Online Appendix:
Discovering Heterogeneous Treatment Effects in Regression Discontinuity Designs

Ágoston Reguly[1,2]

[1]Corvinus University of Budapest
[2]Georgia Institute of Technology

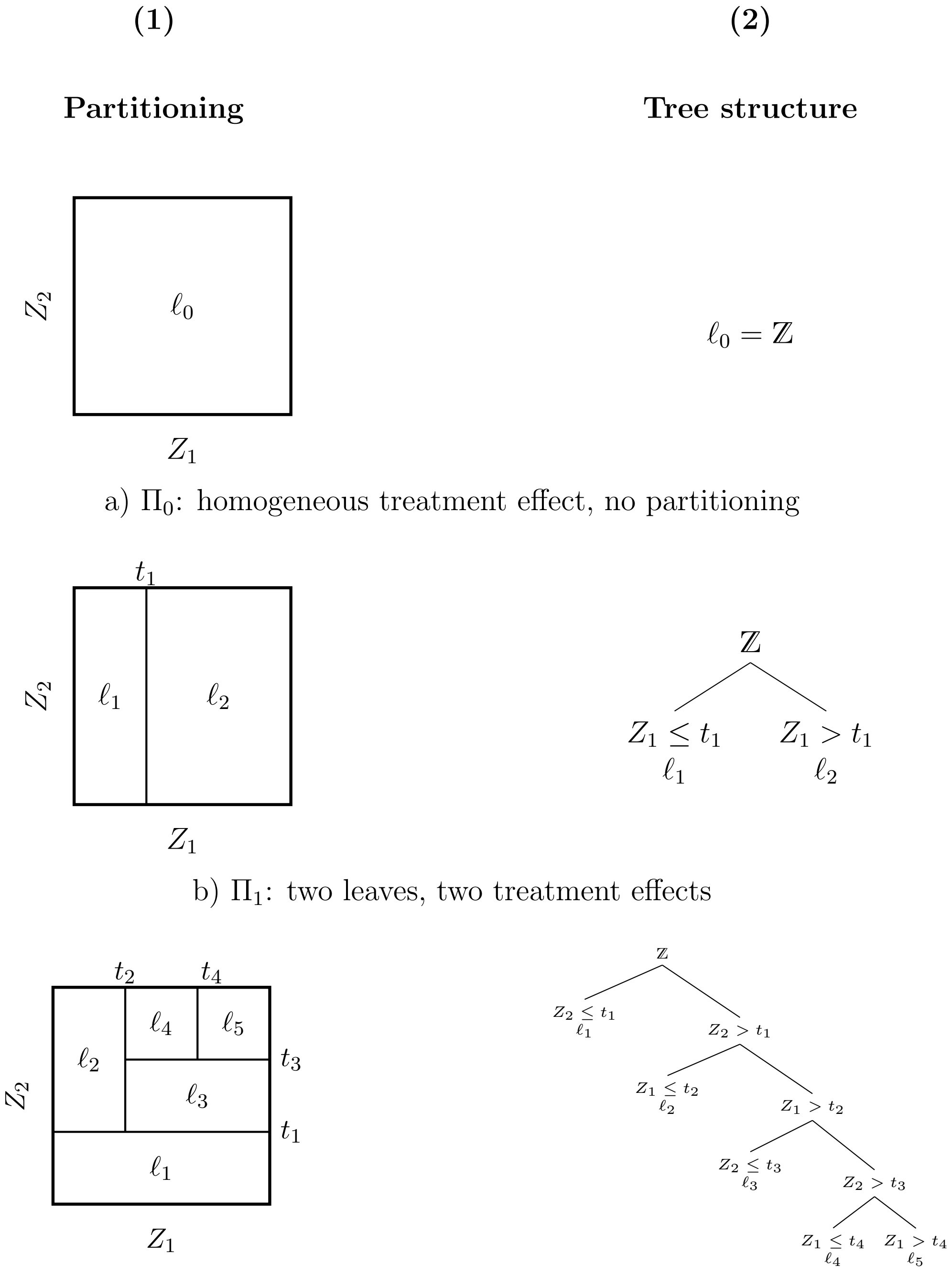
August 29, 2025

A Tree alogorithm details

A.1 Partitioning

For illustrative purposes, consider two features Z_1 and Z_2. Figure A.1 shows three different trees with two partitioning representation and in tree structure. 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 A.1a) shows a tree, where there is only one leaf ℓ_0 containing all the units. This tree corresponds to a homogeneous treatment effect: no matter which values 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 A.1b) has two leaves: ℓ_1 and ℓ_2 resulting in two different treatment effects. Leaf ℓ_1 contains values with $Z_1 \leq t_1$ and ℓ_2 contains $Z_1 > t_1$, where t_1 is the splitting value. Note that Z_2 does not affect the partitioning and irrelevant with respect to treatment heterogeneity. Finally Figure A.1c) shows a tree with five different leaves, resulting in five

Email address: agoston.reguly@uni-corvinus.hu
Codes are available at `https://github.com/regulyagoston/RD_tree`



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 A.1: Different trees and their conditional average treatment effects

different treatment effects depending on both Z_1 and Z_2. In this case if one wants to find the treatment effect for a unit with $Z_1 = z_1$ and $Z_2 = z_2$, one needs to go through the decisions given by the tree. *Example*: $z_1 > t_3$ and $t_2 < z_2 \leq t_4$, corresponds to leaf ℓ_4. Note that the splitting values must satisfy $t_3 > t_1$, $t_1, t_3 \in Supp(Z_1)$ and $t_2, t_4 \in Supp(Z_2)$.

A.2 Finding EMSE optimal RD tree

In this sub-section, discuss the steps to grow the EMSE optimal regression discontinuity tree. We mainly follow the literature on classification and regression trees (CART) and honest causal regression trees (see, e.g., Breiman et al. (1984), Ripley (1996) or Hastie et al. (2011) on CART algorithms and Athey & Imbens (2015, 2016) on honest causal tree algorithm). Let us reiterate the steps of growing discontinuity tree,

Steps for growing discontinuity tree:

1. Split the sample into two independent parts → training and estimation samples.
2. Grow a large tree on the training sample with bandwidth h_r.
3. Prune this large tree to control for over-fitting. This is carried out by cross-validation, where we employ weakest link pruning.
4. Minimize bandwidth h_r such that the pruned tree has the smallest cross-validated EMSE value.
5. Use this EMSE optimal tree with h_r^* to estimate the CATE function on the independent estimation sample.

In the first stage '*honest*' approach randomly assigns the initial sample into two samples to achieve an unbiased CATE estimator. The first sample is called the *'training sample'* ($\mathcal{S}^{tr}$) and its observations are used to grow trees. The second, *'estimation sample'* has a special role. In general, it is locked away from the algorithm, but information on the number of observations is utilized during the tree building phase to control for finding training sample specific patterns. Observation values from the estimation sample are not used until the last stage. This division ensures valid inference for the CATE function in the fifth step. Figure A.2 shows these two samples, which are used to grow a large tree and providing valid inference.

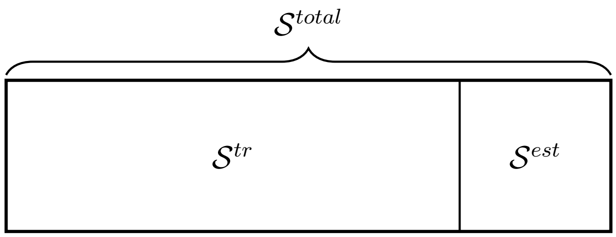


Figure A.2: At stage 2, $\mathcal{S}^{tr}$ is used to grow large tree $\left(\hat{\Pi}^{large}\right)$

In the second stage, a large tree is grown using all the observations from the *training sample* with a given bandwidth h_r. The algorithm recursively partitions the training sample along with the features. For each leaf the method evaluates all candidate features and their possible splits inducing alternative partitions, with the *'honest in-sample criterion'*: $\widehat{EMSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi, h_r)$. This criterion uses additional information from the estimation sample. The treatment effects and the variances are estimated on the training sample only, but they are adjusted with the number of observations (N^{est}) and share of treated and non-treated units within each leaf from the estimation sample ($p^{est}_{\pm,j}$). N^{est} adjusts for the sample shares (how the initial sample is divided into two parts). This does not have a large impact on the in-sample criterion as the value is given by the first step and does not change during the partitioning. Using $p^{est}_{\pm,j}$ instead of $p^{tr}_{\pm,j}$ has larger implication in finite samples. It prevents the algorithm to choose such a feature and splitting value, which is only specific to the training sample. After the split is done, the algorithm iterates the procedure on the newly created leaves. The process repeats itself and stops if the in-sample-criterion does not decrease any further or the magnitude of the reduction is smaller than a pre-set parameter.[1] With this method, one gets a large tree $\left(\hat{\Pi}^{large}\right)$.

