Heterogeneous Treatment Effects in Regression Discontinuity Design

Regression Discontinuity Tree

Ágoston Reguly

February 26, 2021

Central European University

Working paper, comments welcome.

Abstract

This paper constructs an algorithm, which uncover treatment effect heterogeneity in classical regression discontinuity (RD) design. I propose an algorithm: *honest* regression discontinuity tree to identify heterogeneity in the treatment effect across sub-populations using additional features or covariates. The relevant sub-populations are *a prior* unknown, and it is the task of the algorithm to discover them, without invalidating inference.

The paper contributes to two literature: for the regression discontinuity designs with the estimate of conditional average treatment effects (CATE) with potentially many covariates with one running variable and with a known threshold value. For the causal machine learning literature with a new criterion to estimate causal effects, where the effects are estimated by leaf-by-leaf regressions.

To show how the algorithm works in practice, I use Pop-Eleches and Urquiola (2013) data on Romanian school system, and uncover heterogeneous effects on the impact of going to a better school.

JEL: C13, C21, I21

Keywords: Supervised machine learning, regression tree, regression discontinuity design, heterogeneous treatment effect, CATE.

1 Introduction

This paper studies the estimation of treatment effect heterogeneity by features or covariates in regression discontinuity (RD) designs. Treatment effects are identifiable with observational data, when there is discontinuity in the regression function, caused by the treatment. These causal effects are based on comparisons of outcomes from below and above the known threshold, assumed that there is a local randomization around the threshold, or there is no perfect control of the treatment by the individuals (see, Imbens and Lemieux (2008), Lee and Lemieux (2010)). In classical RD settings, one estimates the *average* treatment effect around this threshold value at which the treatment assignment changes, using information on the running variable, the outcome and the threshold point (e.g. Hahn et al., 2001, Calonico et al., 2014). In the last few years, regression discontinuity became extremely popular in theoretical and empirical works, resulting in fair amount of extensions. [1]

This paper contributes to the RD literature with discovering heterogeneous treatment effects in a systematic way, when the treatment is based on one threshold value with one running variable and additional covariates. Researchers are usually interested in potential heterogeneity in the treatment effect to analyse the differences across sub-populations. This type of analysis is important to evaluate the effects of policies for two reasons. First, researchers and policy makers receive a more detailed picture about how the treatment effects different groups. Secondly, the scarcity of resources incentives decision makers for future policies to treat only those groups, where the expected treatment effects are the largest. These questions are not new, in fact there are many application, which repeats the main RD analysis with different sub-samples, defined by the researchers. [2] This method is prone to multiple testing problem and without correction it leads to invalid inference. According to my knowledge, there has not been any other paper, which allow discovering heterogeneity in the treatment effects based on covariates in a systematic way. The closest related paper on this topic is Hsu and Shen (2019) who develop *test*s for possible heterogeneity in the treatment effect (e.g., $H_0 : ATE = CATE$). Their proposed tests allow to reveal possible heterogeneity among individuals with different observed pre-treatment characteristics, however they leave the discovery of groups and the estimation of the conditional average treatment effect function as an open question. My contribution is proposing a data-driven machine learning method, that discover these sub-samples with different treatment effects, using additional features, without invalidating inference. The method helps to answer to the policy related questions, under some mild conditions. I assume that the identification strategy of the treatment effect is valid across all sub-samples and the magnitude of the treatment effect or its variance is different for the found sub-samples. 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 variables. The task of the algorithm

[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.

[2] See e.g., Pop-Eleches and Urquiola (2013). Hsu and Shen (2019) carries out a survey, where they find 17 paper, which uses RD design in top journals in economics (e.g. AER, QJE, JPE, ect) in 2015. They found that 15 out of 17 papers checks for heterogeneity and only 2 addresses the issue with interaction terms. The rests using sub-sample techniques without correcting for multiple testing.

is to find the relevant variables among many others without invalidating the inference due to searching over many possible combinations. The end result gives sub-samples with differences in the treatment effects or in their variances. Finally, this method also can be used as a robustness check for RD, in case of homogeneous treatment effect, the algorithm ends up with a unique treatment effect.

This paper also contributes to the literature of discovering heterogeneous treatment effect with machine learning methods. There is a growing number of papers[3], which uses supervised machine learning (ML) to explore heterogeneity in the treatment effect.[4] These papers combines methods of supervised machine learning[5] and the potential outcome framework to estimate causal effects. All of the below mentioned papers are used in a) randomized experiment and/or b) observational studies with unconfoundedness setup or for estimating c) local average treatment effects (LATE) in observational studies. Imai et al. (2013) uses lasso with two sparsity constraints on the treatment effect heterogeneity parameters and on the pre-treatment effect parameters to identify heterogeneous treatment effects. The idea is to formulate heterogeneity as a variable selection problem in randomized experiments or observational studies with unconfoundedness assumption. Athey and Imbens (2016) also focuses on randomized experiments or observational studies with unconfoundedness, but uses *honest* regression tree to find heterogeneity in the treatment effect. While growing the tree they use conditional means to differentiate the treatment effects from each other. With the honest approach they show that their method gives proper coverage for the treatment effects. Bargagli and Gnecco (2020) follows Athey and Imbens (2016) approach and extend to cases when assignment mechanism is irregular to estimate LATE. Finally there are papers, using random forests to estimate conditional treatment effects for randomized experiment, observational studies with unconfoundedness or to estimate heterogeneity in local average treatment effects. [6] Random forest is out of scope for this paper, but gives a possible extension for the future.

To the best of my knowledge, there is no paper, which considers CATE or CLATE in the quasi-experimental setup of the regression discontinuity design with regression trees. This paper provides an algorithm, which is applicable in these cases. The largest contribution to the causal machine learning literature is the implementation of the running variable and the estimation method inside the algorithm, which gives the base for the identification of the (conditional) treatment effect. The structure of the algorithm is the closest to Athey and Imbens (2016), but instead of estimating simple conditional means given the features, it estimates conditional polynomial regression – using the running variable – while conditioning on the features via regression tree. The difference in the identification strategy not only results in these differences in the algorithm, but in a new criterion as well, which governs the discovery of the proper

[3]E.g. Imai et al. (2013), Athey and Imbens (2016), Wager and Athey (2018), Athey et al. (2019) Bargagli and Gnecco (2020), Friedberg et al. (2020)

[4]Not to confuse with papers, which uses ML technique for estimating high-dimensional nuisance parameter, while the parameter of interest is the average treatment effect.

[5]Hastie et al., 2011 gives a great overview on classical supervised machine learning techniques

[6]E.g. 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 experiment or observational studies with unconfoundedness. Athey et al. (2019) and Friedberg et al. (2020) uses 'generalized random forests' as an adaptive weighting function to express heterogeneity. The main advantage of this method, that it is not prone to the curse of dimensionality as the other classical kernel weighting methods. Athey et al. (2019) proposes 'local' moment conditions – where locality is given by the forest – to estimate conditional local average treatment effect (CLATE). 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 experiment and shows an application with observational study with unconfoundedness assumption.

partitions or trees. With the newly proposed algorithm, one can achieve unbiased estimates for the grouped conditional average treatment effect and their variance. This paper only consider simple tree and not (causal) random forests, thus in cases when the CATE is a continuous function of the features, our method only provides step-approximation.

In addition to Monte Carlo simulations, I use this algorithm to explore heterogeneity in the Romanian school system. Pop-Eleches and Urquiola (2013) shows the average treatment effect on Baccalaureate examination of going to a better school and shows some ad-hoc heterogeneity analysis. I show that using the algorithm I can refine their results, discovering interesting treatment heterogeneity along school average transition scores and number of schools in towns. Furthermore, with a more extensive survey dataset with many socio-economic variables, but with less individuals, I find that gender of the student, education of the mother, availability of the internet or phone or proportion of novice teachers in school had also an effect on the intent-to-treat effect.

The paper is organized as follows: section 2. introduces the basic concepts of sharp RD, regression tree and defines the conditional average treatment effect for regression discontinuity tree. Section 3. shows the honest criterion for RD trees, which governs the discovery of the partitions. Section 4. overviews the specifics of the algorithm for RD trees along with some practical guidance on bandwidth and order of polynomial selection. Section 5. shows the Monte Carlo simulation results with sharp regression discontinuity design for linear and nonlinear in running variable cases. Section 6. demonstrates the usefulness of the algorithm on Pop-Eleches and Urquiola (2013) dataset. Section 7 extends our method to fuzzy RD designs. Section 8. concludes.

2 Regression Discontinuity Tree

With classical regression discontinuity design, researchers are interested in the causal effect of a binary treatment. Let Y be the outcome and using the potential outcome notation, $Y(1)$ denotes the potential outcome, when a unit gets the treatment and $Y(0)$ if no treatment takes place. The outcome corresponding to the treatment received or not can be written as:

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

Getting the treatment is connected to a scalar variable, called the *running variable*, which is denoted by X. Here, I only consider sharp RD[7], where the treatment D is determined solely by the value of the running variable, being on either side of a *fixed* and *known* threshold c, such

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

I am going to use the indicator function instead of D, to have a unified notation and to emphasize that it is referring to the RD setup and not to confuse with randomized experiment of observational studies with unconfoundedness.

Treatment heterogeneity comes in the form of additional characteristics. Let Z be a set of k random variable referring to the possible sources of heterogeneity. Following the machine learning terminology, call it *features*. A natural candidate for defining the conditional average treatment effect would be $\tau(z) = \mathbb{E}\left[Y(1) - Y(0) | X = c, Z = z\right]$, where this function can be continuous,

[7]For fuzzy design, see Section 7

discrete or a mixture in Z. However a simple regression tree does not allow such flexibility. Before defining the CATE function for RD tree, let me introduce some basic notion of regression trees, which highlight the fact, why one needs a different definition for CATE function for RD tree.

2.1 CATE in regression discontinuity tree

Regression trees – also sometimes referred as the partitioning scheme – allows to make a simple, intuitive and easy to interpret step-approximation for the underlying function. 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 and a split value: if observation has larger values for the selected feature than the split value it goes to the first sub-sample, otherwise to the second sample. 2) If needed repeat the split, but now use one of the already split sub-sample for the next candidate split. This way the feature space is partitioned into different complementary regions. These regions are called '*leaves*' or '*partitions*', noted by ℓ_j. A regression tree (Π) has $\#\Pi$ leaves: $j = 1, \ldots, \#\Pi$, which union gives back the complete feature space $\mathbb{Z}$,

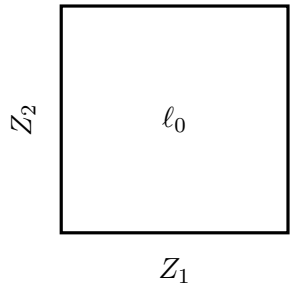
$$\Pi = \{\ell_1, \ldots, \ell_j, \ldots, \ell_{\#(\Pi)}\}, \quad \text{with} \qquad \bigcup_{j=1}^{\#\Pi} \ell_j = \mathbb{Z}$$

Leaf ℓ_j denotes the j'th partition of the feature space and decides which values of Z are part of that particular leaf.

For illustrative purposes, let use only two features Z_1 and Z_2. Figure 1. shows three different trees with two representation. 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.a) shows a tree, where there is no partitions, only one leaf ℓ_0, which contains all the units. This tree defines 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 average treatment effect is the same as the simple average treatment effect. Figure 1.b) 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 $\ell_2 : Z_1 > t_1$, where t_1 is the so called 'splitting value'. It is important, that the splitting values are always within the support of the referred feature, in this case $t_1 \in Supp(Z_1)$. In this case Z_2 does not affect the partitioning and considered irrelevant with respect to the treatment heterogeneity. Figure 1.c) shows a tree, when there are five different leaves, resulting in five different treatment effects depending on both values of Z_1 and Z_2. In this case if one wants to classify the treatment effect for a unit with given values for $Z_1 = z_1$ and $Z_2 = z_2$, one need 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 $t_3 > t_1$, $t_1, t_3 \in Supp(Z_1)$ and $t_2, t_4 \in Supp(Z_2)$).

