and the error term v^m are independent of the regressor's x_i distribution, while the second problem does not appear trivially, since now the class midpoints are proper estimates of the regressor's x_i expected value in the class C_m. From Equation (B.4), we obtain that

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

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

CEU eTD Collection

B.2.1 N (in)consistency

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

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

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

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

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

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

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

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

CEU eTD Collection

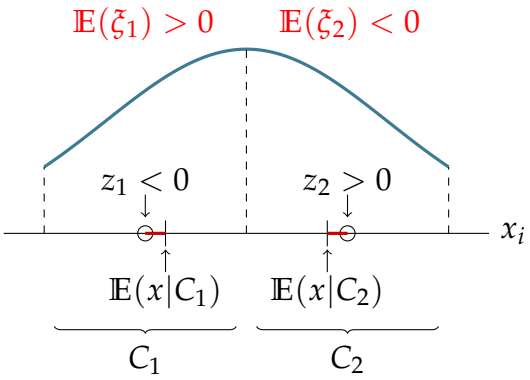


Figure B.2: The estimator is inconsistent even in case of symmetric distributions.

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

The last step in the above derivation can simply be obtained from the definition of the plim operator, i.e., for any $\varepsilon > 0$ given Therefore, to obtain the (in)consistency of the LS estimator $\hat{\beta}^*_{LS}$ in the number of observations N, we only need to calculate the expected value of the truncated random variable x_m, $m = 1, \ldots, M$ and check whether the expression (B.6) equals 0 to satisfy a sufficient condition.

Let us apply these results to the uniform distribution. In this case, there is no consistency issue because the class midpoints coincide with the expected value of the truncated uniform random variable in each class, making the expression (B.6) zero, hence the *LS* estimator is consistent.

Note that the consistency of the LS estimator is not guaranteed even in the case of symmetric distributions and symmetric class boundaries. After appropriate transformations (e.g., demeaning), it can be seen that the sign of the differences between the expectation of the truncated random variables x_m and the class midpoints is opposite to the sign of the class midpoints on either side of the distribution, which implies negative overall asymptotic bias in N (see Figure B.2).

In the case of a (truncated) normal variable, for example, we need to substitute the expected value of the truncated normal random variable x_m for each $m = 1, \ldots, M$ in the consistency formula (B.6). As a result, the difference between the expectation and the class midpoints in general is not zero for all m, hence the formula cannot be made arbitrarily small. Therefore, the LS estimator becomes inconsistent in N (see Table B.3. on the size of the bias).

So far we have focused on the estimation of β in Equation (B.2). But how about γ? It

CEU eTD Collection

can be shown that the bias and inconsistency presented above is contagious. Estimation of all parameters of a model is going to be biased and inconsistent unless the measurement error and x are orthogonal (independent), which is quite unlikely in practice. This is important to emphasize: a single choice type variable in a model is going to infect the estimation of all variables of the model.

B.2.2 M Consistency

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

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

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

CEU eTD Collection

of choices for a survey is relatively small, $M = 3, 5, 7$ or at most $M = 10$.[2]

B.2.3 Some Remarks

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

The measurement error due to discretized choice variables, however, not only induces correlation between the error terms and the observed variables, but it also induces a non-zero expected value for the disturbance terms of the regression in (B.2). Consider a simple example where there is an unobserved variable x_i with an observed discretized choice version:

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

and

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

Using the discretized choice variable means:

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

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

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

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

B.2.4 Estimation Reconsidered

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

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

where $\mathbf{X}$ and $\mathbf{W}$ are $N \times K$ and $N \times J$ data matrices of the explanatory variables, $\mathbf{y}$ is a $N \times 1$ vector containing the data of the dependent variable, $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of disturbance terms, and finally $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $K \times 1$ and $J \times 1$ parameter vectors. $\mathbf{X}$ is not observed, only its

[2] There is an abundant literature about the optimal number of choices (or 'scale points') in a survey, see e.g., (Givon and Shapira, 1984), (Srinivasan and Basu, 1989) or (Alwin, 1992).