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

To solve for the over-fitting – in the third stage – cross-validation is used. The idea is to

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

split the training sample into two further parts: a sample where the tree is independently grown $\mathcal{S}^{(tr,tr)}$ and to a test sample $\mathcal{S}^{(tr,te)}$ where the EMSE criterion can be safely evaluated. This ensures that the tree grown on $\mathcal{S}^{(tr,tr)}$ is exogenous for $\mathcal{S}^{(tr,te)}$, thus the estimated EMSE criterion is unbiased.[2] Figure A.3, shows the splitting of the original training sample into $\mathcal{S}^{(tr,tr)}$ and $\mathcal{S}^{(tr,te)}$.

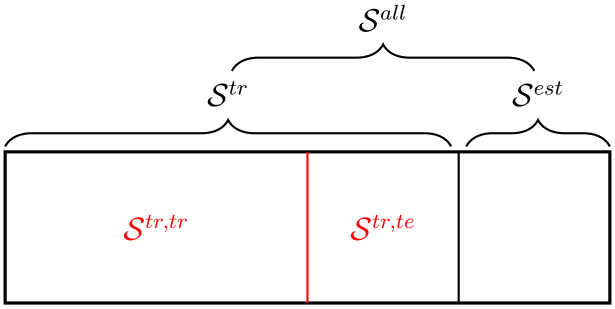


Figure A.3: At stage 3, $\mathcal{S}^{tr,te}$ is used to evaluate the tree $\hat{\Pi}$ grown on $\mathcal{S}^{tr,tr}$

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

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

$$\widehat{EMSE}_{cv}(\gamma, h_r) = \widehat{EMSE}_{\tau}(\mathcal{S}^{(tr,val)}, \mathcal{S}^{(est,val)}, \widehat{\Pi}, h_r) + \gamma\#\widehat{\Pi} \tag{A.1}$$

where, $\widehat{\Pi}$ is an estimator of the tree, grown on the samples of $\{\mathcal{S}^{(tr,tr)}, \mathcal{S}^{(est,tr)}\}$ and the EMSE criterion is evaluated on the independent sample pair of $\{\mathcal{S}^{(tr,te)}, \mathcal{S}^{(est,te)}\}$. To find the optimal complexity parameter (γ_r^*) – hence the EMSE optimal tree – with the bandwidth h_r one calculates the honest cross-validation criterion R times on the alternating test samples, which results in R criteria for each different candidate of γ.[3] Taking the average over the cross-validation samples one can choose γ, which gives the minimum criterion value with

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

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

h_r.[4]

$$\gamma_r^* = \arg\min_{\gamma} R^{-1} \sum_{cv=1}^{R} \widehat{EMSE}_{cv}(\gamma, h_r) \tag{A.2}$$

The next step is to iterate through different h_r and find the optimal bandwidth h^* that solves,

$$h^* = \arg\min_{h} R_h^{-1} \sum_{r=1}^{R_h} \widehat{EMSE}_{cv}(\gamma_r^*, h)$$

This two-step process is possible as for each candidate h_r, γ_r^* is unique.

The next step is to grow a large tree with h^* on the whole training sample and prune back with accompanied γ^* to get the optimal tree $\widehat{\Pi}^*$. The last step uses the locked away estimation sample and the found tree structure $\widehat{\Pi}^*$ to estimate the CATE function for the regression discontinuity tree.

B Derivation of sharp RDD criterion

B.1 (In)feasible EMSE criterion

First, let us manipulate the the infeasible EMSE criterion.

$$\begin{aligned}
EMSE_\tau(\Pi, h) &= \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}}\left[MSE_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}; \Pi, h)\right] \\
&= \mathbb{E}_{X_i, Z_i, \mathcal{S}^{est}}\left\{\left[\tau(Z_i) - \hat{\tau}(z; \Pi, h, \mathcal{S}^{est})\right]^2 - \tau^2(Z_i)\right\} \\
&= \mathbb{E}_{X_i, Z_i, \mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i; \Pi, h, \mathcal{S}^{est}) - 2\hat{\tau}(Z_i; \Pi, h, \mathcal{S}^{est})\tau(Z_i)\right\} \\
&\text{using law of iterated expectations} \\
&= \mathbb{E}_{X_i, Z_i, \mathcal{S}^{est}}\left\{\mathbb{E}\left[\hat{\tau}^2(Z_i; \Pi, h, \mathcal{S}^{est}) - 2\hat{\tau}(Z_i; \Pi, h, \mathcal{S}^{est})\tau(Z_i) \;\middle|\; X_i = c, \mathbb{1}_{\ell_1}(Z_i), \dots, \mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]\right\} \\
&= \mathbb{E}_{X_i, Z_i, \mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i; \Pi, h, \mathcal{S}^{est}) - 2\hat{\tau}(Z_i; \Pi, h, \mathcal{S}^{est})\mathbb{E}\left[\tau(Z_i) \;\middle|\; X_i = c, \mathbb{1}_{\ell_1}(Z_i), \dots, \mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]\right\} \\
&\text{as } \mathbb{E}\left[\tau(Z_i) \;\middle|\; X_i = c, \mathbb{1}_{\ell_1}(Z_i), \dots, \mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right] = \tau(Z_i; \Pi) \\
&= \mathbb{E}_{Z_i, \mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i; \Pi, h, \mathcal{S}^{est}) - 2\hat{\tau}(Z_i; \Pi, h, \mathcal{S}^{est})\tau(Z_i; \Pi)\right\} \\
&= \mathbb{E}_{Z_i, \mathcal{S}^{est}}\left\{\left[\tau(Z_i; \Pi) - \hat{\tau}(Z_i; \Pi, h, \mathcal{S}^{est})\right]^2 - \tau^2(Z_i; \Pi)\right\} \\
&\text{using } Z_i \perp\!\!\!\perp \mathcal{S}^{est} \\
&= \mathbb{E}_{Z_i, \mathcal{S}^{est}}\left\{\left[\tau(Z_i; \Pi) - \hat{\tau}(Z_i; \Pi, h, \mathcal{S}^{est})\right]^2\right\} - \mathbb{E}_{Z_i}\left\{\tau^2(Z_i; \Pi)\right\}
\end{aligned}$$

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

using law of iterated expectations and $Z_i \perp\!\!\!\perp \mathcal{S}^{est}$ yields,

$$EMSE_\tau(\Pi, h) = \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(Z_i;\Pi) - \hat{\tau}(z;\Pi, h, \mathcal{S}^{est})\right)^2 \mid Z_i\right]\right\} - \mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\}. \quad \text{(A.3)}$$

Let us focus on the first part and use the fact that Z_i is independent of $\mathcal{S}^{est}$, thus,

$$\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(Z_i;\Pi) - \hat{\tau}(z;\Pi, h, \mathcal{S}^{est})\right)^2 \mid Z_i\right] \overset{Z_i \perp\!\!\!\perp \mathcal{S}^{est}}{=} \mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(z;\Pi) - \hat{\tau}(z;\Pi, h, \mathcal{S}^{est})\right)^2\right] \Big|_{z=Z_i} .$$

for a fixed z and take expectations w.r.t. the distribution of Z_i in the end. We have

$$\begin{aligned}\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(z;\Pi) - \hat{\tau}(z;\Pi, h, \mathcal{S}^{est})\right)^2\right] &= MSE(\hat{\tau}(z;\Pi, h, \mathcal{S}^{est})) \\ &= \mathbb{B}^2(\hat{\tau}(z;\Pi, h, \mathcal{S}^{est})) + \mathbb{V}(\hat{\tau}(z;\Pi, h, \mathcal{S}^{est})).\end{aligned} \quad \text{(A.4)}$$

Note that the notation $MSE(\cdot)$, $\mathbb{B}^2(\cdot)$, $\mathbb{V}(\cdot)$ means that the bias, and variance are to be computed with respect to the sampling distribution of the estimation sample, but for a fixed z these quantities are constant and do not depend on the particular estimation sample at hand.