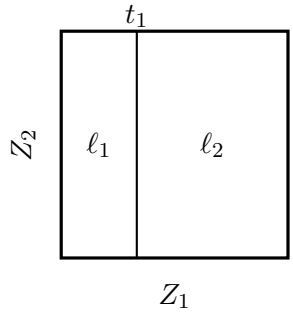
(1) **(2)**

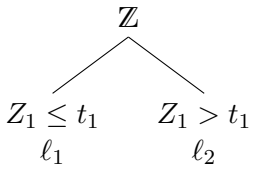
Partitioning **Tree structure**



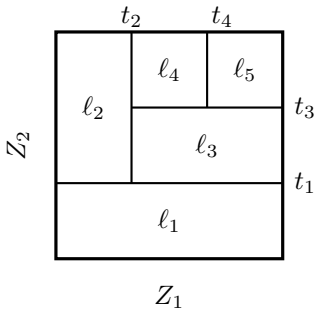
$\ell_0 = \mathbb{Z}$

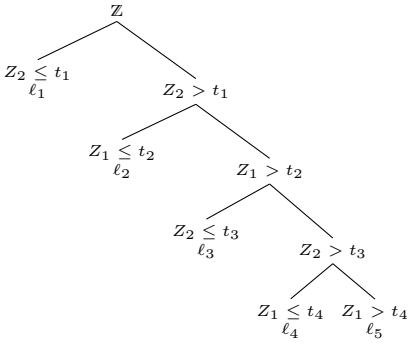
a) Π_0: homogeneous treatment effect, no partitioning





b) Π_1: two leaves, two treatment effects





c) Π_2: five leaves, five treatment effects

Figure 1: Different trees and their conditional average treatment effects

By the nature of recursive splitting of regression tree, it is not possible to construct an estimator for a continuous function, but only a step-approximation. Although this is a drawback of this methodology, on the other hand, the tree structure ensures an intuitive decision based interpretation.

. Until section refsec:criterion, let assume that the true tree Π is known. Using this known tree structure, the treatment effect for leaf ℓ_j can be defined as,

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

Let me use an indicator function, which takes the value of 1, if values of z is part of leaf ℓ_j and zero otherwise.

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

Now, it is possible to formulate the conditional average treatment effect function for regression discontinuity tree,

$$\tau(z;\Pi) = \sum_{j=1}^{\#\Pi} \tau_j \mathbb{1}_{\ell_j}(z;\Pi) \tag{2}$$

2.2 Identification of CATE in sharp RD

To identify the conditional average treatment effect function defined for regression trees in sharp RD, we need the following assumptions referring to each treatment effects given by the leaves (equation 1).

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 leaf in the tree.

(ii) In all leaves, for the density function $F(X=x|Z \in \ell_j(\Pi))$, $x=c$ is an interior point of the support.

Assumption (i) states that the means of the potential outcomes conditional on the running variable in each leaf is 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.[8] Assumption (ii) ensures that density for the running variable is well behaving: it has positive probabilities below and above the threshold value within each leaves. This excludes cases, when there are no realizations of the running variable around 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 $F(Z \in \ell_j(\Pi)|X=x)$ to be continuous in x [9] – 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. Otherwise it is needed. Let me note here, that features (Z) are pre-treatment variables, thus they should not have any effect on the value of the running variable. [10]

If these assumptions hold, the conditional average treatment effect for regression discontinuity

[8] But less restrictive if we assume continuity in $Z=z$ as in e.g. Hsu and Shen (2019).

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

[10] *Note:* Although the used conditional average treatment effect function here is a step-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 paper.

tree is identified as:

$$\begin{aligned}
\tau(z;\Pi) &= \sum_{j=1}^{\#\Pi} \tau_j \mathbb{1}_{\ell_j}(z;\Pi) \\
&= \sum_{j=1}^{\#\Pi} \left\{ \mathbb{E}\left[Y(1)|X=c, Z\in\ell_j(\Pi)\right] - \mathbb{E}\left[Y(0)|X=c, Z\in\ell_j(\Pi)\right] \right\} \mathbb{1}_{\ell_j}(z;\Pi) \\
&= \sum_{j=1}^{\#\Pi} \left\{ \lim_{x\downarrow c} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] - \lim_{x\uparrow c} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] \right\} \mathbb{1}_{\ell_j}(z;\Pi) \\
&= \mu_+(c,z;\Pi) - \mu_-(c,z;\Pi)
\end{aligned} \tag{3}$$

where,

$$\begin{aligned}
\mu_+(x,z;\Pi) &= \sum_{j=1}^{\#\Pi} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] \mathbb{1}_{\ell_j}(z;\Pi) \\
\mu_-(x,z;\Pi) &= \sum_{j=1}^{\#\Pi} \mathbb{E}\left[Y(0)|X=x, Z\in\ell_j(\Pi)\right] \mathbb{1}_{\ell_j}(z;\Pi)
\end{aligned} \tag{4}$$

refers to the conditional expectation function for μ_+: above the threshold (treated) and μ_-: below the threshold units (untreated in sharp RD).

2.3 Parametrization and estimation

The paper uses p'th order polynomial function of X to approximate each conditional expectation functions for each leaves – $\mathbb{E}\left[Y(D)|X=x, Z\in\ell_j(\Pi)\right], D\in\{0,1\}$. This provides a fairly flexible functional form in the running variable. [11] Here we show the parametrization of the conditional expectation function given by equation 4. First, adjust X by c, and let $\boldsymbol{X}'$ be a $1\times(p+1)$ vector such,

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

The conditional expectation function using polynomials with given Π,

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

where, $\boldsymbol{\delta}_j = [\alpha_j, \beta_1, \beta_2, \ldots, \beta_p]'$ is a $(p+1)\times 1$ parameter vector[12] and depends on the partitioning. Note that this definition allows for each leaves thus sub-samples to have different functional forms in X.

To estimate the conditional average treatment effect for regression discontinuity tree let consider a sample $\mathcal{S}$ consists $i=1,\ldots,N$ independent and identically distributed observations of

[11] 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, we recommend to use a restricted sample using a bandwidth, which is estimated on the whole sample.

[12] For RD the main parameter of interest is α_j, thus the difference, also β, should be β_1, j, but we neglect j subscript for convenience.

(Y_i, X_i, Z_i). The paper employs leaf-by-leaf estimation for the parameter vectors $\boldsymbol{\delta}_{+,j}$ and $\boldsymbol{\delta}_{-,j}$, using least squares, which has the advantage of relative fast estimation.[13] The estimator for the parameters are given by,

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

Using these parameter vectors and the identification equation for CATE (equation 3), the estimator for conditional average treatment effect for regression discontinuity trees 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. Notation 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.

3 Honest criterion for discovering RD tree

In this section, the assumption of a known tree is relaxed. Before I propose a unified criterion for regression discontinuity tree, let me introduce some of the basic notations and steps of classification and regression trees (CART).[14] For classical CART algorithms there is a distinction between two i.i.d. samples: training sample and test sample. In the training sample the algorithm is 'trained' – learning the patterns in the data, while in the test sample the performance or predictive power of the algorithm is tested. The CART criterion is constructed in a way that it gives back the required properties of the outcome on the test sample: smallest MSE value for outcomes or in our case unbiased estimate for CATE. In the training sample the algorithm 'learns' the patterns in the data in the following way: 1) it uses the data to grow a large tree with a given criterion. This usually results in a tree, which over-fits the data (over-fitting: less partitions is optimal in the test sample). 2) using this large tree, the algorithm searches for the optimal 'complexity parameter', which controls for over-fitting. The complexity parameter governs the magnitude of 'pruning': it adjust the criterion with the number of leaves grown on the tree. The larger the complexity parameter, the less leaves are in the tree (hence the name 'pruning'). The optimal complexity parameter is found by R-fold cross-validation and results in the lowest criterion value (averaged over the folds) - still using the training sample. This method results in an 'optimal' tree in the sense, it gives the smallest criterion value using only the test sample.

A known property of classical CART is as it grows a tree, spurious correlations between features and outcomes can affect the selected model, due to using only the training sample for grow-

[13] This is computationally 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. However 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.

[14] For a more detailed description see Breiman et al. (1984) or Hastie et al. (2011, Ch 9.).

ing and predicting (or estimating) the outcomes.[15] This leads to biases, which disappear only slowly as the sample size grows. This is not a big problem as long as the goal is prediction and not unbiased parameters. However, as the goal of this paper is to get an unbiased estimator for CATE, the classical method needs refinements. Athey and Imbens (2016) introduces the '*honest*' approach, which achieves unbiased estimates for conditional average treatment effects in experimental design or observational studies with unconfoundedness assumption. By their definition, a regression tree is called 'honest' if it does not use the same information for growing the tree as for estimating the given the tree. This means to split the training sample into two parts: one for constructing the tree (including the cross-validation step) and one for estimating the treatment effects within leaves of the 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 one can eliminate the biases. The cost of honest approach is the loss in precision – less observation used – due to sample splitting (Athey and Imbens, 2016, p. 7353-7354).[16].

In this paper, I propose a unified criterion for growing a regression discontinuity tree and for cross-validate to avoid over-fitting the data, while using the honest approach to get valid inference. Let introduce three *independent*[17] samples to derive the honest target for regression discontinuity tree. The first sample is the '*training sample*': $\mathcal{S}^{tr}$, which is used to find the proper tree Π^* from all the possible alternatives. This includes find proper splitting values, growing a large tree and usage of cross-validation to prune back excess leaves which results from over-fitting. The second sample is '*estimation sample*': $\mathcal{S}^{est}$. Using the estimation sample and take estimated tree Π^* as given, one get an unbiased estimator for CATE and for its variance, due to independence of the training and estimation sample. Finally, define the '*test sample*': $\mathcal{S}^{te}$, which is used to evaluate the selected tree and its unbiased estimates and compare it to the alternatives. I propose such criterion, which satisfies that the selected tree Π^* and its estimated CATE is the closest to the true (step-approximated) CATE function on the independent test sample.

Before introducing the criterion, there is a corollary assumption from the step-approximation nature of CATE, given a sample. Let call it a correct specification assumption for the conditional average treatment effect function,

Correct specification assumption

1. $\exists\, \Pi^{cs} | \mathcal{S} = \{\ell_1, \ldots, \ell_j, \ldots, \ell_{\#\Pi^{cs}}\}$
   such that,
2. $\ell'_j \subset \ell_j \Longrightarrow \tau'_j = \tau_j$
3. $\forall \ell_j, \ell_\beth,\ j \neq \beth \Longrightarrow \tau_j \neq \tau_\beth$

Correct specification assumption ensures that even though the true CATE function is continuous, 1. there exists a best step-approximation given by the tree Π^{cs} on a given sample. 2. this tree Π^{cs}

[15] In the initial splits the CART tend to group together observations, with similar, extreme outcomes. Thus after the training data have 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 spurious correlations leads to biases in the predictions/estimates.

[16] With the hones 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.

[17] Independence is a crucial assumption, when deriving the properties of the criterion. However, this independence assumption can be satisfied easily, by randomly assigning units to each sample.

is such that within leaf j, a subset noted by ℓ'_j has the same treatment value. 3. for all leaves, which are different from each other the treatment values are not the same.[18] This assumption is important to compare the grown partition to the correctly specified step-approximation and not the continuous CATE function.

A natural – but in-feasible criterion – for evaluating the regression discontinuity tree would be minimizing the mean squared error of the estimated conditional average treatment effect on the test sample with respect to the correctly specified tree, Π^{cs}. The estimated CATE uses the tree grown on $\mathcal{S}^{tr}$ and estimated on $\mathcal{S}^{est}$. Note, in this formulation, there is an extra adjustment term, $\tau^2(Z_i;\Pi^{cs})$ – an independent scalar from the optimization, thus does not have any effect on the result – that allows for theoretical derivations.

$$MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = \frac{1}{N^{te}} \sum_{i\in\mathcal{S}^{te}} \left\{ \left[\tau(Z_i;\Pi^{cs}) - \hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2 - \tau^2(Z_i;\Pi^{cs}) \right\} \tag{5}$$

Minimizing this criterion with respect to Π, would give a MSE-optimal regression discontinuity tree, however $\tau(z;\Pi^{cs})$ is unknown. Instead – following Athey and Imbens (2016) – we minimize the *expected* MSE over the test and estimation sample. This formulation has two advantages: a) it gives the best fitting tree for the *expected* estimation and test sample. This is favourable, while, when the tree is grown, both of these samples are unknown for the algorithm. b) using this formulation, a feasible criterion can be derived for the CATE. The expected MSE criterion is given by,

$$EMSE_\tau(\Pi) = \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}}\left[MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi)\right] \tag{6}$$

The goal of the algorithm is to construct a tree – grown on the training sample – which minimize this honest criterion.

$$\Pi^* = \arg\min_{\Pi} \left[EMSE_\tau(\Pi)\right]$$

Before showing the main results from deriving a feasible EMSE criterion, let me introduce some further assumptions, necessary for the derivation.