CEU eTD Collection

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

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

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

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

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

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

The LS Estimator

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

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

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

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

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

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

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

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

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

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

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

The LS estimator in this case becomes:

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

CEU eTD Collection

The LS is unbiased if $(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{E}\mathbf{X} = \mathbf{I}$. Note that $\mathbf{Z}'$ and $\mathbf{E}$ are of size $K \times MK$ and $MK \times N$, respectively. This means $\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}$ are invertible as long as $N > K$, which is a standard assumption in classical regression analysis. Let us consider a typical element in $\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}$ first. Since $\mathbf{Z}$ is non-stochastic as it contains only all the pre-defined choice values, it is sufficient to examine $\mathbf{E}\mathbf{E}'$:

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

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

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

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

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

where $f(x_k, x_l)$ denotes the joint distribution of x_k and x_l and $\Omega = [c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}]$ defines the region for integration. Thus, $N^{-1}b_{mn}$ should converge into Equation (B.15) under the usual WLLN.

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

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

Now,

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

where $\Omega_1 = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}}$ with $\Omega_{\mathbf{X}}$ denotes the sample space of x_k and x_l. Thus, $N^{-1}a_{kl}$ should converge into Equation (B.16) under the usual WLLN.

In the case when Equations (B.13) and (B.14) do not hold, the analysis becomes more tedious algebraically, but it does not affect the result that LS is biased. Recall Equation (B.12), and let ω_{ij}

CEU eTD Collection

be the (i, j) element in $\mathbf{M_W}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, then following the same argument as above, $\mathbf{EM_WE}'$ can be expressed as a $M \times M$ block partition matrix with each entry a $K \times K$ matrix. The typical (m, n) element in the (k, l) block is

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

with its expected value being

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

where $\mathbf{w} = (w_1, \ldots, w_J)$, $d\mathbf{w} = \prod_{i=1}^{J} dw_i$ and $\Omega = [c_{km-1}, c_{km}] \times [c_{ln-1}, c_{ln}] \times \Omega_{\mathbf{w}}$ where $\Omega_{\mathbf{w}}$ denotes the sample space of $\mathbf{w}$. Note that ω_{ij} is a nonlinear function of $\mathbf{w}$, and so the condition of existence for Equation (B.18) is complicated. However, under the assumption that the integral in Equation (B.18) exits, then $N^{-1} g_{kl}$ should converge to Equation (B.18) 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 (B.18) to Equation (B.15).

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

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

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

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

where u_l denotes the random variable corresponding to the i^{th} column of $\mathbf{M_W X}$ and $\Omega_m = [c_{km-1}, c_{km}] \times \Omega_{\mathbf{X}} \times \Omega_{\mathbf{w}}$ with $\Omega_{\mathbf{w}}$ denotes the sample space of $\mathbf{W}$. Note that $u_l = x_l$ under the assumption of independence, which reduces Equation (B.20) to Equation (B.16).

B.2.5 Extension to Panel Data

So far, we have dealt with cross-sectional data. Next, let us see what changes if we have panel data at hand, which is closer to the reality of data gathering through surveys. We can extend our basic model using Equation (B.2) to

$$y_{it} = w_{it}'\gamma + x_{it}^{*'}\beta + \varepsilon_{it}, \tag{B.21}$$

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

$$y_{it} = w_{it}'\gamma + x_{it}'\beta + u_{it}, \tag{B.22}$$

CEU eTD Collection

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

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

The bias of the panel data within estimator can be easily shown. Let us re-write Equation (B.11) in a panel data context

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

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

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

or equivalently

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

where

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

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

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

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

CEU eTD Collection

B.3 Technical Proofs

B.3.1 Proof of Proposition 1

Recall

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

We can reformulate Equation (B.23) by considering the number of observations up to a certain boundary point, rather than the number of observations in a particular class. That is checking for

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

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

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

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

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

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

B.3.2 Proof of Proposition 2

Recall

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

As $S \to \infty$, $\exists c_b^{WS} = c$ for any $c \in (a_l, a_u)$, by construction. Furthermore, for any c_b^{WS}, $\exists l \in [1, S]$, $m \in [1, M]$ such that $c_b^{WS} = c_m^{(l)}$. Also note that as $S \to \infty$, we need $N \to \infty$ as well. Now consider $\Pr(x^\dagger < c_b^{WS}) = \Pr(x^\dagger < c_m^{(l)})$, given $\Pr(x \in \mathcal{S}_s) = 1/S$ and using equation (B.24)