Taking the expectations over Z_i yields,

$$\begin{aligned}\mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(z;\Pi) - \hat{\tau}(z;\Pi, h, \mathcal{S}^{est})\right)^2\right] \Big|_{z=Z_i}\right\} &= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\mathbb{B}^2(\hat{\tau}(z;\Pi, h, \mathcal{S}^{est}))\right] \Big|_{z=Z_i}\right\} \\ &+ \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\mathbb{V}(\hat{\tau}(z;\Pi, h, \mathcal{S}^{est}))\right] \Big|_{z=Z_i}\right\}\end{aligned} \quad \text{(A.5)}$$

that finishes the derivation of the first component in Equation (A.3). For the second component, $\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\}$ lets assume here that $\tilde{\tau}(z;\Pi, h, \mathcal{S}^{est})$ is an unbiased and consistent estimator for $\tau(z;\Pi)$, for any fixed z. (In Section B.2. we discuss this more in detail.) Now, $\tau^2(z;\Pi) = \mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}(z;\Pi, h, \mathcal{S}^{est})])^2$, thus

$$\begin{aligned}\tau^2(z;\Pi) &= \mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}(z;\Pi, h, \mathcal{S}^{est})])^2 \\ &= \mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}^2(z;\Pi, h, \mathcal{S}^{est})] - \mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}(z;\Pi, h, \mathcal{S}^{est})]\end{aligned}$$

This leads to an expression to the expected CATE square,

$$\mathbb{E}_{Z_i}[\tau^2(Z;\Pi)] = \mathbb{E}_{Z_i}\{\mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}^2(Z;\Pi, \mathcal{S}^{est})]\} - \mathbb{E}_{Z_i}\{\mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}(Z;\Pi, \mathcal{S}^{est})]\}. \quad \text{(A.6)}$$

Putting these parts together completes our derivation of the EMSE criterion,

$$\begin{aligned} EMSE_\tau(\Pi, h) = \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\mathbb{B}^2(\hat{\tau}(z;\Pi,h,\mathcal{S}^{est}))\right]\big|_{z=Z_i}\right\} + \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\mathbb{V}(\hat{\tau}(z;\Pi,h,\mathcal{S}^{est}))\right]\big|_{z=Z_i}\right\} \\ \mathbb{E}_{Z_i}\{\mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}^2(Z_i;\Pi,h,\mathcal{S}^{est})]\} + \mathbb{E}_{Z_i}\{\mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}(Z_i;\Pi,h,\mathcal{S}^{est})]\}. \quad \square \end{aligned} \tag{A.7}$$

B.2 Nonparametric estimator for EMSE criterion

Next we propose nonparametric estimator for each components of the derived EMSE criterion given by Equation (A.7). First, let us take the tree structure Π as given, $\hat{\tau}$ has the following form:

$$\hat{\tau}(z;\Pi,h,\mathcal{S}^{est}) = \sum_{\ell=1}^{\#\Pi} \hat{\tau}_j \mathbb{1}_{\ell_j}(z;\Pi),$$

where $\hat{\tau}_j$, $j = 1, \dots, \#\Pi$ represent the nonparametric within-leaf RDD estimators. Utilizing this fact, let us express the bias and variance terms for each leaves:

$$\mathbb{B}[\hat{\tau}_j] = \mathcal{B}_j h^{2q+2} + o(h^{2q+2}) \tag{A.8}$$

and

$$\mathbb{V}[\hat{\tau}_j(\Pi, h)] = \mathcal{V}_j (N_j^{est} h)^{-1} + o((N_j^{est} h)^{-1}) \tag{A.9}$$

where N_j^{est} is the number of observations in the estimation sample falling into leaf j and $\mathcal{B}_j$ and $\mathcal{V}_j$ are leaf-specific constants, that we express in Section B.3. We'll ignore the remainder terms from this point on, leading to an expression for Equation (A.4),

$$MSE(\hat{\tau}_j) = \mathcal{B}_j h^{2q+2} + \mathcal{V}_j (N_j^{est} h)^{-1} \,.$$

Next, the expected variance of $\hat{\tau}(\cdot)$, w.r.t. the distribution of Z_i is given by

$$\mathbb{E}_{Z_i}\big\{\mathbb{V}_{\mathcal{S}^{est}}[\hat{\tau}(z;\Pi,h,\mathcal{S}^{est})]\big\} = \sum_{j=1}^{\#\Pi} \frac{\mathcal{V}_j}{N_j^{est} h} P(Z_i \in \ell_j) \approx \frac{1}{N^{est} h} \sum_{j=1}^{\#\Pi} \mathcal{V}_j$$

if we estimate $P(Z_i \in \ell_j)$ by $p_j^{est} = N_j^{est}/N^{est}$. Similarly, the expected squared bias is given by,

$$\mathbb{E}_{Z_i}\big\{\mathbb{B}^2_{\mathcal{S}^{est}}[\hat{\tau}(z;\Pi,h,\mathcal{S}^{est})]\big\} = \sum_{j=1}^{\#\Pi} h^{2q+2} \mathcal{B}_j^2 P(Z_i \in \ell_j) \approx h^{2q+2} \sum_{j=1}^{\#\Pi} \mathcal{B}_j^2 p_j^{est}.$$

We can use these as estimators for the components of Equation (A.5). Note that these quantities are to be estimated over the *test sample* using nonparametric RDD estimation theory as $\mathcal{S}^{est}$ and $\mathcal{S}^{te}$ are independent from each other. However, one need to adjust with the number of observations in the estimation sample.

The term $\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\}$ is not straightforward to handle because $\mathbb{E}_{\mathcal{S}^{est}}[\hat{\tau}(z;\Pi,h,\mathcal{S}^{est})] \neq \tau(z;\Pi)$ even for a fixed z (the estimator is biased within each leaf). Therefore, to estimate $\mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}(z;\Pi,h,\mathcal{S}^{est})]$, that is unbiased. We can use the bias corrected estimator proposed by Calonico et al. (2014)

$$\tilde{\tau}(z;\Pi,h,\mathcal{S}^{est}) := \sum_{\ell=1}^{\#\Pi}\left(\hat{\tau}_j - \hat{\mathcal{B}}_j h^2\right)\mathbb{1}_{\ell_j}(z;\Pi)\,. \tag{A.10}$$

We can also use the bias corrected variance

$$\mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}(z;\Pi,h,\mathcal{S}^{est})] = \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\mathbb{V}\left[\tilde{\tau}_j(\Pi,h)\right]\,.$$

Note that both estimators $\mathbb{V}\left[\tilde{\tau}_j(\Pi,h)\right]$ require to estimate the bias. Now, lets turn to the expected values over the Z_i for these quantities, where we use simple averages,

$$\mathbb{E}_{Z_i}\{\mathbb{E}_{\mathcal{S}^{est}}[\tilde{\tau}^2(Z;\Pi,\mathcal{S}^{est})]\} = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\tilde{\tau}^2(Z_i;\Pi,\mathcal{S},h)$$

and

$$\mathbb{E}_{Z_i}\{\mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}(Z;\Pi,\mathcal{S}^{est})]\} = \sum_{j=1}^{\#\Pi}p_j^{te}\mathbb{V}\left[\tilde{\tau}_j(\Pi,h)\right]$$

Putting these together, we get a feasible EMSE criterion for nonparametric case.

$$\begin{aligned}\widehat{EMSE}_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi,h) =& h^{2q+2}\sum_{j=1}^{\#\Pi}\hat{\mathcal{B}}_j^2 p_j^{est} + \frac{1}{N^{est}h}\sum_{j=1}^{\#\Pi}\hat{\mathcal{V}}_j \\ &- \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\tilde{\tau}^2(Z_i;\Pi,\mathcal{S},h) + \sum_{j=1}^{\#\Pi}p_j^{te}\mathbb{V}\left[\tilde{\tau}_j(\Pi,h)\right] \quad \square\end{aligned}$$

As a last note, let us mention that asymptotically it is true, $\mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}(z;\Pi,h,\mathcal{S}^{est})] \overset{a}{=} \mathbb{V}_{\mathcal{S}^{est}}[\hat{\tau}(z;\Pi,h,\mathcal{S}^{est})]$, hence the bias corrected and the non-corrected versions are asymptotically the same. One may utilize this fact and further simplify the EMSE expression, by assuming $\mathbb{V}\left[\tilde{\tau}_j(\Pi,h)\right] \overset{a}{=}$

$\mathbb{V}\left[\hat{\tau}_j(\Pi, h)\right] = (n_j h)^{-1}\mathcal{V}_j$. This leads to simplification referring to the variances,

$$\begin{aligned}\widehat{EMSE}_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi, h) =& h^{2q+2}\sum_{j=1}^{\#\Pi}\hat{\mathcal{B}}_j^2 p_j^{est} + \frac{1}{h}\left(\frac{1}{N^{est}} + \frac{1}{N^{te}}\right)\sum_{j=1}^{\#\Pi}\hat{\mathcal{V}}_j \\ & - \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\tilde{\tau}^2(Z_i; \Pi, \mathcal{S}, h)\,.\end{aligned}$$

Note, however, that here we must use $P(Z_i \in \ell_j) = p_j^{te} = N_j^{te}/N^{te}$ for scaling $\mathbb{V}\left[\tilde{\tau}_j(\Pi, h)\right]$.