Assumptions for deriving estimator for EMSE criterion

1. Samples $\mathcal{S}^{tr},\mathcal{S}^{est},\mathcal{S}^{te}$ are independent from each other.
2. 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.
3. The share of units below and above the threshold within each leaf are the same for the estimation and test sample.

These assumptions are rather mild assumption: if there is large enough number of observations in the samples, then by randomly assigning observations to each samples satisfy these assumptions.

[18] This assumption can be relaxed with causal forest for RD.

Under these assumptions, the EMSE criterion (equation 6) can be decomposed into two parts – see the derivations in Appendix A,

$$EMSE_\tau(\Pi) = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi^{cs})\right] \tag{7}$$

The second part ($\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi^{cs})\right]$) is the workhorse of the algorithm: the expected value of the squared correctly specified CATE for the test sample or with different phrasing the average value of the squared treatment effects given by the leaves. This part decreases the EMSE value if the treatment effects are different in sub-samples (leaves). Furthemore, this term is increasing with the number of splits, therefore responsible for potential over-splitting of the true treatment effects. The first part is the expected value over the features in the test sample ($\mathbb{E}_{Z_i}\{\cdot\}$) for the variance of CATE ($\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]$) or put it differently: the averaged CATE variance in the test sample. Minimizing this part of the EMSE, results in a tree which provides the smallest averaged CATE variance, thus the less uncertainty in the estimator the better. This first part offsets the potential over-splitting coming from the second part: as the average value of the squared treatment effects are always increasing as the tree is larger, the number of observations within each leaf with each splitting is decreasing. This results in higher variance, increases the EMSE value during the minimization. It is also important to add that the expected variance part may lead the algorithm to discover sub-samples which has the same treatment effect, but they are different in their variances, resulting overall in a smaller EMSE value.

In the following I propose estimators for both parts. Here I only present the results, but see the complete derivations in Appendix B.
The expected variance term can be estimated as a weighted average of the variance estimators for the conditional expectation function for each leaf, evaluated at the threshold value. The weights are the share of units above and below the threshold within each leaf. This is an intuitive measure: it weights the uncertainty in the treatment effect in each leaf by the share of the observations. Furthermore, these weights incorporates the fact that the variance of the treatment is the same for each units within the leaf, thus there is no need to estimate the variance for each unit of observations. For simplicity, let assume that within each leaf, the disturbance term has the same finite variance (homoscedastic errors within each leaf).[19]

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

where, N^{est} is the number of observations in the estimation sample, $e_1 = [1, 0\ldots, 0]$ is a $1\times(p+1)$ selector-vector. Inside the square brackets everything refers to within leaf quantities for above $(+)$ or below $(-)$ the threshold. $\hat{\sigma}^2_{\pm,j}$ is the variance estimator and $M^{-1}_{\pm,j}$ is the inverse of the running variable's cross-product, both estimated on the test sample. $p^{est}_{\pm,j}$ is the share of units above $(+)$ or below $(-)$ the threshold within leaf j. The scalar N^{est} and the weights $p^{est}_{\pm,j}$ refers to the estimation sample, which comes from Assumption 2. and 3. for EMSE derivation. (For the detailed description of the notations see Appendix B.1.)

The estimator for the expected value of the squared correctly specified CATE (second part of equation 7.), uses variance decomposition. After some manipulation (see Appendix B.2), one gets

[19]Note: a) this only assumed for each leaf and not for the whole partitioning, thus disturbance terms for all leaves does not need to be homoscedastic. b) this assumption can be relaxed – see Appendix D – but here for simplicity we use the homoscedastic case.

the square of the expected treatment effects estimated on the test sample minus the average of the weighted variances.

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

The squared expectation part of the estimator leads the discovery of different treatment effects, it is the sample analogue for finding sub-samples with different treatment effects. This part always increases as $\#\Pi$ increases, while the average of the sum of squared treatment effect for two (or more) sub-samples is always smaller than the average of the sum of different squared treatment effects. The second part is similar to the derived expected variance, but here the scaling comes from the test sample, due to estimating on the test sample only. The weights refers to the estimation sample, utilizing Assumption 3. for EMSE. This is favourable, while then the only difference is in the scaling term.
Putting together the two parts, yields in the following estimable EMSE criterion for regression discontinuity trees,

$$\begin{aligned}\widehat{EMSE}_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{te}) \\ &+\left(\frac{1}{N^{te}}+\frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\left(\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}\right)}{p^{est}_{+,j}} + \frac{\left(\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}\right)}{p^{est}_{-,j}}\right]e_1\right\}\end{aligned} \tag{8}$$

Minimizing this criterion leads to splits, where a) treatment effects are different from each other in the sub-samples – first part; and b) the variance of the treatment effect for each leaf becomes smaller than the variance without a split – second part. Furthermore, using $p^{est}_{\pm}$ as weights in the second part, allows the criterion to avoid differences in variance estimators – coming from specificity of the used sample. This is beneficial, when using this criterion in the tree growing stage, the algorithm uses observations only from $\mathcal{S}^{tr}$, but with this additional knowledge on the share of observations within leaves from the estimation sample. This refinement of the algorithm helps to avoid finding extreme values only present in the training sample.

4 Algorithm specification

The implementation of growing a regression tree is based on the aforementioned two distinct stages: a) initial tree building, where an over-fitted/large tree is grown b) cross-validation to select the complexity parameter, which is used for 'pruning'. The second stage eliminates the over-fitting of the first stage. Each of these parts uses the introduced EMSE criterion. In this paper I focus on the specifics of the RD designs and use the building blocks for finding the best tree, described in (Hastie et al., 2011, Ch. 9), Athey and Imbens (2015) or Athey and Imbens (2016).

4.1 Criteria, estimation and bucketing

The main challange for the algorithm is during the tree-building phase to find the optimal splitting values for each feature. For this purpose – evaluating the splitting criterion – I only use the training sample and evaluates $\widehat{EMSE}_\tau(\mathcal{S}^{tr},\mathcal{S}^{tr},\Pi)$. Furthermore, I use a shortcut only considering the candidate splitting values for each feature, while the other parts of the criterion remains the same. As the algorithm consider each possible splits given the feature values the inversion of $M_{\pm,j}$ is computationally challenging. Instead calculating the inverse each time, I use

the Sherman-Morison formula. This recursive estimation enables to calculate the inverse only once per splitting candidate feature.[20] Following Athey and Imbens (2015), I use 'bucketing' of the observations as well, when considering a split. This ensures that each candidate split has enough treated and non-treated units. This is an important nuisance, because the criterion may vary too much without this technique. Bucketing ensures that there is no 'better' split value only due to adding one treated or untreated units, without the other.[21]

The complexity parameter governs the depth of the tree. I use a simple scalar complexity parameter (γ) that represents the cost of an additional leaf on the tree, $\widehat{EMSE}_\tau(\mathcal{S}, \mathcal{S}, \Pi) - \gamma \# \Pi$, which is a common method in the literature. The optimal complexity parameter gives the minimum cross-validation criteria, which is calculated using separate cross-validation (CV) samples $\widehat{EMSE}_\tau(\mathcal{S}^{tr,cv}, \mathcal{S}^{tr,tr}, \Pi)$. The final choice of the complexity parameter - similarly to the literature - can use the value, which gives the smallest cross-validation criteria or in case of a flat cross-validation criterion function the 1-standard-deviation rule, which can result in a smaller tree.

4.2 Bandwidth and polynomial selection

In RD the main focus is properly estimate a specific point in the conditional expectation function without being interested in the function itself or in the interpretability of the regression coefficients. For this reason non-parametric methods are usually preferred to estimate the average treatment effect. This paper only considers a parametric specification with respect to the running variable within each leaf. Even though higher order polynomials work fine in case of non-linear functional form in the running variable (see Monte-Carlo results in Section 5.), in some cases it can cause noisy estimates. Thus in practice I advise to use a pre-estimated (under-smoothing) bandwidth on the full sample, then restrict the used sample and employ the algorithm in this restricted sample.[22]

Another important question is selecting the order of polynomials used during the estimation. There is a recent discussion on this issue, where the main recommendations are to use low order (local) polynomials. (see e.g. Gelman and Imbens (2019) or Pei et al. (2020)) This paper offers a natural approach to select the order for the polynomials: use the cross-validation procedure jointly with the complexity parameter to select the order of polynomial as well.

5 Monte-Carlo simulations

For Monte Carlo simulation I created five different designs all mitigating different aspects of heterogeneous treatment effects in RD.

DGP-1: The first and simplest design mitigates a simple tree structure: there is two distinct treatment effects, given by two relevant groups and two irrelevant groups. It shows how the algorithm performs finding the true tree, given the noise in the data generating process (DGP).

- $\tau(Z_1 = 1) = 1, \tau(Z_1 = 0) = -1$

[20] Note that there is a trade-off: if there are multiple repetition 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.

[21] I improve the algorithm, proposed by Athey and Imbens (2016) – available at `https://github.com/susanathey/causalTree` – by carrying out the bucketing after the criterion is calculated and using the last valid split value instead of taking the average.

[22] An interesting research avenue would be to add bandwidth selection thus non-parametric estimators to the EMSE criterion. The tree growing could handle this extension naturally, but the bias-variance trade-off would alter, thus result in a different scaling.

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) and two continuous (Z_3, Z_4). This design mitigates the step-approximation behavior of the algorithm with heterogeneity in the expectation functions as well.

- $\tau(Z_3) = 2Z_3$

The last three designs use non-linear functional forms, proposed in Calonico et al. (2014), mitigating different RD applications. The value for the variance for the disturbance term is also used in Calonico et al. (2014) ($\sigma = 0.125 \longrightarrow \sigma^2 = 0.0.015625$). Furthermore, I add a heterogeneous treatment effects to the original simulation setup

DGP-3: Imitates Lee (2008) vote-shares application. I assume two treatment effects and heterogeneous conditional expectation function. I use 52 dummy variables representing political parties and states. The political party dummy (Z_1) is relevant and has an effect on both treatment and functional form. States are irrelevant.

- $\tau(Z_1 = 1) = 0.02, \tau(Z_1 = 0) = 0.07$

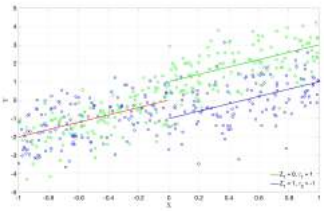
DGP-4: Imitates Ludwig and Miller (2007), who studied the effect of Head Start funding to identify the program's effects on health and schooling. I assume a continuous treatment effect based on the age of participants (Z_3), while adding (irrelevant) dummies representing the different continents.

- $\tau(Z_3) = -0.45 - Z_3$

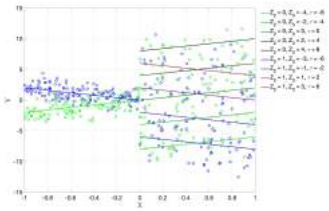
DGP-5: Is an alternative DGP by Calonico et al. (2014), which add extra curvature to the functional form. This design is exactly the same as in Calonico et al. (2014), thus there is only one homogeneous treatment effect.

- $\tau = 0.04$

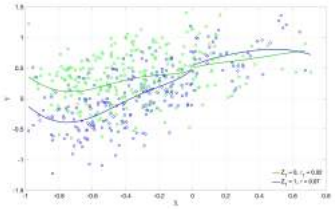
For the complete specification see the appendix, section E. Figure 2 shows the different sharp RD designs.