CEU eTD Collection

gives

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

Note that the summation over the different classes in Equation (B.24) is being replaced as we are considering the cumulative probability and that no value greater than $c_m^{(l)}$ will be used as a candidate in the working sample for c_b^{WS}. Under the shifting method, $c_m^{(s)} \leq c_m^{(l)}$ for $s < l$ and using the definition of conditional probability gives

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

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

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

This completes the proof.

B.3.3 Speed of Convergence for the Shifting Method

Recall

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

For each of the conditions in Equation (B.25), the corresponding expression is $o(1)$. To see this, note that $f(x)$ is a density, so the integral is less than 1. First, consider the case of $s \neq 1$ and

CEU eTD Collection

$m = 1$,

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

As $S \to \infty$, the ratio in the last line goes to 0. This is expected if the widths of the classes in the working sample go to zero. This is straightforward, while the probability that an observation belongs to a point is 0. The same derivations applies to the case when $m = M$. Now, consider the case of $1 < m < M$,

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

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

B.4 Summary of the Notation Used in the Paper

Scalars:

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

CEU eTD Collection

c_b^{WS} – working sample b'th class's lower boundary point
h – working sample's class widths
Δ – size of shift for the shifting method

Running indexes
i – refers to individual $i = 1, \ldots, N$, and in some places it is a running index.
t – refers to time $t = 1, \ldots, T$
m – refers to class $m = 1, \ldots, M$
k – refers to a DOC variables in matrix formulation, $k = 1, \ldots, K$
j – refers to a control variables in matrix notation, $j = 1, \ldots, J$, and in some places it is a running index.
b – working sample classes, $b = 1, \ldots, B$
s – split sample index
i_m – running index, where m is the indication in which class that observation is (M consistency)
m_i – i-th observation in the m-th class (M consistency)

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

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

CEU eTD Collection

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

Functions

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

Intervals

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

Sets

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

Matrix notations

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

CEU eTD Collection

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

CEU eTD Collection

10.14754/CEU.2021.06

Appendix C

Appendix for Chapter 3

C.1 Additional Monte Carlo simulations

We extend the Monte Carlo simulations in five different ways. The basic setup is the same as in Section 3.4, but we change each time one parameter. All the following tables shows 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.

Remark: In 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 model, * 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.

C.1.1 Moderate sample size

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

[1] We have not decreased our sample size further while for magnifying method in case of $N = 100$ it would mean 10 number effective observations on average.

CEU eTD Collection

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.1, 1.15]$ (0.06),(0.07)	$[-1.09, 1.15]$ (0.08),(0.08)	$[-1.09, 1.16]$ (0.07),(0.07)	$[-1.07, 1.17]$ (0.09),(0.09)	$[-1.06, 1.18]$ (0.08),(0.09)	$[-1.09, 1.15]$ (0.05),(0.06)
Ordered probit*	0.1978 (0.0810)	0.0690 (0.0797)	0.2138 (0.0827)	0.0181 (0.0763)	0.0965 (0.0795)	0.4484 (0.0908)
Ordered logit*	0.6523 (0.1479)	0.3828 (0.1431)	0.6967 (0.1561)	0.2419 (0.1364)	0.4309 (0.1455)	1.2109 (0.1682)
Interval regression	0.0254 (0.0618)	0.0329 (0.0784)	0.0398 (0.0694)	0.0512 (0.0882)	0.0638 (0.0825)	0.0396 (0.0505)
Midpoint regression	0.0209 (0.0643)	0.0293 (0.0786)	0.0310 (0.0733)	0.0453 (0.0895)	0.2029 (0.0426)	0.0275 (0.0526)
Magnifying $(S = 10)$	0.0145 (0.1781)	0.0117 (0.2222)	0.0127 (0.1988)	-0.0184 (0.2538)	0.0757 (0.1023)	0.0330 (0.1358)
Shifting $(S = 10)$	0.0016 (0.0682)	-0.0026 (0.0771)	-0.0031 (0.0696)	-0.0053 (0.0872)	0.0050 (0.0441)	-0.0010 (0.0498)