B.3 Additional quantities

We follow Calonico et al. (2014) to express the bias and variance components. First, let us express the leaf-by-leaf bias component $\mathcal{B}_j$.

$$\begin{aligned}\mathcal{B}_j &= \mathcal{B}_{+,j} - \mathcal{B}_{-,j} \\ \mathcal{B}_{\pm,j} &= e'\Gamma_{j,\pm}^{-1}(h)\nu_{j,\pm}(h)\end{aligned}$$

where e is a selector vector $[1, 0, \ldots, 0]'$ with $q+1$ elements in case of sharp RD, and $\Gamma_{j,\pm}(h)$ and $\nu(h)$ are defined as,

$$\begin{aligned}\Gamma_{j,\pm}(h) &= n_j^{-1}\dot{\boldsymbol{X}}'\boldsymbol{W}_{j,\pm}(h)\dot{\boldsymbol{X}} \\ \nu_\pm(h) &= n_j^{-1}\dot{\boldsymbol{X}}'\boldsymbol{W}_{j,\pm}(h)\dot{\boldsymbol{S}}\,.\end{aligned}$$

Note that $\dot{\boldsymbol{X}}, \dot{\boldsymbol{S}}$ are modified versions of $\boldsymbol{X}$, whereas $\boldsymbol{W}_{j,+}(h), \boldsymbol{W}_{j,-}(h)$ refer to weights below or above the threshold c using the kernel function $k(\cdot)$. Let $r_q(x) = (1, x, \ldots, x^q)$. Then

$$\begin{aligned}\dot{\boldsymbol{X}} &= [r_q(X_1/h), \ldots, r_q(X_n/h)]' \\ \dot{\boldsymbol{S}} &= [(X_1/h)^q, \ldots, (X_n/h)^q]' \\ \boldsymbol{W}_{j,+}(h) &= diag\left(\mathbb{1}(X_1 \geq c, \boldsymbol{Z}_1 \in \ell_j)k_h(X_1), \ldots, \mathbb{1}(X_n \geq c, \boldsymbol{Z}_n \in \ell_j)k_h(X_n)\right) \\ \boldsymbol{W}_{j,-}(h) &= diag\left(\mathbb{1}(X_1 < c, \boldsymbol{Z}_1 \in \ell_j)k_h(X_1), \ldots, \mathbb{1}(X_n < c, \boldsymbol{Z}_n \in \ell_j)k_h(X_n)\right)\,.\end{aligned}$$

These quantities defines the scaler used in Equation (A.8) and their sample counterparts for the bias correction term in Equation (A.10).

Next, let us express the leaf-by-leaf variance, but for simplicity instead of $\tau_j(\Pi, h)$, let us drop Π here.

$$\mathbb{V}\left[\hat{\tau}_j(h_n)\right] = \mathbb{V}_{j,+}(h_n) + \mathbb{V}_{j,-}(h_n)$$

where

$$\begin{aligned}\mathbb{V}_{j,\pm}(h) &= (n_j h)^{-1}\mathcal{V}_{j,\pm}(h)\\ \hat{\mathcal{V}}_{j,\pm}(h) &= e\Gamma^{-1}_{j,\pm}(h)\Psi_{j,\pm}(h,h)\Gamma^{-1}_{j,\pm}(h)e\,.\end{aligned}$$

where $\Psi_{j,\pm}(h,b) = (n_j)^{-1}\,\dot{\boldsymbol{X}}'\boldsymbol{W}_{j,\pm}(h)\Sigma_j\boldsymbol{W}_{j,\pm}(b)\ddot{\boldsymbol{X}}$ with $\ddot{\boldsymbol{X}}$ is similar to $\dot{\boldsymbol{X}}$, but using $q+1$ order polynomials if $b \neq h$, otherwise it is the same. Here we use the same, whereas for the bias-corrected expression we utilize the second bandwidth b as well. $\Sigma_{j,\pm}$ is the variance-covariance estimator in leaf j for each X_i, below or above the threshold. Note that $\mathcal{V}_j = \mathcal{V}_{j,+}(h) + \mathcal{V}_{j,-}(h)$, thus these expressions defines the leaf-by-leaf constant used in Equation (A.9).

As a final estimator, let us express the bias-corrected variance estimator for $\tilde{\tau}_j$.

$$\mathbb{V}\left[\tilde{\tau}_j(h_n,b_n)\right] = \mathbb{V}_{j,+}(h_n,b_n) + \mathbb{V}_{j,-}(h_n,b_n)\,,$$

where,

$$\begin{aligned}\mathbb{V}_{j,\pm}(h,b) =& (n_j h)^{-1}\dot{\mathcal{V}}_{j,\pm}(h) - 2h^{q+1}(n_j b^{q+1})^{-1}\mathcal{C}_{j,\pm}(h,b)\frac{\mathcal{B}_{j,\pm}(h)}{(q+1)!}\\ &+ h^{2(q+1)}(n_j b)^{-1}\ddot{\mathcal{V}}_{j,\pm}(b)\frac{\mathcal{B}^2_{j,\pm}(h)}{(q+1)!^2}\\ \mathcal{C}_{j,\pm}(h,b) =& (q+1)e'\dot{\Gamma}^{-1}_{j,\pm}(h)\Psi_{j,\pm}(h,b)\ddot{\Gamma}^{-1}_{j,\pm}(b)e\,,\end{aligned}$$

and $\dot{\mathcal{V}}$ and $\ddot{\mathcal{V}}$ and $\dot{\Gamma}$ and $\ddot{\Gamma}$ refers to the q and $q+1$ version of polinomials accordingly.

C Fuzzy designs

This section contains additional derivations for the fuzzy design. First, notice that the CLATE for RD tree is identified as,

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

Notice that the theoretical EMSE criterion is the same as for the sharp case, expressed in Equation (A.7). The difference is in the proposed estimator, where there is an additional component, weightening by the probability of the participation. We follow Calonico et al. (2014), who proposes estimators for both the bias and the variance via a linearization of the final quantity, hence $\tau^{FRD}(z;\Pi,h,\mathcal{S}^{est})$, instead of estimating the bias and the variance separately for the nominator and denominator.

First, let us write the parameter of interest with a tree structure Π,

$$\hat{\tau}^{FRD}(z;\Pi,h,\mathcal{S}^{est}) = \sum_{\ell=1}^{\#\Pi} \hat{\tau}_j^{FRD} \mathbb{1}_{\ell_j}(z;\Pi) = \sum_{\ell=1}^{\#\Pi} \frac{\hat{\tau}_j^y}{\hat{\tau}_j^t} \mathbb{1}_{\ell_j}(z;\Pi),$$

where $\hat{\tau}_j^y$, $j = 1,\dots,\#\Pi$ represent the nonparametric within-leaf RD estimators for the outcome and $\hat{\tau}_j^t$ for the participation. The bias term after linearization for each leaves is given:

$$\mathbb{B}[\hat{\tau}_j^{FRD}] = \mathcal{B}_j^{FRD} h^2 + o(h^2)$$

and the variance is,

$$\mathbb{V}[\hat{\tau}_j^{FRD}(\Pi,h)] = \mathcal{V}_j^{FRD} (N_j^{est} h)^{-1} + o((N_j^{est} h)^{-1})$$

where N_j^{est} is the number of observations in the estimation sample falling into leaf j. $\mathcal{B}_j^{FRD}$ and $\mathcal{V}_j^{FRD}$ are leaf-specific constants, defined by Calonico et al. (2014) in the supplemental appendix, under S.2.3. Similarly, as with sharp design we'll ignore the remainder terms from this point on. Using the same logic the estimator for the expected bias squared term is given by,

$$\mathbb{E}_{Z_i}\left\{ \left(\mathbb{B}_{\mathcal{S}^{est}}[\hat{\tau}^{FRD}(z;\Pi,h,\mathcal{S}^{est})]\right)^2 \right\} \approx h^{2q+2} \sum_{j=1}^{\#\Pi} \left(\mathcal{B}_j^{FRD}\right)^2 p_j^{est}.$$

using that $P(Z_i \in \ell_j)$ is approximated by $p_j^{est} = N_j^{est}/N^{est}$. Similarly, the expected variance is,

$$\mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}[\hat{\tau}^{FRD}(z;\Pi,h,\mathcal{S}^{est})]\right\} \approx \frac{1}{N^{est}h} \sum_{j=1}^{\#\Pi} \mathcal{V}_j^{FRD}\,.$$

For the term $\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\}$, we can use the bias corrected estimator proposed by Calonico et al. (2014) and to assess the behavior of the criterion we can disentangle the estimator that refers to the outcome and for the participation equation.

$$\tilde{\tau}^{FRD}(z;\Pi,h,\mathcal{S}^{est}) := \sum_{\ell=1}^{\#\Pi} \left(\hat{\tau}_j^{FRD} - \mathcal{B}_j^{FRD} h^2\right) \mathbb{1}_{\ell_j}(z;\Pi) = \sum_{\ell}^{\#\Pi} \left(\frac{\hat{\tau}_j^y}{\hat{\tau}_j^t} - \mathcal{B}_j^{FRD} h^2\right) \mathbb{1}_{\ell_j}(z;\Pi)\,. \tag{A.11}$$