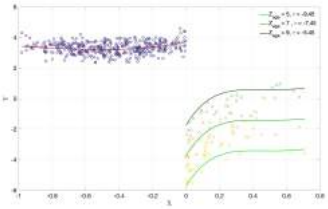
(a) DGP-1: Two treatment effects with homogeneous CEFs



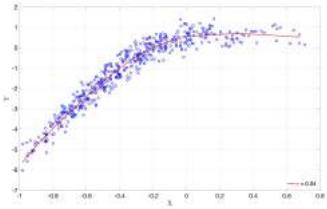
(b) DGP-2: heterogeneous CEF and continuous CATE



(c) DGP-3: Two treatment effects, with non-linear CEF



(d) DGP-4: Continuous CATE, with non-linear CEF



(e) DGP-5: Homogeneous treatment effect, with non-linear CEF

Figure 2: Simulation designs. DGP-1 and 2 uses $\sigma^2\epsilon = 1$, while DGP-3, 4 and 5 uses $\sigma_\epsilon = 0.125$ (similarly to the referenced papers).

During the simulations I am using a moderate sample size $N^{all} = 10,000$, from which the training and estimation sample contains $N^{tr} = N^{est} = 5,000$ observations. For evaluating the criterion, I create the test sample with $N^{te} = 10,000$ to minimize sampling variation. This is used after the tree is grown and the CATE is estimated. I also calculate and report the infeasible criterion based on equation (5) using this test sample. During the simulations $1,000$ Monte Carlo sample is used, where the variation comes only from the disturbance term.
Table 1. reports the infeasible MSE value along with the Monte Carlo averages for the number of leaves and the percentage of DGP found. The latter one is only feasible, when the DGP has a tree-structure: for DGP-1, 3 and 5. The other two is only a step-approximation for the continuous CATE. (Thus in these cases the more leaf we have, the better approximation is provided by the algorithm.)

DGP	noise	inf. MSE	#Π	DGP found
DGP-1	$\sigma^2 = 0.1$	0.0003	2.1580	0.8580
	$\sigma^2 = 1$	0.0032	2.0440	0.9580
	$\sigma^2 = 10$	0.0337	2.0270	0.9690
DGP-2	$\sigma^2 = 0.1$	0.6692	14.7330	-
	$\sigma^2 = 1$	0.7679	13.8670	-
	$\sigma^2 = 10$	1.3126	11.7060	-
DGP-3	$\sigma = 0.05$	0.0001	2.0000	1.0000
	$\sigma = 0.125$	0.0008	1.2370	0.2370
DGP-4	$\sigma = 0.05$	0.3135	4.8850	-
	$\sigma = 0.125$	0.3030	5.1050	-
DGP-5	$\sigma = 0.05$	0.0000	1.0430	0.9570
	$\sigma = 0.125$	0.0002	1.0170	0.9830

Table 1: Monte Carlo averages, using HCE-1, smallest CV value, $p = 1$ for DGP 1 and 2 and $p = 5$ for DGP 3,4 and 5

Table 1. shows that the algorithm works considerably well, it finds the DGP, when it has a tree structure more than 95% of the time. One exception is DGP-1, $\sigma^2 = 0.1$, where the cross-validation criteria function is rather flat. In this case with 1-standard deviation cross-validation rule the DGP is found in every cases. (See appendix, section F.) When the CATE function is continuous and there is only moderate noise, we got many leaves, which provides support for step-approximation of the continuous CATE function. The criterion works well not only in the linear case, but in non-linear cases as well, when the noise is not too large compared to the treatment effects. In the continuous treatment case it provides even less leaves which is again the result of the increased variance part in the EMSE, thus our algorithm tend to over-prune in these cases.[23]

An other important result is the Monte Carlo averages for the estimated conditional average treatment effects and their standard errors. Table 2 reports the Monte Carlo average of the treatment estimates for a given leaf $\left(\bar{\hat{\tau}}\right)$, the Monte Carlo average of the estimated standard errors for the treatment $\left(\overline{SE\left[\hat{\tau}\right]}\right)$ and the Monte Carlo standard deviation of the estimated treatment effect $(SD\left[\hat{\tau}\right])$. In case of continuous CATE, five distinct point is chosen from the relevant variable.[24]

[23]Note, as the pruning is a separate step in case of possible over-pruning the researcher can use a more lenient complexity parameter to prune the tree. It is advised to check the cross-validation criterion function before deciding on which rule to use to select the complexity parameter.

[24]For continuous irrelevant variables we choose 100 random points, get the CATE for each of them and average over them.

DGP 1	true value	Leaf 1: $Z_1 = 1$, $\tau_1 = 1$			Leaf 2: $Z_1 = 0$, $\tau_2 = -1$		
	estimates	$\overline{\hat{\tau_1}}$	$\overline{SE\left[\hat{\tau_1}\right]}$	$SD\left[\hat{\tau_1}\right]$	$\overline{\hat{\tau_2}}$	$\overline{SE\left[\hat{\tau_2}\right]}$	$SD\left[\hat{\tau_2}\right]$
	$\sigma = 0.1$	1.0001	0.0188	0.0175	-0.9997	0.0180	0.0178
	$\sigma = 1$	1.0003	0.0575	0.0555	-0.9991	0.0563	0.0563
	$\sigma = 10$	0.9991	0.1809	0.1821	-0.9950	0.1777	0.1823
DGP 2	true value	Leaf 1: $Z_3 = -4$, $\tau(-4) = -8$			Leaf 2: $Z_3 = -2$, $\tau(-2) = -4$		
	estimates	$\overline{\hat{\tau_1}}$	$\overline{SE\left[\hat{\tau_1}\right]}$	$SD\left[\hat{\tau_1}\right]$	$\overline{\hat{\tau_2}}$	$\overline{SE\left[\hat{\tau_2}\right]}$	$SD\left[\hat{\tau_2}\right]$
	$\sigma^2 = 0.1$	-7.8416	0.1066	0.3110	-4.1587	0.1409	0.2823
	$\sigma^2 = 1$	-7.8858	0.1705	0.2924	-4.2376	0.1879	0.2905
	$\sigma^2 = 10$	-7.8626	0.4684	0.6265	-4.3864	0.4193	0.8544
	true value	Leaf 3: $Z_3 = 0$, $\tau(0) = 0$			Leaf 4: $Z_3 = 2$, $\tau(2) = 4$		
	estimates	$\overline{\hat{\tau_3}}$	$\overline{SE\left[\hat{\tau_3}\right]}$	$SD\left[\hat{\tau_3}\right]$	$\overline{\hat{\tau_4}}$	$\overline{SE\left[\hat{\tau_4}\right]}$	$SD\left[\hat{\tau_4}\right]$
	$\sigma^2 = 0.1$	0.0386	0.1315	0.2030	4.8247	0.1021	0.0658
	$\sigma^2 = 1$	-0.0330	0.1953	0.3745	4.7359	0.1696	0.3361
	$\sigma^2 = 10$	-0.0595	0.4390	0.9167	4.0316	0.3983	0.9172
	true value	Leaf 5: $Z_3 = 4$, $\tau(2) = 8$					
	estimates	$\overline{\hat{\tau_5}}$		$\overline{SE\left[\hat{\tau_5}\right]}$		$SD\left[\hat{\tau_5}\right]$	
	$\sigma^2 = 0.1$	8.3832		0.1322		0.1328	
	$\sigma^2 = 1$	8.2082		0.1850		0.2205	
	$\sigma^2 = 10$	7.9984		0.3924		0.3638	
DGP 3	true value	Leaf 1: $Z_1 = 0$, $\tau_1 = 0.07$			Leaf 2: $Z_1 = 1$, $\tau_2 = 0.02$		
	estimates	$\overline{\hat{\tau_1}}$	$\overline{SE\left[\hat{\tau_1}\right]}$	$SD\left[\hat{\tau_1}\right]$	$\overline{\hat{\tau_2}}$	$\overline{SE\left[\hat{\tau_2}\right]}$	$SD\left[\hat{\tau_2}\right]$
	$\sigma = 0.05$	0.0701	0.0082	0.0082	0.0195	0.0084	0.0083
	$\sigma = 0.125$	0.0557	0.0174	0.0186	0.0432	0.0175	0.0212
DGP 4	true value	Leaf 1: $Z_{age} = 5$, $\tau(5) = -5.45$			Leaf 2: $Z_{age} = 6$, $\tau(6) = -6.45$		
	estimates	$\overline{\hat{\tau_1}}$	$\overline{SE\left[\hat{\tau_1}\right]}$	$SD\left[\hat{\tau_1}\right]$	$\overline{\hat{\tau_2}}$	$\overline{SE\left[\hat{\tau_2}\right]}$	$SD\left[\hat{\tau_2}\right]$
	$\sigma = 0.05$	-5.8180	0.0410	0.0152	-6.3044	0.0442	0.0158
	$\sigma = 0.125$	-5.8169	0.0548	0.0424	-6.3033	0.0579	0.0409
	true value	Leaf 3: $Z_{age} = 7$, $\tau(7) = -7.45$			Leaf 4: $Z_{age} = 8$, $\tau(8) = -8.45$		
	estimates	$\overline{\hat{\tau_3}}$	$\overline{SE\left[\hat{\tau_3}\right]}$	$SD\left[\hat{\tau_3}\right]$	$\overline{\hat{\tau_4}}$	$\overline{SE\left[\hat{\tau_4}\right]}$	$SD\left[\hat{\tau_4}\right]$
	$\sigma^2 = 0.05$	-7.8591	0.2076	0.0886	-7.8591	0.2076	0.0886
	$\sigma^2 = 0.1295$	-7.7760	0.1679	0.1410	-7.7760	0.1679	0.1410
	true value	Leaf 5: $Z_{age} = 9$, $\tau(9) = -9.45$					
	estimates	$\overline{\hat{\tau_5}}$		$\overline{SE\left[\hat{\tau_5}\right]}$		$SD\left[\hat{\tau_5}\right]$	
	$\sigma^2 = 0.05$	-9.1891		0.0388		0.0168	
	$\sigma^2 = 0.1295$	-9.1857		0.0555		0.0487	
DGP 5	true value	Homogeneous Treatment, $\tau = 0.04$					
	estimates	$\overline{\hat{\tau}}$		$\overline{SE\left[\hat{\tau}\right]}$		$SD\left[\hat{\tau}\right]$	
	$\sigma = 0.05$	0.0398		0.0059		0.0056	
	$\sigma = 0.125$	0.0398		0.0152		0.0146	

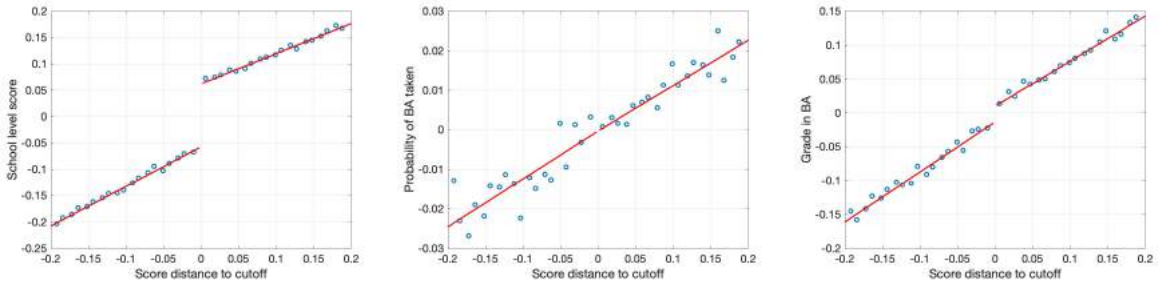
Table 2: Estimated Monte Carlo averages and standard deviation of parameters for treatment and standard errors by each leaf

Table 2 shows that the (averaged) point estimates reflects well the true treatment effects in general, and if the DGP is found the true value lies within $\pm 1SE$. During the estimation I use heteroscedastic consistent standard errors within each leaf (HCE-1[25]) and report the Monte Carlo averages. In cases, when DGP has a tree structure and the algorithm found the DGP, the average SE and the MC standard deviation of treatment effect are the same. In case of continuous CATE, the reported average of SEs and MC standard deviation of the treatments somewhat miss-leading, while with the regression tree one can only do a step approximation of the continuous CATE and the $SE\left[\hat{\tau}\right]$ refers to the whole interval, not only for the chosen specific point estimator.