Table C.1: Monte Carlo average bias and standard deviation with moderate sample size, $N = 1,000$

Shifting method always outperforms the alternatives. Magnifying method gives better results, except in the exponential and weibull cases where it has similar magnitude of bias as the interval regression (exponential case) or the midpoint regression (weibull case). Note that in these cases interval regression and midpoint regression are not superior to the magnifying method. They only outperform magnifying method 'at random'. As we will show in Table C.4 these methods are inconsistent in N, however magnifying method does converge to the true parameter value.

C.1.2 Symmetric boundaries

Next, we investigate symmetric boundary cases. We set the domain for y_i 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.

CEU eTD Collection

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Set identification†	$[-1.11, 1.13]$ (0.02),(0.02)	$[-1.15, 1.10]$ (0.02),(0.02)	$[-1.09, 1.16]$ (0.02),(0.02)	$[-1.07, 1.17]$ (0.03),(0.03)	$[-1.06, 1.19]$ (0.03),(0.03)	$[-1.09, 1.15]$ (0.02),(0.02)
Ordered probit*	0.0890 (0.0252)	0.0029 (0.0243)	0.2085 (0.0262)	0.0158 (0.0234)	0.0986 (0.0241)	0.4461 (0.0295)
Ordered logit*	0.4513 (0.0446)	0.3198 (0.0427)	0.6862 (0.0499)	0.2379 (0.0422)	0.4338 (0.044)	1.2085 (0.0546)
Interval regression	0.0085 (0.022)	-0.0267 (0.0234)	0.0371 (0.0221)	0.0491 (0.0271)	0.0663 (0.0249)	0.0397 (0.0166)
Midpoint regression	0.0070 (0.0211)	0.0240 (0.0242)	0.0362 (0.0216)	0.0490 (0.0273)	0.2077 (0.0128)	0.0314 (0.0157)
Magnifying $(S = 10)$	-0.0323 (0.0606)	-0.0336 (0.0694)	-0.0072 (0.0616)	-0.0332 (0.0781)	0.0213 (0.0333)	0.0066 (0.0417)
Shifting $(S = 10)$	-0.0028 (0.0222)	0.0002 (0.0234)	-0.0004 (0.0209)	-0.0023 (0.0278)	-0.0034 (0.0131)	-0.0008 (0.0153)

Table C.2: Monte Carlo average bias and standard deviation with symmetric boundary points: $a_l = -3, a_u = 3$

As we expected the maximum likelihood methods, where they have a closer fit to the assumed distribution the distortion is somewhat smaller in case of ordered probit model[2]. 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 mis-specified cases. Magnifying method gives slightly worse results in the normal and logistic cases, but the shifting method performs similarly good.

C.1.3 Number of choices

Another question is how the number of choices (M) effect the bias. We investigated $M = 3$ case, where questionnaire only 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 other even worse. Shifting method gives fairly accurate estimates.

[2]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 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.

CEU eTD Collection

	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)
Magnifying $(S = 10)$	0.0138	0.0022	0.0179	-0.0101	0.0397	0.0299
	(0.0534)	(0.0676)	(0.0606)	(0.0813)	(0.0341)	(0.0418)
Shifting $(S = 10)$	-0.0234	-0.0015	0.0003	-0.0094	-0.0128	-0.0019
	(0.0247)	(0.0236)	(0.0213)	(0.0292)	(0.0154)	(0.0166)

Table C.3: Monte Carlo average bias and standard deviation with small number of choice options, $M = 3$

Also note that with split sampling methods 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.

C.1.4 Convergence in N

Table C.4 shows the (asymptotic) reduction in the bias with the split sampling methods. We use now only normal distribution's setup for ϵ_i. Here we added 'Magnifying with replacement' method, which uses the magnifying method to get the DTOs, then it uses these DTO values to calculate conditional averages to replace the NDTO values.