We can also use the bias corrected variance for the fuzzy estimate,

$$\mathbb{V}_{\mathcal{S}^{est}}[\tilde{\tau}^{FRD}(z;\Pi,h,\mathcal{S}^{est})] = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\mathbb{V}\left[\tilde{\tau}_j^{FRD}(\Pi,h)\right] .$$

Note that the leaf-by-leaf variance $\mathbb{V}\left[\tilde{\tau}_j^{FRD}\right]$ is a scaled function of the variance of the outcome and participation estimators. Putting these together, we get a feasible EMSE criterion for the fuzzy case,

$$\begin{aligned}\widehat{EMSE}_\tau^{FRD}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi,h) =& h^{2q+2}\sum_{j=1}^{\#\Pi}\left(\mathcal{B}_j^{FRD}\right)^2 p_j^{est} + \frac{1}{N^{est}h}\sum_{j=1}^{\#\Pi}\mathcal{V}_j^{FRD} \\ & - \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left(\frac{\hat{\tau}_j^y}{\hat{\tau}_j^t} - \mathcal{B}_j^{FRD}h^2\right)^2(Z_i;\Pi,\mathcal{S},h) + \sum_{j=1}^{\#\Pi} p_j^{te}\mathbb{V}\left[\tilde{\tau}_j^{FRD}(\Pi,h)\right] .\end{aligned}$$

D Parametric Estimator

Nonparametric estimation is a data intensive method. Using a parametric estimator instead provide some benefits with smaller sample size, however comes with the cost of less flexible functional form. We derive the properties of the parametric OLS estimator here as well.

D.1 Sharp design with parametric estimator

First, let us assume a q-th order polynomial functional form in X for each leaf to identify τ_j. Each conditional expectation function – $\mathbb{E}\left[Y(d)|X=x,Z\in\ell_j(\Pi)\right], d\in\{0,1\}$ – is given by a q-th order polynomial, which ensures a flexible functional form. For a given Π, one can then write

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

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

The OLS estimator for the parameters is given by

$$\begin{aligned}\hat{\boldsymbol{\delta}}_j^{+,LS} &= \arg\min_{\boldsymbol{\delta}_j^{+,LS}} \sum_{i\in\mathcal{S}} \left\{ \mathbb{1}_c(X_i)\mathbb{1}_{\ell_j}(Z_i;\Pi)\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^{+,LS}\right)^2\right\} \\ \hat{\boldsymbol{\delta}}_j^{-,LS} &= \arg\min_{\boldsymbol{\delta}_j^{-,LS}} \sum_{i\in\mathcal{S}} \left\{ \left[1-\mathbb{1}_c(X_i)\right]\mathbb{1}_{\ell_j}(Z_i;\Pi)\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^{-,LS}\right)^2\right\} \ , \qquad \forall j\end{aligned} \tag{A.12}$$

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

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

In order to analyze the variance for the proposed estimator, let us write the model as

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

where ϵ_i is the idiosyncratic disturbance term. Now,

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

are the within-leaf variance estimators for the disturbance terms in leaf j with $N_{+,j}^{te}, N_{-,j}^{te}$ number of observations within the same leaf for above and below the threshold, respectively. $\hat{\epsilon}_i$ are the OLS residuals. For simplicity, here we assume the same finite variance within the leaves when deriving these estimators.[5] Furthermore, let the cross-product of the running variable above and below the threshold for leaf j be

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

[5] This can be seen as a leaf-level clustered homoskedastic error assumption.

Derivation of EMSE for parametric case

Now, let us turn to the derivation of the EMSE criterion. We use the same steps, however with parametric model assumption we can go further:

$$
\begin{aligned}
EMSE_\tau(\Pi) = \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}} & \left[MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi)\right] \\
& \vdots \\
& = \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(Z_i;\Pi) - \hat{\tau}^{LS}(Z_i;\Pi,\mathcal{S}^{est})\right)^2 \mid Z_i\right]\right\} - \mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
& \text{Note: } \mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}^{LS}(z;\Pi,\mathcal{S}^{est})\right] = \tau(z;\Pi) \text{ where } z \text{ is fixed, thus} \\
& \tau(Z_i;\Pi) = \mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}^{LS}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i} \overset{Z_i \perp\!\!\!\perp \mathcal{S}^{est}}{=} \mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}^{LS}(Z_i;\Pi,\mathcal{S}^{est}) \mid Z_i\right] \\
& = \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}^{LS}(Z_i;\Pi,\mathcal{S}^{est}) \mid Z_i\right] - \hat{\tau}^{LS}(Z_i;\Pi,\mathcal{S}^{est})\right)^2 \mid Z_i\right]\right\} - \mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
& = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^{LS}(Z_i;\Pi,\mathcal{S}^{est}) \mid Z_i\right]\right\} - \mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
& = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^{LS}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \qquad \square
\end{aligned}
$$

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

Expected variance of CATE

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

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

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

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

where

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

Same applies for the components of $\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,est}\right]$, but using observations, below the threshold, selected by $1-\mathbb{1}_c(X_i)$ instead of $\mathbb{1}_c(X_i)$.
Using these estimates, leads to the following expression for the variance, with scalar z,

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

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}^{LS}(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 for same leaf shares: }\frac{N^{te}_j}{N^{te}}\approx\frac{N^{est}_j}{N^{est}}\\
&\approx \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}/N^{est}_j}+\frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}/N^{est}_j}\right]e_1\right\}\\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left[e_1'\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}}+\frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}
\end{aligned}$$

where, N^{te}_j, N^{est}_j are the number of observations within leaf j for the test sample and estimation sample, respectively, and $p^{est}_{\pm,j}$ is the share of units above (+) and below (-) the threshold. This derivation uses the fact that observations are randomly assigned to the test sample and to the estimation sample, thus the leaf shares in the test sample (N^{te}_j/N^{te}) is approximately the same as in the estimation sample, (N^{est}_j/N^{est}).

Remarks:

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

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

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

Expected square of CATE

The second part of the EMSE criterion is the estimator for the expected squared of the true CATE, $\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right]$ over the test sample's features. Using, $\tau(z;\Pi) = \mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}^{LS}(z;\Pi,\mathcal{S}^{te})\right]$, where z is fixed, therefore $\tau(Z_i;\Pi) = \mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}^{LS}(z;\Pi,\mathcal{S}^{te})\right]\big|_{z=Z_i}$. Based on this fact, it follows:

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

The two parts can be estimated by two natural candidates. The expected square CATE is just the average of the squared CATE estimator given by the test sample. The expected variance term is similar to the previous, but note that the variance is estimated purely on the test sample. This means that the scaling factor for the number of observations is coming only from the test sample.

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

This expression is the same as the expected variance using the test sample, the only difference is the scalar N^{te} is used instead of N^{est}. The assumption for the same observation shares is used here to make the weights the same for the variance estimators.

The estimator for the expected value of the true squared CATE function over the test sample is given by,

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

The averaged squared treatment estimator prefers trees with many leaves. It is the sample analog for finding groups with different treatment effects. This term always increases as the number of leaves increases, while the average of the sum of squared treatment effects for two (or more) groups is always greater than the average of the sum of one averaged squared treatment effect. The second part is similar to the derived expected variance, but here the scaling for the average (N^{te}) comes from the test sample, as the estimator refers to the expected value over the test sample.[6] The weights of $p^{est}_{+,j}$ and $p^{est}_{-,j}$ come from the estimation sample, and they help the algorithm to avoid sample-specific splits.

Estimator for EMSE

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

$$\begin{aligned}\widehat{EMSE}^{LS}_{\tau}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left(\hat{\tau}(Z_i;\Pi,\mathcal{S}^{te})\right)^2 \\ &+\left(\frac{1}{N^{te}}+\frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}\end{aligned} \tag{A.13}$$

D.2 Relation to Athey & Imbens (2016)

(Athey & Imbens 2016, p. 7357) proposes similar criterion, but with unconfoundedness assumption, which leads to the following EMSE criterion:

$$\begin{aligned}\widehat{EMSE}^{AI}_{\tau}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left(\hat{\tau}^{AI}(Z_i;\Pi,\mathcal{S}^{te})\right)^2 \\ &+\left(\frac{1}{N^{te}}+\frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi}\left\{\frac{\hat{\sigma}^2_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}}{p^{est}_{-,j}}\right\},\end{aligned} \tag{A.14}$$

where $\hat{\tau}^{AI}(Z_i;\Pi,\mathcal{S}^{te})$ is defined as the estimator for within leaf differences for treated and control units (sample average differences) and $\hat{\sigma}^2_{\pm,j}$ are the sample variances for the treated

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

and control unit averages. Comparing Equation A.13 to A.14, it is easy to see that if the parametric regression model contains intercept only, we get back the criterion provided by A.14.