6 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.
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 preference for schools. 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 observation) and a survey dataset (more than 6,000 observation) from Romania to study the impact of attending a more selective high school during the period of 2003-2007. Using the administrative dataset, they find that attending a better school significantly improves a student's performance on the Baccalaureate exam, but does not affect the exam take-up rate.

Figure 3 summarizes the classic mean RD results from Pop-Eleches and Urquiola (2013). In all three graphs the x-axis represents the running variable, which is a student's standardized transition score subtracting the school admission cut-off. The y-axis in the left graph represents the *peer quality* students experience, measured by the average transition score at their respective school. The middle graph shows the probability of a student taking the Baccalaureate exam, while the right graph 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 y-axis is centered around 0 for all plotted outcome. Both the left and the right graphs shows a jump in the average outcome at the discontinuity point, but the jump in the exam-taking rate is quite noisy and seemingly insignificant.



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

Figure 3: Bin-scatter for main (pooled) RD results of Pop-Eleches and Urquiola (2013), using school fixed effects

[25] See MacKinnon and White (1985)

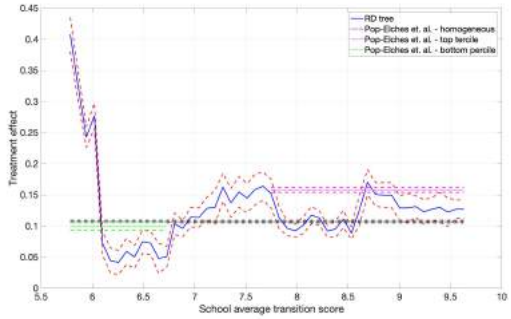
First, I revisit Pop-Eleches and Urquiola (2013) heterogeneity analysis on the intent-to-treat effects using peer quality (level of school average transition score) and the number of schools in town as the sources of heterogeneity, using the same administrative data between 2003 to 2005. Similarly, I restrict the sample to observations, which lies within the ± 0.1 interval for the running variable and use the same linear specification in the running variable. Pop-Eleches and Urquiola (2013) inspect heterogeneity in the treatment effect with pre-specified sub-samples: those who are in the top or bottom tercile for the the student's transition score, or to selected towns sub-samples: a) four or more schools in the towns, b) three schools or c) two schools only. Instead of using pre-specified sub-samples, I use the algorithm to find the relevant sub-samples. I also use this two variables[26] to explore the heterogeneity, but use them simultaneously allowing for finding different treatment effects in the student's transition score and in the number of schools in towns. This means that in competitive environment (high transition score) treatment effects can be different in towns with large or small number of schools. Figure 4 shows these marginalized treatment effects[27] along with treatment values reported in Pop-Eleches and Urquiola (2013). Left column shows the treatment effects conditional on level of school average transition score[28], while the right column conditioning on the number of schools.

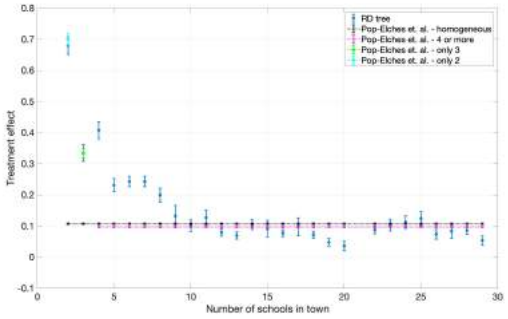
[26]For number of towns I add dummy variables as well to easily search for certain number of schools within the towns, not only using number of schools in town variable.

[27]I have calculated the treatment effect for each students than averaged over the non-plotted variable, thus in case of level of school average transition score, I take students with the same school average transition score and averaged them over the number of schools variables, reporting this average.

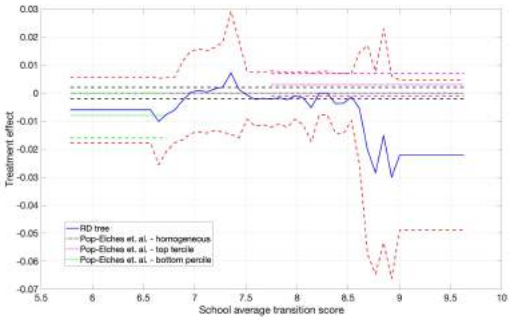
[28]I used 50 equal sized bins to group school average values.

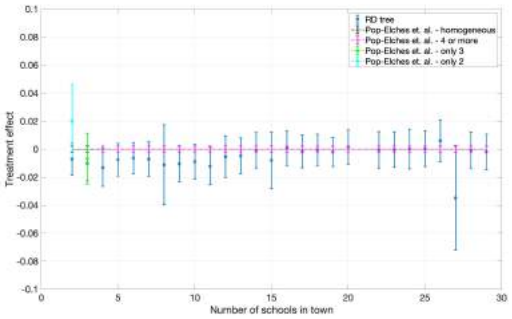
(1)	(2)
Level of school avg. transition score	**Number of schools**



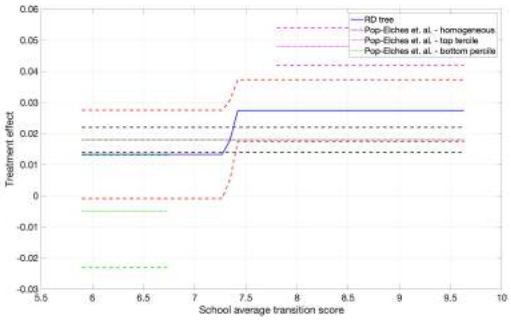


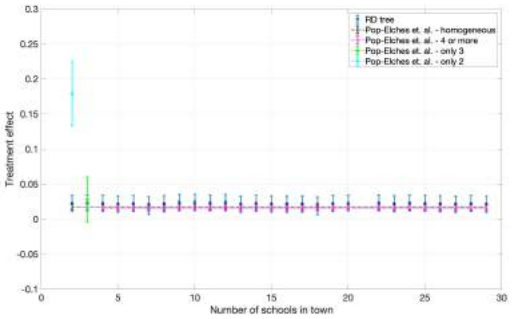
a) Schools average transition score effect





b) Probability of taking Baccalaureate effect





c) Effect on Baccalaureate grade

Figure 4: CATE for different outcomes, using school fixed effects

The treatment effects are aligned with the results of Pop-Eleches and Urquiola (2013), but shows a more sophisticated relation. On level of school average transition score effect – Figure 4 a, I find a clear indication of continuous treatment effect along with both features. The conditional effect is the largest for schools with low average transition scores and towns with few schools. For the probability of taking the Baccalaureate exam, the results are similarly noisy and seems insignificant if I use only the marginalized treatment effects given by Figure 4 b). However, looking at the tree structure, there is a unique treatment group, given by a combination of level of average transition score and number of schools, who's treatment effect is much lower than the average. This group has has more than 8.6 school average transition score - top 10% - and there are more than 15 schools in town. These are one of the most competitive schools in large towns,

where – after taking school fixed effect – there is lower participation rate on the Baccalaureate exam: the treatment effect for this group is -0.06 and its standard error is 0.014. This is aligned with the negative peer effect that Pop-Eleches and Urquiola (2013) reports. Finally the found heterogeneous effects for Baccalaureate grade is the simplest, if the school average transition score is above the median (7.4 is the 44'th percentile), students can expect a larger grade if going to a better school. Note that these effects are not significantly different at 95% from the homogeneous treatment effect.
Overall there is significant heterogeneity in the intent-to-treat effects, and with the help of the algorithm, one can identify these sub-populations in a systematic way.
Finally, I use Pop-Eleches and Urquiola (2013) survey dataset, which contains less observations, but contains a rich variety of socio-economic factors (geneder, ethnicity, education, accessibility of internet, phone, ect.) and school and study behavior specific questions (parents pay for tutoring, parents helps students, child does homework every day, peer ranking, teacher characteristics, ect.). In the survey there are only 135 schools located in 59 towns with 2 to 4 schools and the year is 2005 to 2007. Overall I use 32 different features to search for heterogeneity. As the survey corresponds to later years, the data includes only observations on schools average transition scores, but on the other two outcomes.

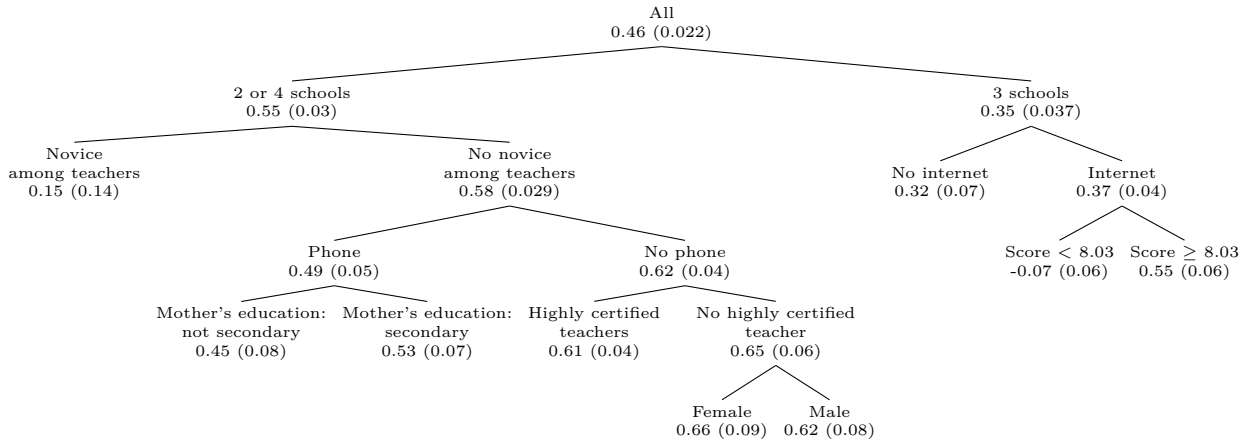


Figure 5: Exploring heterogeneous groups for peer quality - intent-to-treat effects, standard errors are clustered on student level.

The fitted tree (see Figure 5 suggests an informative result: for towns with 2 or 4 schools, admitted students (on average) has 0.55 higher scores. If one go further the tree suggests that having a novice among teachers (less than 2 years of experience) the treatment effect may disappear (note, only 319 observations fall into this category). It is interesting that in 2005-07 having a phone would somewhat reduce the peer quality effect. During this time it is not common to have phones among 12-14 years old. We find also interesting splits such as education of the mothers, or if there are teachers with highest state certification in the school and if the students gender is male or female.[29] These are indeed interesting splits, but statistically non-distinguishable. Heterogeneity with respect to towns, if there are three schools, the accessibility of the internet seems important. Although the effect on that level is the same, with an additional split, one can find an interesting result: the level of transition scores matter for these groups. For schools, which has students score above the first tercile (8.03 is the 35% percentile to be exact), the peer quality effect is similar to 2 or 4 schools students (0.55 higher scores). However

[29] As the cross-validation criterion is quite steep with the 1-SD rule, these splits are pruned back.

for students in these schools with internet access, below the first tercile, the peer quality effect is insignificant. This effect suggest potential segregation of the discovered group.

7 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, but allows for a smaller jump in the probability of assignment to the treatment at the threshold. This means that there are compliers, always takers and never takers during the observations. Let T, be the treatment variable - whether a unit gets indeed treated or not. To identify treatment effect for compliers or the conditional local average treatment effect (CLATE) in regression discontinuity tree, the following assumptions are needed

Identifying assumptions of CATE in fuzzy RD

1. $\lim_{x\downarrow c} \mathbb{P}\left[T(1)|X=x\right] \geq \lim_{x\uparrow c} \mathbb{P}\left[T=1|X=x\right]$
   (a) There is a positive discontinuity in the probability of taking the treatment around the threshold.
   (b) This is not only an assumption, but a built-in restriction for the algorithm: if this condition's sample analogue is not satisfied, it is not a valid split.
2. $\mathbb{E}\left[Y(d)|T(1)-T(0)=d', X=x, Z\in\ell_j(\Pi)\right]$ are continuous in $x=c$ for all pairs of $d,d'\in\{0,1\}$ and exists $\forall j$ and $z\in \text{Supp}[Z]$.
   (a) Continuity of average potential outcomes for always-takers, compliers and never-takers with respect to the running variable within each leaf.
3. $\mathbb{P}\left[T(1)-T(0)=1|X=x, Z\in\ell_j(\Pi)\right]$ and $\mathbb{P}\left[T(1)-T(0)=d|X=x, Z\in\ell_j(\Pi)\right]$ are continuous in x for $d\in\{0,1\}$, $\forall j$ and exists for $z\in\text{Supp}[Z]$.
   (a) Continuity of the probability of an individual to belong any of the (above mentioned) group within all leaves.
4. $\mathbb{E}\left[T(1)-T(0)|X=c, Z\in\ell_j(\Pi)\right]>0\,,\ \forall z$
   (a) No non-trivial presence of compliers or defiers along with assumption 1.
5. In all leaves, for the density function $F(X=x|Z\in\ell_j(\Pi))$, $x=c$ is an interior point of the support.