As the table suggests, as we increase the number of observations the bias vanishes for the split sampling methods. Also if we increase the number of split samples the bias tend to decrease. It is important to highlight that in the magnifying case the effective number of observation is decreasing in S, therefore if we do not increase N the variance of the estimator is increasing. This shows the trade-off between small sample bias and observing the values more precisely. Based on this table we suggest, in case of magnifying method to use only a moderate number of split samples $(3 - 10)$ in case of moderate sample size. For the shifting method there is no such trade-off, however the results are not much better as we increase the number of split samples. 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.

CEU eTD Collection

		$N = 1,000$	$N = 10,000$	$N = 100,000$
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)
Magnifying only DTO	$S = 3$	-0.0526 (0.0916)	-0.0070 (0.0275)	0.0003 (0.0086)
	$S = 5$	0.0116 (0.1226)	0.0379 (0.0363)	-0.0045 (0.0115)
	$S = 10$	0.0217 (0.1694)	-0.0110 (0.0545)	0.0069 (0.0165)
	$S = 25$	0.0939 (0.2522)	-0.0196 (0.0835)	-0.0074 (0.0276)
	$S = 50$	-0.0761 (0.4768)	-0.0050 (0.1233)	0.0075 (0.0392)
	$S = 100$	0.0382 (0.6889)	0.0175 (0.1781)	-0.0033 (0.0557)
Magnifying with replacement	$S = 3$	-0.0597 (0.0986)	-0.0060 (0.0279)	0.0004 (0.0086)
	$S = 5$	-0.0065 (0.1385)	0.0373 (0.0374)	-0.0040 (0.0115)
	$S = 10$	0.0534 (0.1988)	-0.0103 (0.0575)	0.0066 (0.0165)
	$S = 25$	0.1100 (0.3004)	-0.0165 (0.0947)	-0.0075 (0.0280)
	$S = 50$	-0.1135 (0.5403)	-0.0098 (0.1381)	0.0079 (0.0402)
	$S = 100$	0.0918 (0.8123)	0.0189 (0.2061)	-0.0033 (0.0585)
Shifting	$S = 3$	-0.0038 (0.0629)	-0.0018 (0.0198)	-0.0008 (0.0061)
	$S = 5$	-0.0002 (0.0621)	-0.0024 (0.0194)	-0.0001 (0.0059)
	$S = 10$	0.0024 (0.0603)	-0.0016 (0.0189)	-0.0008 (0.0058)
	$S = 25$	0.0013 (0.0592)	-0.0016 (0.0186)	-0.0007 (0.0058)
	$S = 50$	0.0004 (0.0587)	0.0000 (0.0185)	0.0001 (0.0058)
	$S = 100$	0.0004 (0.0596)	-0.0002 (0.0183)	-0.0003 (0.0056)

Table C.4: Bias reduction for split sampling methods: different sample sizes and number of split samples

CEU eTD Collection

C.1.5 Magnifying method with replacement

Finally we report our Monte Carlo experiment with magnifying method using all observations using replacement technique based on DTO. These results are slightly worse than using only DTOs with magnifying method.

	Normal	Logistic	Log-Normal	Uniform	Exponential	Weibull
Base	-0.0043	0.0719	0.0645	0.0813	0.0345	0.0433
	(0.0551)	(0.0719)	(0.0645)	(0.0813)	(0.0345)	(0.0433)
$N = 1000$	0.0476	0.0342	0.0445	-0.0021	0.0868	0.0648
	(0.2043)	(0.2552)	(0.2241)	(0.2895)	(0.1223)	(0.1565)
Symmetric	-0.0323	-0.0328	-0.0081	-0.0317	0.0176	0.0053
	(0.0639)	(0.0720)	(0.0645)	(0.0813)	(0.0345)	(0.0433)
$M = 3$	0.0164	0.0026	0.0181	-0.0093	0.0394	0.0290
	(0.056)	(0.0713)	(0.0638)	(0.0850)	(0.0354)	(0.0434)

Table C.5: Magnifying all observation with replacement using DTOs

Note that this method can be easily computed with the already collected data to check the 'robustness' of the magnifying method.

CEU eTD Collection