D.3 Fuzzy design with parametric estimator

Let us recall the identification of the CLATE parameter in the fuzzy design,

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

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

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

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

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

where $\boldsymbol{\delta}_j^{+,t}, \boldsymbol{\delta}_j^{-,t}, \boldsymbol{\delta}_j^{+,y}$ and $\boldsymbol{\delta}_j^{-,y}$ estimated by OLS:

$$\begin{aligned}
\hat{\boldsymbol{\delta}}_j^{+,t} &= \arg\min_{\boldsymbol{\delta}_j^{+,t}} \sum_{i\in\mathcal{S}} \mathbb{1}_c(x)\mathbb{1}_{\ell_j}(z;\Pi)\left(T_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^{+,t}\right)^2\\
\hat{\boldsymbol{\delta}}_j^{-,t} &= \arg\min_{\boldsymbol{\delta}_j^{-,t}} \sum_{i\in\mathcal{S}} (1-\mathbb{1}_c(x))\mathbb{1}_{\ell_j}(z;\Pi)\left(T_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^{-,t}\right)^2\\
\hat{\boldsymbol{\delta}}_j^{+,y} &= \arg\min_{\boldsymbol{\delta}_j^{+,y}} \sum_{i\in\mathcal{S}} \mathbb{1}_c(x)\mathbb{1}_{\ell_j}(z;\Pi)\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^{+,y}\right)^2\\
\hat{\boldsymbol{\delta}}_j^{-,y} &= \arg\min_{\boldsymbol{\delta}_j^{-,y}} \sum_{i\in\mathcal{S}} (1-\mathbb{1}_c(x))\mathbb{1}_{\ell_j}(z;\Pi)\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^{-,y}\right)^2
\end{aligned}$$

Estimator for CLATE parameter based on these polynomial functions is given by

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

and its variance:

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

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

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

We use the same expected MSE criterion for fuzzy design as well,

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

One can construct estimators for these two terms. The variance part from the expected variance is

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

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

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

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 $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, one can estimate all the variances and covariances using the observations from the test sample and use only the additional knowledge on the number of observations in the estimation sample. In the simplest – finite variances of the error terms within each leaf – one can write the following sample analogs (below threshold units it is similar).

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

where

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

Remark: the number of observations and the inverse of the running variable's product is the same for both treatment and outcome equations. It is also easy to use other variance estimators (e.g., heteroscedastic-robust versions or clustered), see Appendix D.4.

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,LS}(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] \\
&\quad + \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] \\
&\quad - 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. \\
&\qquad + \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] \\
&\qquad \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

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

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

The expected value of this variance over Z_i from the test sample can be calculated similarly as in the sharp RDD case.

$$\begin{aligned}\hat{\mathbb{E}}_{Z_i}\left\{\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}^{FRD,LS}(z;\Pi,\mathcal{S}^{est})\right]\big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left\{\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\left(\frac{\mathcal{V}_{+,j}}{N^{est}_{+,j}} + \frac{\mathcal{V}_{-,j}}{N^{est}_{-,j}}\right)e_1\right\}\\ &\approx \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left(\frac{\mathcal{V}_{+,j}}{p^{est}_{+,j}} + \frac{\mathcal{V}_{-,j}}{p^{est}_{-,j}}\right)e_1\right\}\end{aligned}$$

The second part of the EMSE criterion is the estimator for the expected squared $\tau^2_{FRD}(Z_i;\Pi)$. Similarly to sharp RD, one can construct the following estimator,

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

Note, here everything is estimated on the test sample and we used the assumption, that the number of unit shares for below and above the threshold – for all leaf – are approximately the same in the estimation and test sample ($p^{te}_{+,j} \approx p^{est}_{+,j}, p^{te}_{-,j} \approx p^{est}_{-,j}$). The feasible criteria for fuzzy design for EMSE:

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

where

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

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 ($\hat{\tau}^y_j(\cdot)$). $\sigma^{2,t}_{\pm,j}$, $\sigma^{2,y}_{\pm,j}$ and $C^{y,t}_{\pm,j}$ are estimators for the variances and co-variance for the leaf-by-leaf disturbance terms.

D.4 Different variance estimators for parametric criterion

Homoscedastic error assumption is rather a strong assumption in RD context, thus the use of different heteroscedastic consistent estimators is favorable. First, we show derivation of $\widehat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est}\right]$ – the other parts can be calculated similarly – then we put together with the other parts.

General case:

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

Estimators are different in how to calculate $\hat{\Sigma}_{+,j}$:

White's estimator ('HCE0'):

$$\hat{\Sigma}^{HCE0}_{+,j} = \frac{1}{N^{te}_{+,j}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\boldsymbol{X}_i(\epsilon_i^y)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

Adjusted 'HCE1':

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

In case of clustered SE, with HC1

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

where $G^{te}_{+,j}$ is the number of clusters in leaf j above the threshold in the test sample. The variance estimators are similarly constructed for parameters below the threshold.

In sharp RD, one gets the variance estimator as,

$$\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{SRD}(z;\Pi,\mathcal{S}^{est})\right] = \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(Z_i;\Pi)e_1'\left\{\frac{\hat{M}_{+,j}^{-1}\hat{\Sigma}_+\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}+\frac{\hat{M}_{-,j}^{-1}\hat{\Sigma}_-\hat{M}_{-,j}^{-1}}{N_{-,j}^{est}}\right\}e_1$$

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 CLATE parameters,

$$\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_F(z;\Pi,\mathcal{S}^{est})\right] &= \left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_-^y(c,z;\Pi,\mathcal{S}^{est})\right]\right)\\
&+ A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_-^t(c,z;\Pi,\mathcal{S}^{est})\right]\right)\\
&- 2A_3(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right]\\
&\qquad\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}}_j^{+,y,est}\right]e_1 + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,y,est}\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}}_j^{+,t,est}\right]e_1 + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,t,est}\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}}_j^{+,y,est},\hat{\boldsymbol{\delta}}_j^{+,t,est}\right]e_1\right.\\
&\qquad\qquad\left.+\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,y,est},\hat{\boldsymbol{\delta}}_j^{-,t,est}\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}}_j^{+,y,est}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,y,est}\right]\right)\right.\\
&\qquad + A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,t,est}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,t,est}\right]\right)\\
&\qquad \left.-2A_3\left(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est},\hat{\boldsymbol{\delta}}_j^{+,t,est}\right] + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,y,est},\hat{\boldsymbol{\delta}}_j^{-,t,est}\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}}_j^{+,y,est}\right] + A_2\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,t,est}\right] - 2A_3\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est},\hat{\boldsymbol{\delta}}_j^{+,t,est}\right]\right.\\
&\qquad\qquad \left.+A_1\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,y,est}\right] + A_2 + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,t,est}\right] - 2A_3\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,y,est},\hat{\boldsymbol{\delta}}_j^{-,t,est}\right]\right\}e_1\\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{\frac{1}{N_{+,j}^{est}}\hat{M}_{+,j}^{-1}\left(A_1\hat{\Sigma}_{+,j}^y + A_2\hat{\Sigma}_{+,j}^t - 2A_3\hat{C}_{+,j}\right)\hat{M}_{+,j}^{-1}\right.\\
&\qquad\qquad \left.+\frac{1}{N_{-,j}^{est}}\hat{M}_{-,j}^{-1}\left(A_1\hat{\Sigma}_{-,j}^y + A_2\hat{\Sigma}_{-,j}^t - 2A_3\hat{C}_{-,j}\right)\hat{M}_{-,j}^{-1}\right\}e_1\\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{\frac{1}{N_{+,j}^{est}}\hat{M}_{+,j}^{-1}\hat{\Sigma}_+^*\hat{M}_{+,j}^{-1} + \frac{1}{N_{-,j}^{est}}\hat{M}_{-,j}^{-1}\hat{\Sigma}_-^*\hat{M}_{-,j}^{-1}\right\}e_1
\end{aligned}$$

This result is quite useful: there is no need to calculate and multiply with $\hat{M}_{\pm,j}^{-1}$ multiple times during calculating the variances, but they can be 'added up', using 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.