Under these assumptions, the CLATE for RD tree is identified as,

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

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,

$$\mu^t_+(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}^t_{-,j} \ , \quad \mu^t_-(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}^t_{+,j} \ , \quad \boldsymbol{\delta}^t_{j,\pm} = \left[\alpha^t_{j,\pm}, \beta^t_{j,1,\pm}, \ldots, \beta^t_{j,p,\pm}\right]'$$

$$\mu^y_+(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}^y_{-,j} \ , \quad \mu^y_-(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}^y_{+,j} \ , \quad \boldsymbol{\delta}^y_{j,\pm} = \left[\alpha^y_{j,\pm}, \beta^y_{j,1,\pm}, \ldots, \beta^y_{j,p,\pm}\right]'$$

The sample estimates for fuzzy design is provided in appendix, section C.
Using the same logic to find the proper (step-approximated) 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 sample estimator for fuzzy designs is given by,

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

where,

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

is the within leaf variance of the outcome equation at the threshold, estimated from above (+) or below (-) and $\hat{\tau}^t_j(\cdot), \hat{\tau}^y_j(\cdot)$ are the j'th leaf treatment effect estimated on the participation equation ($\hat{\tau}^t_j(\cdot)$) and on the outcome equation ($\hat{\tau}^y_j(\cdot)$). See the derivations in the Appendix, Section C.
The EMSE criterion for the fuzzy design combines the jumps in the outcome and in the participation equation along with the variance. It means if there is a difference in two subgroups in the participation probabilities at the threshold or in the outcome equation then the EMSE criterion will have a lower value, thus the algorithm will pick this difference. Also if the variance of the outcome equation at the threshold gets lower by a split the EMSE criterion will be lower, thus even if there is no big changes in the LATE value, but in the variances, the algorithm is doing a split. One feature of this criterion if the changes in the jump in the outcome equation and in the participation equations are with the same magnitude, then the EMSE criterion does not changes. This is due to the fact the LATE does not changes either in this case. 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 two trees.

8 Conclusions

This paper proposes an algorithm, which uncover treatment effect heterogeneity in classical regression discontinuity (RD) designs. The introduced honest regression discontinuity tree algorithm identifies heterogeneity in the treatment effect across sub-samples using additional features or covariates. The relevant sub-samples are to be discovered with by algorithm, without invalidating inference. The paper contributes to two literature: for the regression discontinuity designs with the estimate of conditional average treatment effects (CATE) with potentially many covariates with one running variable and with a known threshold value. For the causal machine learning literature with a new criterion to estimate causal effects, where the effects are estimated by leaf-by-leaf polynomial regressions and not by conditional means.

The properties of the CATE function for sharp regression design is analysed in details and the paper shows the strength and weaknesses of the step-approximation nature of regression trees. The criterion which leads to feasible CATE function for the regression discontinuity tree is shown in a separate section discussing how the algorithm finds the true (correctly specified) tree.
Monte Carlo simulation results show that the proposed algorithm and criterion works fine and discovers the true tree more than 95% of the cases. The estimated conditional treatment effects - if the true tree is found - are unbiased and the standard errors give proper estimates for coverage.
To show 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 confirms the results for the heterogeneity analysis done by Pop-Eleches and Urquiola (2013), but extends their results with a more refine conditional treatment effect. Using more features – in a smaller dataset – I have found interesting patterns on the treatment effect for peer quality in the 2005-2007 sample. One of the discovered groups is students with low score on the national assessment test in towns with three schools with internet access. The conditional treatment effects shows that there is no positive peer quality effect for this groups, while other groups have statistically significant positive peer quality effect. This result may show a sign for segregation: low scored students in such environment goes to the same school.

References

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

Athey, S. and Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *Stat*, 1050(5).

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

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

Becker, S. O., Egger, P. H., and von Ehrlich, M. (2013). Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy*, 5(4):29–77.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Caetano, C., Caetano, G., and Escanciano, J. C. (2017). Over-identified regression discontinuity design. *Unpublished, University of Rochester*.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., and Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, pages 1–15.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, 37(3):447–456.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Hsu, Y.-C. and Shen, S. (2019). Testing treatment effect heterogeneity in regression discontinuity designs. *Journal of Econometrics*, 208(2):468–486.

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

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Ludwig, J. and Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122(1):159–208.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Pei, Z., Lee, D. S., Card, D., and Weber, A. (2020). Local polynomial order in regression discontinuity designs. Working Paper 27424, National Bureau of Economic Research.

Pop-Eleches, C. and Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, 103(4):1289–1324.

Robson, M., Doran, T., Cookson, R., et al. (2019). Estimating and decomposing conditional average treatment effects: the smoking ban in england. Technical report, HEDG, Department of Economics, University of York.

Toda, T., Wakano, A., and Hoshino, T. (2019). Regression discontinuity design with multiple groups for heterogeneous causal effect estimation. Technical report, arXiv.

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

Xu, K.-L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics*, 201(1):1–18.

Appendix

A Decomposition of EMSE criterion

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

$$
\begin{aligned}
EMSE_\tau(\Pi) &= \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}}\left[MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi)\right] \\
&\quad \text{using } \mathcal{S}^{te} \perp\!\!\!\perp \mathcal{S}^{est} \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i;\Pi^{cs}) - \hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2 - \tau^2(Z_i;\Pi^{cs})\right\} \\
&\quad \text{using law of iterated expectations} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(Z_i;\Pi^{cs}) - \hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right)^2 - \tau^2(Z_i;\Pi^{cs}) \mid Z_i\right]\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est}) \mid Z_i\right] - \hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right)^2 \mid Z_i\right] - \tau^2(Z_i;\Pi^{cs})\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est}) \mid Z_i\right] - \tau^2(Z_i;\Pi^{cs})\right\} \\
&\quad \text{using } Z_i \perp\!\!\!\perp \mathcal{S}^{est} \text{ with more expressive notation, } \big|_{z=Z_i} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi^{cs})\right] \qquad \blacksquare
\end{aligned}
$$

B Derivation of honest sharp RDD criterion

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

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

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

B.1 Expected variance of CATE

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

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

where $e_1 = [1, 0\ldots, 0]$ is still a $1 \times (p+1)$ selector-vector. Because $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{est}_{+,j}\right]$ and $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{est}_{-,j}\right]$ 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,

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

where,

$$\begin{aligned}
N^{est}_{+,j} &= \sum_{i\in\mathcal{S}^{est}} \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \ , \qquad N^{te}_{+,j} = \sum_{i\in\mathcal{S}^{te}} \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \\
\hat{M}_{+,j} &= \frac{1}{N^{te}_{+,j}} \sum_{i\in\mathcal{S}^{te}} \boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \\
\hat{\sigma}^2_{+,j} &= \frac{1}{N^{te}_{+,j}-p-1} \sum_{i\in\mathcal{S}^{te}} \left(Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\delta}}^{te}_{+,j}\right)^2 \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \ , \\
j &= 1,2,\ldots,\#\Pi
\end{aligned}$$

Same applies for the components of $\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{est}_{-,j}\right]$, 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, we 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\}\\
&\text{using Assumption 2. for EMSE:}\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^{te}_j/N^{te}).

B.2 Expected square of CATE

The second part of the EMSE criterion is the estimator for the expected squared of the true CATE, $\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi^{cs})\right]$ over the test sample's features.
A natural estimator for $\tau(z;\Pi^{cs})=\mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}^2(z;\Pi,\mathcal{S}^{te})\right]$, taking z as given and using the assumption for correctly specified tree. Using the law of iterated expectation,

$$\begin{aligned}
\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi^{cs})\right] &= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}^2(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\}\\
&\text{using variance decomposition}\\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}^2_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}-\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\}\\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}^2_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\}-\mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\}
\end{aligned}$$

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 the scaling factor for number of observations are coming purely from the test sample.

$$\begin{aligned}\hat{\mathbb{E}}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{te}_{+,j}}+\frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{te}_{-,j}}\right]e_1\right\}\\ &\text{using Assumption 3. for EMSE: } p^{te}_{+,j}\approx p^{est}_{+,j},\ p^{te}_{-,j}\approx p^{est}_{-,j},\ \forall j\\ &= \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}}+\frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}\end{aligned}$$

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 3. for deriving the EMSE condition 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,

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

B.3 Estimator for EMSE

Plugging the two parts together yields an estimator for the EMSE criterion,

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

C Derivation of honest fuzzy RDD leaf-by-leaf LS criterion

Let assume, that there is a sample $\mathcal{S}$, with $i=1,\dots,N$ with identically and independently distributed observations of (Y_i,X_i,T_i,Z_i). For leaf-by-leaf estimation, we use that $\mathbb{1}_{\ell_j}(z;\Pi)$ creates disjoint sets, and we can estimate the parameters and their variances consistently in each leaf separately. The conditional mean estimator is given by,

$$\begin{aligned}\hat{\mu}^t_+(x,z;\Pi,\mathcal{S}^{est}) &= \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\hat{\boldsymbol{\delta}}^{t,est}_{+,j} \quad , \quad \hat{\mu}^t_-(x,z;\Pi,\mathcal{S}^{est}) = \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\hat{\boldsymbol{\delta}}^{t,est}_{-,j}\\ \hat{\mu}^y_+(x,z;\Pi,\mathcal{S}^{est}) &= \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\hat{\boldsymbol{\delta}}^{y,est}_{+,j} \quad , \quad \hat{\mu}^y_-(x,z;\Pi,\mathcal{S}^{est}) = \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\hat{\boldsymbol{\delta}}^{y,est}_{-,j}\end{aligned}$$

where, $\hat{\boldsymbol{\delta}}^{t,est}_{+,j}$, $\hat{\boldsymbol{\delta}}^{t,est}_{-,j}$, $\hat{\boldsymbol{\delta}}^{y,est}_{+,j}$ and $\hat{\boldsymbol{\delta}}^{y,est}_{-,j}$ solve for,

$$\hat{\boldsymbol{\delta}}^{t,est}_{+,j} = \arg\min_{\boldsymbol{\delta}^t_{+,j}}\sum_{i\in\mathcal{S}^{est}}\mathbb{1}_c(x)\mathbb{1}_{\ell_j}(z;\Pi)\left(T_i-\boldsymbol{X}_i'\boldsymbol{\delta}^t_{+,j}\right)^2 \quad , \quad \hat{\boldsymbol{\delta}}^{t,est}_{-,j} = \arg\min_{\boldsymbol{\delta}^t_{-,j}}\sum_{i\in\mathcal{S}^{est}}\mathbb{1}_c(x)\mathbb{1}_{\ell_j}(z;\Pi)\left(T_i-\boldsymbol{X}_i'\boldsymbol{\delta}^t_{-,j}\right)^2$$

$$\hat{\boldsymbol{\delta}}^{y,est}_{+,j} = \arg\min_{\boldsymbol{\delta}^y_{+,j}}\sum_{i\in\mathcal{S}^{est}}\mathbb{1}_c(x)\mathbb{1}_{\ell_j}(z;\Pi)\left(Y_i-\boldsymbol{X}_i'\boldsymbol{\delta}^y_{+,j}\right)^2 \quad , \quad \hat{\boldsymbol{\delta}}^{y,est}_{-,j} = \arg\min_{\boldsymbol{\delta}^y_{-,j}}\sum_{i\in\mathcal{S}^{est}}\mathbb{1}_c(x)\mathbb{1}_{\ell_j}(z;\Pi)\left(Y_i-\boldsymbol{X}_i'\boldsymbol{\delta}^y_{-,j}\right)^2$$