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

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 $k = 1, \dots, K$ features such that $Z_{i,k}$ is independent across k and independent from ϵ_i, X_i. The source of variation comes from ϵ_i during the simulations, thus $X_i, Z_{i,k}$ are the same across the Monte Carlo samples. All the other terms are dependent on the setup.
We report three Monte Carlo average statistics to evaluate the performance of the algorithm:

1. Average of the infeasible MSE

$$infMSE := MC^{-1} \sum_{mc=1}^{MC} \left(N^{eval}\right)^{-1} \sum_{i=1}^{N^{eval}} (\kappa(Z_{i,k}) - \tilde{\tau}_{mc}(Z_i; \hat{\Pi}(\mathcal{S}^{tr}), h^*, \mathcal{S}^{est}_{mc}))^2$$

2. Monte-carlo average bias,

$$Bias := MC^{-1} \sum_{mc=1}^{MC} \left(N^{eval}\right)^{-1} \sum_{i=1}^{N^{eval}} \left[\kappa(Z_{i,k}) - \tilde{\tau}_{mc}(Z_i; \hat{\Pi}(\mathcal{S}^{tr}), h^*, \mathcal{S}^{est}_{mc})\right]$$

3. 95 % empirical coverage ratio

$$Coverage := MC^{-1} \sum_{mc=1}^{MC} \sum_{i=1}^{N^{eval}} \mathbb{1}\left(\tau(Z_{i,k}; \widehat{\Pi}_{mc}) \in \left[\tilde{\tau}_{i,mc} - z_{0.025}\widehat{SE}[\tilde{\tau}_{i,mc}], \tilde{\tau}_{i,mc} + z_{0.975}\widehat{SE}[\tilde{\tau}_{i,mc}]\right]\right),$$

   where $\tilde{\tau}_{i,mc} = \tau(Z_i; \hat{\Pi}(\mathcal{S}^{tr}), h^*, \mathcal{S}^{est}_{mc})$ and $\tau(Z_{i,k}; \widehat{\Pi}_{mc})$ is the CATE in step-function format, where the step-function is given by the estimated tree $\widehat{\Pi}_{mc}$.

4. Average number of leaves in the final tree.

$$\#\bar{\Pi} := MC^{-1} \sum_{mc=1}^{MC} \#\hat{\Pi}_{mc}$$

DGP 1-3 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 setup to investigate the effect of excess curvature. For all three designs the running variable is generated by $X_i \sim (2\mathcal{B}(2,4)-1)$, where $\mathcal{B}$ denotes a beta distribution and the disturbance term has the variance of $\sigma_\epsilon^2 = 0.05$. The threshold value is set to $c = 0$.

DGP 1: 'An Alternative DGP' by Calonico et al. (2014) adds extra curvature to the functional form. This design is exactly the same as in Calonico et al. (2014), thus it has a homogeneous treatment effect. 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 $Z_{i,1} = 1$. Treatment effect and $\eta(\cdot)$ is homogeneous.

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

$$\kappa = 0.04$$

DGP 2: Imitating Lee (2008) vote-shares. We assume two treatment effects and heterogeneous $\eta(\cdot)$. The features are the same as in DGP-1. We set the functional form as in Calonico et al. (2014) first setup, but tweak the CEFs, such that the two different subgroups has slightly different CEFs as well both below and above the treshold.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 0.48 + 1.27X_i + 7.18X_i^2 + 20.21X_i^3 + 21.54X_i^4 + 7.33X_i^5, & \text{if } X_i < 0, Z_{i,1} = 1 \\ 0.48 + 2.35X_i + 8.18X_i^2 + 22.21X_i^3 + 24.14X_i^4 + 8.33X_i^5, & \text{if } X_i < 0, Z_{i,1} = 0 \\ 0.48 + 0.84X_i - 3.00X_i^2 + 7.99X_i^3 - 9.01X_i^4 + 3.56X_i^5, & \text{if } X_i \geq 0, Z_{i,1} = 1 \\ 0.48 + 1.21X_i - 2.90X_i^2 + 6.99X_i^3 - 10.01X_i^4 + 4.56X_i^5, & \text{if } X_i \geq 0, Z_{i,1} = 0 \end{cases}$$

$$\kappa(Z_{i,1}) = 0.02 \times Z_{1,i} + 0.08 \times (1 - Z_{1,i})$$

DGP 3: Ludwig & 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 involved in the analysis.

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

$$\kappa(Z_{i,1}) = -0.45 + 0.5Z_1 - 0.25Z_1^2 + 0.1Z_1^3;$$

E.1 Monte Carlo simulation for fuzzy design

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

$$Pr(T_i = 1|X_i) = \Phi\left(2X_i - 6X_i^2 + 3X_i^3 + 10 \times \mathbb{1}_c(X_i)\right)$$

This definition gives a nonlinear CEF for the participation equation, while there is positive probability of taking the treatment below c, but zero probability of not-taking the treatment above $X_i = 0$. This is not needed in general, but simplifies our analysis. For simplicity we use 'DGP-x-f' expression for referring these fuzzy setups. Table A.1 shows the same algorithm performance measures and the evidence on valid inference similarly to the sharp design. The results are aligned with the conclusion reported in Section 4, but the fuzzy design is even more data intensive.

Table A.1: Monte Carlo averages for performance measures in fuzzy designs

		Tree				Forest		
Design	Sample Size	infMSE	Bias	Coverage	$\#\Pi$	infMSE	Bias	Coverage
	$N = 1000$	479.21	2.2574	0.4180	1.00	115,378	130.71	0.0380
DGP-1	$N = 5000$	0.0030	-0.0019	0.9640	1.06	0.0031	0.0039	0.9440
	$N = 10000$	0.0094	0.0008	0.9270	1.16	4.1823	0.0516	0.8434
	$N = 50000$	0.0008	-0.0042	0.9355	1.09	0.0003	-0.0008	0.9196
	$N = 1000$	284.65	-0.7618	0.5710	1.00	228,947	-67.926	0.8950
DGP-2	$N = 5000$	0.0037	-0.0008	0.9530	1.23	0.0022	0.0000	0.9035
	$N = 10000$	0.0133	0.0209	0.8710	1.55	0.0035	0.0293	0.8225
	$N = 50000$	0.0011	-0.0055	0.9600	1.55	0.0007	0.0008	0.9358
	$N = 1000$	96,050	309.57	1.0000	1.00	$2{,}973 \times 10^8$	545,252	0.0000
DGP-3	$N = 5000$	188.426	-4.0452	1.0000	2.00	205.15	-1.8787	0.8234
	$N = 10000$	95.2737	-2.6172	0.8551	7.09	79.1468	-0.0362	0.9829
	$N = 50000$	27.9040	0.6607	0.9563	9.60	14.0076	0.5399	0.9692

Number of true leaves: $\#\Pi_{DGP-1} = 1$, $\#\Pi_{DGP-2} = 2$, $\#\Pi_{DGP-3} = Inf$. Algorithm setup: using the smallest cross-validation value to select the pruning parameter $\gamma^*_{h_r}$. For bandwidth selection (h^*) we use grid-search around the bandwidth for the homogeneous treatment effect with coverage error optimal bandwidth with 10 candidates. We use local linear model ($q = 1$) for each leaf and $q + 1$ for bias correction. Bandwidth for higher order polynomial is selected by $\rho = 1$. Leaf-by-leaf variance estimators are using the HCE-1 formula.

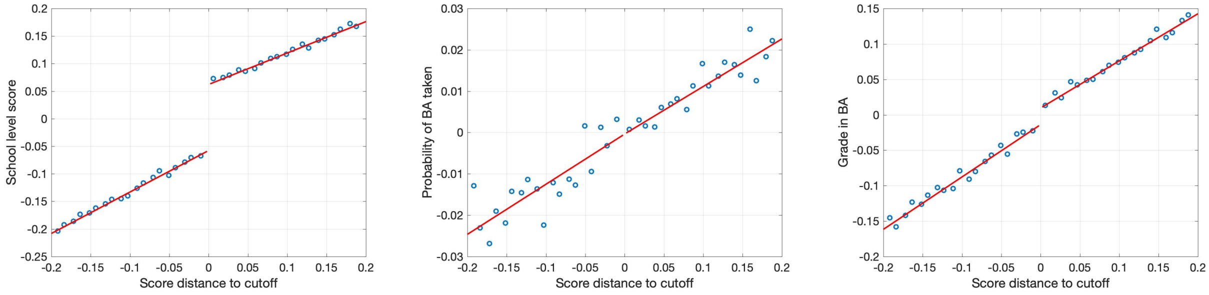
F Additional results on the empirical exercise

This part adds additional information to the empirical analysis. Table A.2 shows the descriptives for the used variables in the heterogeneity analysis. Here, we only present the variables used for revisiting the heterogeneity analysis by Pop-Eleches & Urquiola (2013).