Estimator for LATE parameter is given by,

$$\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est}) = \frac{\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est}) - \hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est})}{\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est}) - \hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})} = \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})} = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\frac{\hat{\alpha}^y_{+,j} - \hat{\alpha}^y_{-,j}}{\hat{\alpha}^t_{+,j} - \hat{\alpha}^t_{-,j}}$$

and its variance,

$$\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right] &= \frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})\right] \\
&+ \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})\right] \\
&- 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}^{est}),\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})\right]
\end{aligned}$$

where, $\mathbb{C}_{\mathcal{S}^{est}}[\cdot,\cdot]$ is the covariance of two random variable. Each part can be decomposed one step further,

$$\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est})\right] \\
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right] \\
\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}^{est}),\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] \\
&\quad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]
\end{aligned}$$

We have same expected MSE criterion for fuzzy design as well. After the same manipulations as in Section A, we get:

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

We 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 $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right]$,

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

and $\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right]$,

$$\begin{aligned}\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{C}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\boldsymbol{\delta}}_{+,j}^{y,est}, e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\boldsymbol{\delta}}_{+,j}^{t,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\mathbb{C}_{\mathcal{S}^{est}}\left[e_1'\hat{\boldsymbol{\delta}}_{+,j}^{y,est}, e_1'\hat{\boldsymbol{\delta}}_{+,j}^{t,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_{+,j}^{y,est}, \hat{\boldsymbol{\delta}}_{+,j}^{t,est}\right]e_1\end{aligned}$$

All the other variances/covariance have the same form with the appropriate parameter vector. Because $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, we 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 – homoscedastic case – we can write the following sample analogues (below threshold units it is similar).

$$\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}_{+,j}^{y,est}\right] = \frac{\hat{\sigma}_{+,j}^{2,y}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}} \quad , \quad \widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}_{+,j}^{t,est}\right] = \frac{\hat{\sigma}_{+,j}^{2,t}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}} \quad , \quad \widehat{\mathbb{C}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}_{+,j}^{y,est}, \hat{\boldsymbol{\delta}}_{+,j}^{t,est}\right] = \frac{\hat{C}_{+,j}^{y,t}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}$$

where,

$$\begin{aligned}N_{+,j}^{est} &= \sum_{i\in\mathcal{S}^{est}}\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \quad , \quad N_{+,j}^{te} = \sum_{i\in\mathcal{S}^{te}}\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\\ \hat{M}_{+,j} &= \frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\\ \hat{\sigma}_{+,j}^{2,y} &= \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\left[(\epsilon_i^y)^2\,\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right] \ , \qquad \epsilon_i^y = Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\delta}}_{+,j}^{y,te}\\ \hat{\sigma}_{+,j}^{2,t} &= \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\left[(\epsilon_i^t)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right] \ , \qquad \epsilon_i^t = T_i - \boldsymbol{X}_i'\hat{\boldsymbol{\delta}}_{+,j}^{t,te}\\ \hat{C}_{+,j}^{y,t} &= \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\left(\epsilon_i^y\epsilon_i^t\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right) \ , \qquad j=1,2,\ldots,\#\Pi\end{aligned}$$

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 Section D.

Putting together the variances, in the homoscedastic case we have the following expression,

$$
\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 $\tau^2_{FRD}(Z_i;\Pi)$. Similarly to sharp RD, we 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,

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

D 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, we show derivation of $\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{+,j}\right]$ – the other parts can be calculated similarly – then we put together with the other parts.
General case:

$$\begin{aligned}
\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{+,j}\right] &= \frac{1}{N^{est}_{+,j}} \left(\frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \boldsymbol{X}_i \boldsymbol{X}_i' \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)\right)^{-1} \left[\frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \boldsymbol{X}_i' \hat{\Omega} \boldsymbol{X}_i \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)\right] \\
&\qquad \left(\frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \boldsymbol{X}_i \boldsymbol{X}_i' \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)\right)^{-1} \\
&= \frac{1}{N^{est}_{+,j}} \hat{M}^{-1}_{+,j} \left[\frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \boldsymbol{X}_i' \hat{\Omega} \boldsymbol{X}_i \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)\right] \hat{M}^{-1}_{+,j} \\
&= \frac{1}{N^{est}_{+,j}} \hat{M}^{-1}_{+,j} \hat{\Sigma}_{+,j} \hat{M}^{-1}_{+,j}
\end{aligned}$$

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':

$$\hat{\Sigma}^{HCE1}_{+,j} = \frac{1}{N^{te}_{+,j} - p - 1} \sum_{i \in \mathcal{S}^{te}} \boldsymbol{X}_i' \boldsymbol{X}_i (\epsilon^y_i)^2 \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)$$

In case of clustered SE, with HC1

$$\hat{\Sigma}^{C}_{+,j} = \frac{N^{te}_{+,j} - 1}{(N^{te}_{+,j} - p - 1)^2} \frac{N_c}{N_c} \sum_{i \in \mathcal{S}^{te}} \left(\sum_{c=1}^{N_c} \boldsymbol{X}'_{i,c} \boldsymbol{X}_{i,c} (\epsilon^y_{i,c})^2\right) \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i)$$

where, N_c is the number of clusters in leaf j above the threshold in the test sample.

In sharp RD, we get 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 together the variance for LATE,

$$
\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(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{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\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{\boldsymbol{\delta}}^{y,est}_{+,j}, \hat{\boldsymbol{\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{\boldsymbol{\delta}}^{y,est}_{-,j}, \hat{\boldsymbol{\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{\boldsymbol{\delta}}^{y,est}_{+,j}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{-,j}\right]\right)\right. \\
&\qquad + A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{t,est}_{+,j}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{t,est}_{-,j}\right]\right) \\
&\qquad \left. -2A_3\left(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{+,j}, \hat{\boldsymbol{\delta}}^{t,est}_{+,j}\right] + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{-,j}, \hat{\boldsymbol{\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{\boldsymbol{\delta}}^{y,est}_{+,j}\right] + A_2 \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{t,est}_{+,j}\right] - 2A_3 \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{+,j}, \hat{\boldsymbol{\delta}}^{t,est}_{+,j}\right]\right. \\
&\qquad \left. + A_1 \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{-,j}\right] + A_2 + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{t,est}_{-,j}\right] - 2A_3 \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{y,est}_{-,j}, \hat{\boldsymbol{\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 M^{-1} several time, and the variance coming from the outcome, treatment equation and their covariance can be added up, based on only the test sample.

E Monte Carlo simulation setup

For Monte Carlo simulations we use a general formulation for the DGPs and change the appropriate parts for each specific setup.

$$Y_i = \eta(X_i, Z_{i,k}) + \mathbb{1}_{X_i \geq 0} \times \kappa(Z_{i,k}) + \epsilon_i$$

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 $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. We generate $k = 1, \ldots, K$ features such that $Z_{i,k}$ it is independent across k and from $\epsilon_i, Z_{i,k}$ and X_i. The source of variation is only in ϵ_i during the simulations, thus $X_i, Z_{i,k}$ are the same across the $1,000$ Monte Carlo samples. All the other terms are dependent on the setup.
We report three Monte Carlo average statistics to evaluate the performance of our algorithm:

1. Infeasible MSE: $MSE^{infeasible} = \frac{1}{N^{te}} \sum_{i=1}^{N^{te}} (\tau(Z_i; \Pi(\hat{\mathcal{S}}^{tr}, \mathcal{S}^{est}) - \kappa(Z_{i,k}))^2$
2. 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 the same as for the DGP. If the grown tree is smaller (less leaves) or larger (more leaves) than the original DGP, then it is not found.[30]

For DGP 1 and 2, we use linear in X_i DGPs with $X_i \sim U[-1, 1]$ where the threshold value is $c = 0$. For the features, we use four variables, two binary ($Z_{i,1-2}$) with 0.5 marginal probability of being 1 and two uniformly distributed continuous variables ($Z_{i,3-4} \sim U[-5, 5]$). We use different values $(0.1, 1, 10)$ as the standard deviation of the disturbance term (σ_ϵ^2) to investigate the effect of the noise for the performance of the algorithm.

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.

$$\begin{aligned} \eta(X_i) &= 2 \times X_i \\ \kappa(Z_{i,3}) &= Z_{1,i} - (1 - Z_{1,i}) + \end{aligned}$$

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.

$$\begin{aligned} \eta(X_i, Z_{i,2}) &= 2 \times Z_{i,2} \times X_i - 2 \times (1 - Z_{i,2}) \times X_i \\ \kappa(Z_{i,3}) &= 2 \times Z_{i,3} \end{aligned}$$

DGP 3-5 uses nonlinear specification for X_i. We 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 to model to investigate the effect of extra 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 standard deviation of 0.05 or 0.1295.

[30] We allow the splitting value for each feature to be within 0.2 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.

DGP 3: Imitating Lee (2008) vote-shares. We assume two treatment effects and heterogeneous $\eta(\cdot)$. We 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 $Z_{i,1} = 1$ we 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. We 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 we add dummies representing different continents.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 3.71 + 2.30X_i + 3.28X_i^2 + 1.45X_i^3 + 0.23X_i^4 + 0.03X_i^5, & \text{if } X_i < 0 \\ 3.71 + 18.49X_i - 54.81X_i^2 + 74.30X_i^3 - 45.02X_i^4 + 9.83X_i^5, & \text{if } X_i < 0 \end{cases}$$

$$\kappa(Z_{i,1}) = -5.45 - (Z_{1,i} - 5);$$

DGP 5: 'An Alternative DGP' by Calonico et al. (2014) which add 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 covariates/features are the same as DGP 4 and they are all irrelevant: treatment effect and $\eta(\cdot)$ is homogeneous.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 0.48 + 1.27X_i - 0.5 \times 7.18X_i^2 + 0.7 \times 20.21X_i^3 \\ \quad +1.1 \times 21.54X_i^4 + 1.5 \times 7.33X_i^5, & \text{if } X_i < 0 \\ 0.48 + 0.84X_i - 0.1 \times 3.00X_i^2 - 0.3 \times 7.99X_i^3 \\ \quad -0.1 \times 9.01X_i^4 + 3.56X_i^5, & \text{if } X_i < 0 \end{cases}$$

$$\kappa(Z_{i,1}) = 0.04$$

F Additional Monte Carlo simulations

DGP	σ_ϵ	inf. MSE	#Π	DGP found
DGP-1	$\sigma^2 = 0.1$	0.0003	2.0000	1.0000
	$\sigma^2 = 1$	0.0031	2.0010	0.9990
	$\sigma^2 = 10$	0.0411	1.9940	0.9860
DGP-2	$\sigma^2 = 0.1$	0.8150	9.9210	-
	$\sigma^2 = 1$	0.9781	9.3690	-
	$\sigma^2 = 10$	1.7554	7.0120	-
DGP-3	$\sigma = 0.05$	0.0001	1.9790	0.9790
	$\sigma = 0.125$	0.0008	1.1220	0.1220
DGP-4	$\sigma = 0.05$	0.3141	4.0010	-
	$\sigma = 0.125$	0.3131	4.1890	-
DGP-5	$\sigma = 0.05$	0.0000	1.0040	0.9960
	$\sigma = 0.125$	0.0002	1.0050	0.9950

Table 3: Sharp RDD: Monte Carlo averages, using HCE-1, 1SE CV value, p=1 for DGP 1 and 2 and p=5 for DGP 3, 4 and 5

DGP 1	true value	Leaf 1: $Z_1 = 1$, $\tau_1 = 1$			Leaf 2: $Z_1 = 0$, $\tau_2 = -1$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE\left[\hat{\tau}_1\right]}$	$SD\left[\hat{\tau}_1\right]$	$\overline{\hat{\tau}_2}$	$\overline{SE\left[\hat{\tau}_2\right]}$	$SD\left[\hat{\tau}_2\right]$
	$\sigma^2 = 0.1$	1.0001	0.0180	0.0175	-0.9997	0.0177	0.0178
	$\sigma^2 = 1$	1.0004	0.0569	0.0555	-0.9991	0.0559	0.0563
	$\sigma^2 = 10$	0.9910	0.1799	0.2017	-0.9872	0.1764	0.2033
DGP 2	true value	Leaf 1: $Z_3 = -4$, $\tau(-4) = -8$			Leaf 2: $Z_3 = -2$, $\tau(-2) = -4$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE\left[\hat{\tau}_1\right]}$	$SD\left[\hat{\tau}_1\right]$	$\overline{\hat{\tau}_2}$	$\overline{SE\left[\hat{\tau}_2\right]}$	$SD\left[\hat{\tau}_2\right]$
	$\sigma^2 = 0.1$	-7.8777	0.1070	0.1483	-4.1426	0.1094	0.2684
	$\sigma^2 = 1$	-7.8913	0.1570	0.2337	-4.2041	0.1560	0.3175
	$\sigma^2 = 10$	-7.7451	0.3846	0.6610	-4.3175	0.3136	1.3157
	true value	Leaf 3: $Z_3 = 0$, $\tau(0) = 0$			Leaf 4: $Z_3 = 2$, $\tau(2) = 4$		
	estimates	$\overline{\hat{\tau}_3}$	$\overline{SE\left[\hat{\tau}_3\right]}$	$SD\left[\hat{\tau}_3\right]$	$\overline{\hat{\tau}_4}$	$\overline{SE\left[\hat{\tau}_4\right]}$	$SD\left[\hat{\tau}_4\right]$
	$\sigma^2 = 0.1$	-0.0285	0.1233	0.4610	4.7862	0.1060	0.1424
	$\sigma^2 = 1$	-0.0569	0.1600	0.7723	4.5139	0.1466	0.5551
	$\sigma^2 = 10$	0.0596	0.3095	1.3716	3.8246	0.2789	1.1685
	true value	Leaf 5: $Z_3 = 4$, $\tau(2) = 8$					
	estimates	$\overline{\hat{\tau}_5}$		$\overline{SE\left[\hat{\tau}_5\right]}$		$SD\left[\hat{\tau}_5\right]$	
	$\sigma^2 = 0.1$	8.1072		0.1137		0.1642	
	$\sigma^2 = 1$	8.0670		0.1502		0.1626	
	$\sigma^2 = 10$	7.8669		0.3182		0.4771	
DGP 3	true value	Leaf 1: $Z_1 = 0$, $\tau_1 = 0.07$			Leaf 2: $Z_1 = 1$, $\tau_2 = 0.02$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE\left[\hat{\tau}_1\right]}$	$SD\left[\hat{\tau}_1\right]$	$\overline{\hat{\tau}_2}$	$\overline{SE\left[\hat{\tau}_2\right]}$	$SD\left[\hat{\tau}_2\right]$
	$\sigma = 0.05$	0.0697	0.0082	0.0086	0.0202	0.0084	0.0093
	$\sigma = 0.125$	0.0534	0.0162	0.0169	0.0469	0.0169	0.0186
DGP 4	true value	Leaf 1: $Z_{age} = 5$, $\tau(5) = -5.45$			Leaf 2: $Z_{age} = 6$, $\tau(6) = -6.45$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE\left[\hat{\tau}_1\right]}$	$SD\left[\hat{\tau}_1\right]$	$\overline{\hat{\tau}_2}$	$\overline{SE\left[\hat{\tau}_2\right]}$	$SD\left[\hat{\tau}_2\right]$
	$\sigma = 0.05$	-5.8235	0.0411	0.0373	-6.2973	0.0444	0.0461
	$\sigma = 0.125$	-5.8425	0.0554	0.0788	-6.2708	0.0582	0.0978
	true value	Leaf 3: $Z_{age} = 7$, $\tau(7) = -7.45$			Leaf 4: $Z_{age} = 8$, $\tau(8) = -8.45$		
	estimates	$\overline{\hat{\tau}_3}$	$\overline{SE\left[\hat{\tau}_3\right]}$	$SD\left[\hat{\tau}_3\right]$	$\overline{\hat{\tau}_4}$	$\overline{SE\left[\hat{\tau}_4\right]}$	$SD\left[\hat{\tau}_4\right]$
	$\sigma^2 = 0.05$	-7.6403	0.0809	0.0409	-7.6403	0.0809	0.0409
	$\sigma^2 = 0.1295$	-7.6739	0.0833	0.1085	-7.6739	0.0833	0.1085
	true value	Leaf 5: $Z_{age} = 9$, $\tau(9) = -9.45$					
	estimates	$\overline{\hat{\tau}_5}$		$\overline{SE\left[\hat{\tau}_5\right]}$		$SD\left[\hat{\tau}_5\right]$	
	$\sigma^2 = 0.05$	-9.1891		0.0388		0.0168	
	$\sigma^2 = 0.1295$	-9.1410		0.0566		0.02320	
DGP 5	true value	Homogeneous Treatment, $\tau = 0.04$					
	estimates	$\overline{\hat{\tau}}$		$\overline{SE\left[\hat{\tau}\right]}$		$SD\left[\hat{\tau}\right]$	
	$\sigma = 0.05$	0.0398		0.0058		0.0056	
	$\sigma = 0.125$	0.0396		0.0151		0.0145	

Table 4: Sharp RDD: Estimated Monte Carlo averages and standard deviation of parameters for treatment and standard errors by each leaf, with 1-SE CV value

For fuzzy designs, we use the same functional forms for the DGPs, but add a homogeneous first-stage for getting the treatment,

$$T_i = \begin{cases} \mathbb{1}\left(0.5 + 0.8X_i + \epsilon_t > 0\right) \ , & \text{if } X_i \geq 0 \\ 0 & \text{if } X_i < 0 \end{cases}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$. For clarity we uses 'DGP-x-f' for these fuzzy setups.

DGP	σ_ϵ	inf. MSE	#Π	DGP found
DGP-1-f	$\sigma^2 = 0.1$	0.0007	2.2010	0.8220
	$\sigma^2 = 1$	0.0063	2.0610	0.9450
	$\sigma^2 = 10$	0.1860	1.9230	0.8320
DGP-2-f	$\sigma^2 = 0.1$	1.4523	7.4300	-
	$\sigma^2 = 1$	1.5132	7.9020	-
	$\sigma^2 = 10$	2.5411	5.5020	-
DGP-3-f	$\sigma = 0.05$	0.0005	1.3210	0.3210
	$\sigma = 0.125$	0.0008	1.0970	0.0970
DGP-4-f	$\sigma = 0.05$	0.1049	6.4470	-
	$\sigma = 0.125$	0.1481	6.1190	-
DGP-5-f	$\sigma = 0.05$	0.0000	1.0270	0.9730
	$\sigma = 0.125$	0.0003	1.0240	0.9760

Table 5: Fuzzy RDD: Monte Carlo averages, using HCE-1, smallest CV value, p=1 for DGP 1 and 2 and p=5 for DGP 3, 4 and 5

DGP-1-f	true value	Leaf 1: $Z_1 = 1$, $\tau_1 = 1$			Leaf 2: $Z_1 = 0$, $\tau_2 = -1$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE[\hat{\tau}_1]}$	$SD[\hat{\tau}_1]$	$\overline{\hat{\tau}_2}$	$\overline{SE[\hat{\tau}_2]}$	$SD[\hat{\tau}_2]$
	$\sigma^2 = 0.1$	1.0010	0.0247	0.0249	-0.9990	0.0257	0.0248
	$\sigma^2 = 1$	1.0031	0.0778	0.0786	-0.9968	0.0800	0.0785
	$\sigma^2 = 10$	0.8863	0.2453	0.3996	-0.8616	0.2453	0.4208
DGP-2-f	true value	Leaf 1: $Z_3 = -4$, $\tau(-4) = -8$			Leaf 2: $Z_3 = -2$, $\tau(-2) = -4$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE[\hat{\tau}_1]}$	$SD[\hat{\tau}_1]$	$\overline{\hat{\tau}_2}$	$\overline{SE[\hat{\tau}_2]}$	$SD[\hat{\tau}_2]$
	$\sigma^2 = 0.1$	-5.4033	0.1354	0.1578	-5.3992	0.1354	0.1552
	$\sigma^2 = 1$	-5.4150	0.1706	0.2910	-5.3977	0.1706	0.2789
	$\sigma^2 = 10$	-5.5869	0.3495	0.9250	-5.2159	0.3488	0.9787
	true value	Leaf 3: $Z_3 = 0$, $\tau(0) = 0$			Leaf 4: $Z_3 = 2$, $\tau(2) = 4$		
	estimates	$\overline{\hat{\tau}_3}$	$\overline{SE[\hat{\tau}_3]}$	$SD[\hat{\tau}_3]$	$\overline{\hat{\tau}_4}$	$\overline{SE[\hat{\tau}_4]}$	$SD[\hat{\tau}_4]$
	$\sigma^2 = 0.1$	0.7288	0.1416	0.3947	3.6186	0.1358	0.3264
	$\sigma^2 = 1$	0.6226	0.2062	0.5639	3.5404	0.1934	0.3799
	$\sigma^2 = 10$	0.8128	0.4005	1.3599	3.0606	0.3758	1.1807
	true value	Leaf 5: $Z_3 = 4$, $\tau(2) = 8$					
	estimates	$\overline{\hat{\tau}_5}$		$\overline{SE[\hat{\tau}_5]}$		$SD[\hat{\tau}_5]$	
	$\sigma^2 = 0.1$	7.6433		0.1301		0.1014	
	$\sigma^2 = 1$	7.6255		0.1895		0.1455	
	$\sigma^2 = 10$	7.5325		0.4034		0.5159	
DGP-3-f	true value	Leaf 1: $Z_1 = 0$, $\tau_1 = 0.07$			Leaf 2: $Z_1 = 1$, $\tau_2 = 0.02$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE[\hat{\tau}_1]}$	$SD[\hat{\tau}_1]$	$\overline{\hat{\tau}_2}$	$\overline{SE[\hat{\tau}_2]}$	$SD[\hat{\tau}_2]$
	$\sigma = 0.05$	0.0513	0.0085	0.0150	0.0348	0.0087	0.0128
	$\sigma = 0.125$	0.0447	0.0170	0.0187	0.0393	0.0175	0.0181
DGP-4-f	true value	Leaf 1: $Z_{age} = 5$, $\tau(5) = -5.45$			Leaf 2: $Z_{age} = 6$, $\tau(6) = -6.45$		
	estimates	$\overline{\hat{\tau}_1}$	$\overline{SE[\hat{\tau}_1]}$	$SD[\hat{\tau}_1]$	$\overline{\hat{\tau}_2}$	$\overline{SE[\hat{\tau}_2]}$	$SD[\hat{\tau}_2]$
	$\sigma = 0.05$	-6.6026	0.1791	0.0987	-6.6026	0.1791	0.0987
	$\sigma = 0.125$	-6.5917	0.1836	0.1919	-6.5917	0.1836	0.1919
	true value	Leaf 3: $Z_{age} = 7$, $\tau(7) = -7.45$			Leaf 4: $Z_{age} = 8$, $\tau(8) = -8.45$		
	estimates	$\overline{\hat{\tau}_3}$	$\overline{SE[\hat{\tau}_3]}$	$SD[\hat{\tau}_3]$	$\overline{\hat{\tau}_4}$	$\overline{SE[\hat{\tau}_4]}$	$SD[\hat{\tau}_4]$
	$\sigma = 0.05$	-7.2928	0.0670	0.0377	-8.9366	0.1560	0.0341
	$\sigma = 0.1295$	-7.2459	0.0739	0.1793	-8.9088	0.1623	0.0829
	true value	Leaf 5: $Z_{age} = 9$, $\tau(9) = -9.45$					
	estimates	$\overline{\hat{\tau}_5}$		$\overline{SE[\hat{\tau}_5]}$		$SD[\hat{\tau}_5]$	
	$\sigma = 0.05$	-8.9366		0.1542		0.0341	
	$\sigma = 0.1295$	-8.9093		0.1504		0.0831	
DGP-5-f	true value	Homogeneous Treatment, $\tau = 0.04$					
	estimates	$\overline{\hat{\tau}}$		$\overline{SE[\hat{\tau}]}$		$SD[\hat{\tau}]$	
	$\sigma = 0.05$	0.0396		0.0062		0.0061	
	$\sigma = 0.125$	0.0391		0.0160		0.0157	

Table 6: Fuzzy RDD: Estimated Monte Carlo averages and standard deviation of parameters for treatment and standard errors by each leaf, with smallest CV value