	School level average transition score	Baccalaureate taken	Baccalaureate grade	Scaled School admission score	Number of schools in town
Mean	7.65	0.74	8.12	0.10	17.50
Median	7.55	1.00	8.15	0.15	17.00
Std deviation	0.75	0.44	0.90	0.55	7.49
Min	5.78	0.00	5.19	-1.00	2.00
Max	9.63	1.00	10.00	1.00	29.00
N	1,857,376	1,857,376	1,256,038	1,857,376	1,857,376

Table A.2: Descriptive statistics of the variables used in heterogeneity analysis of Pop-Eleches & Urquiola (2013)

Figure A.4 and Table A.3 summarizes the original RD results from Pop-Eleches & Urquiola (2013). In Figure A.4 for all three graphs, the horizontal axis represents the running variable, which is a student's standardized transition score subtracting the school admission cut-off. The vertical axis in Figure A.4a) represents the *peer quality*, that each admitted student experiences, when going to school. Peer quality is defined as the average transition score for the admitted students in each school. This indicates that the higher the level of average transition score is (e.g., the admitted students performed great in the nationwide test), the better the peer quality. Figure A.4b) shows the probability of a student taking the Baccalaureate exam, while Figure A.4c) plots the Baccalaureate exam grade among exam-takers. In all outcomes, school fixed effects are used as in Pop-Eleches & Urquiola (2013), thus the vertical axis is centered around 0 for all plotted outcomes. Both left and right graphs show a jump in the average outcome at the discontinuity point, but the jump in the exam-taking rate is quite noisy and seemingly insignificant. As a next step we revisit the same heterogeneity analysis, but now use nonparametric estimators instead of parametric RD. We use `rdrobust` package to estimate the effects. The bandwidth is MSE optimal and ρ is estimated, thus not fixed at 1. It is interesting to compare the differences between Table A.3 and A.4. The sign does not changes, and significance does not disappear, thus parametric estimates are fairly robust, however the point estimates are significantly different from each other and with nonparametric estimation they are higher for school level average transition score and mixed for BA grade. Next, we employ our algorithm to the probability of taking the BA and we find 16 leaves, producing slightly different estimates. Out of these 16 leaves 6 subgroups has positive effects on the probabilities and the other 10 are negatives. Figure A.5 shows the different point estimates along the two dimensions: school average transition score and number of schools in town. Note that the unified bandwidth is 0.0836. We have also checked for the outcome of BA grade, but have not find



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

Figure A.4: Bin-scatter for main (pooled) RD results of Pop-Eleches & Urquiola (2013), using school fixed effects

	School level average transition score	Baccalaureate taken	Baccalaureate grade
Full sample			
τ_0	0.107***	0.000	0.018***
$SE(\tau_0)$	(0.001)	(0.001)	(0.002)
N	1,857,376	1,857,376	1,256,038
Top tercile			
τ_1	0.158***	0.003	0.048***
$SE(\tau_1)$	(0.002)	(0.002)	(0.003)
N_1	756,141	756,141	579,566
Bottom tercile			
τ_2	0.099***	−0.008*	−0.005
$SE(\tau_2)$	(0.003)	(0.004)	(0.009)
N_2	392,475	392,475	212,282
Towns with four or more schools			
τ_1	0.097***	0.000	0.016***
$SE(\tau_1)$	(0.001)	(0.001)	(0.002)
N_1	1,806,411	1,806,411	1,223,341
Towns with three schools			
τ_2	0.333***	−0.007	0.028*
$SE(\tau_2)$	(0.007)	(0.009)	(0.016)
N_2	31,149	31,149	19,877
Towns with two schools			
τ_3	0.697***	0.020	0.179***
$SE(\tau_3)$	(0.010)	(0.013)	(0.023)
N_3	19,816	19,816	12,820

Notes: All regressions are clustered at the student level and include cutoff fixed effects. Standard errors are in parentheses. All estimates present reduced form specifications where the key independent variable is a dummy for whether a student's transition score is greater than or equal to the cutoff.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table A.3: Heterogeneity in Baccalaureate Effects - Pop-Eleches & Urquiola (2013), Table 5

	School level average transition score	Baccalaureate taken	Baccalaureate grade
Full sample			
τ_0	0.127***	−0.002	0.027***
$SE(\tau_0)$	(0.003)	(0.003)	(0.006)
h	0.207	0.289	0.213
N	1,857,376	1,857,376	1,256,038
Top tercile			
τ_1	0.132***	0.002	0.037***
$SE(\tau_1)$	(0.004)	(0.004)	(0.008)
h	0.263	0.359	0.221
N_1	756,141	756,141	579,566
Bottom tercile			
τ_2	0.117***	−0.007	0.035**
$SE(\tau_2)$	(0.007)	(0.008)	(0.016)
h	0.156	0.273	0.228
N_2	392,475	392,475	212,282
Towns with four or more schools			
τ_1	0.114***	−0.002	0.024***
$SE(\tau_1)$	(0.003)	(0.003)	(0.006)
h	0.206	0.307	0.217
N_1	1,806,411	1,806,411	1,223,341
Towns with three schools			
τ_2	0.414***	−0.013	0.083**
$SE(\tau_2)$	(0.018)	(0.021)	(0.038)
h	0.251	0.250	0.251
N_2	31,149	31,149	19,877
Towns with two schools			
τ_3	0.715***	0.028	0.171***
$SE(\tau_3)$	(0.023)	(0.0024)	(0.043)
h	0.319	0.325	0.366
N_3	19,816	19,816	12,820

Notes: All regressions are clustered at the student level and include cutoff fixed effects. Standard errors are in parentheses. All estimates present reduced form specifications where the key independent variable is a dummy for whether a student's transition score is greater than or equal to the cutoff. Using local linear regression for estimation and MSE optimal bandwidths. Estimates are bias-corrected values.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table A.4: Replication of Heterogeneity in Baccalaureate Effects with nonparametric estimation

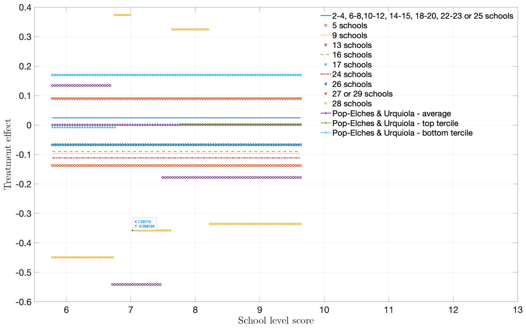


Figure A.5: CATE estimate for probability of taking the examination

any relevant heterogeneity. On the estimation subsample we have estimated 0.0343(0.0119) effect, that is statistically the same as for nonparametric estimator. (The bandwidths are also similar, our optimal bandwidth is 0.2247.

	Mean	Median	Std. dev.	Min	Max	N
Outcome and running variables						
School level average transition score	8.20	8.29	0.60	6.53	9.41	11,931
Scaled Admission score	0.85	0.82	0.97	-2.07	3.91	11,931
Socioeconomic characteristics of households						
Female head of household (d)	0.89	1	0.32	0	1	11,931
Age of head of household	46.75	45	7.15	13	97	11,843
Romanian (d)	0.94	1	0.24	0	1	11,931
Hungarian (d)	0.05	0	0.22	0	1	11,931
Gypsy (d)	0.01	0	0.06	0	1	11,931
Other Ethnicity (d)	0.01	0	0.09	0	1	11,931
HH's Primary education (d)	0.66	1	0.47	0	1	11,840
HH's Secondary education (d)	0.20	0	0.40	0	1	11,840
HH's Tertiary education (d)	0.13	0	0.34	0	1	11,840
Socioeconomic characteristics of students						
Gender of student (d)	0.42	1	0.49	0	1	11,931
Age of student	18.08	18	0.94	14	23	11,866
Accessibility of households to goods						
Car (d)	0.57	1	0.49	0	1	11,820
Internet (d)	0.73	1	0.44	0	1	11,829
Phone (d)	0.47	0	0.50	0	1	11,807
Computer (d)	0.87	1	0.34	0	1	11,851
Parental and Child responses to survey questions						
Parent volunteered (d)	0.11	0	0.31	0	1	11,868
Parent paid tutoring (d)	0.24	0	0.42	0	1	11,931
Parent helps HW (d)	0.20	0	0.40	0	1	11,815
Child does HW every day - Parent (d)	0.75	1	0.43	0	1	11,779
Negative interactions with peers	0.12	0	0.37	0	5	11,838
Child does HW every day - Child (d)	0.63	1	0.48	0	1	11,908
HW percieved easy	5.45	5.60	1.02	1	7	9,628
Characteristics of schools						
No. schools	2.33	2	0.50	2	4	11,931
2 schools (d)	0.69	1	0.46	0	1	11,931
3 schools (d)	0.29	0	0.45	0	1	11,931
4 schools (d)	0.02	0	0.13	0	1	11,931
Highest certification teacher in school (d)	0.61	1	0.49	0	1	11,169
Novice teacher in school (d)	0.06	0	0.24	0	1	11,169

(d) indicates it is a dummy variable. 'HH' stands for household, 'HW' for homework.

Table A.5: Descriptive statistics of the used variables for exploring heterogeneity in a survey-based dataset

Table A.5 shows the descriptives for each candidate features used to find the tree shown by Figure

3.

Finally, let us note that the results are quite robust to randomization of the observations in the training/estimation sample. Some of the splits may vary but the main conclusion is similar in most of the cases.

References

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

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

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

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

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

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

Ludwig, J. & 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.

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

Ripley, B. D. (1996), *Pattern recognition and neural networks*, Cambridge university press.