UNIVERSITÀ DEGLI STUDI DI MACERATA

CORSO DI DOTTORATO DI RICERCA IN
Quantitative Methods for Policy Evaluation

CICLO XXXVIII

TITOLO DELLA TESI
REGRESSION DISCONTINUITY DESIGN:
THEORETICAL INSIGHTS AND EMPIRICAL APPLICATIONS TO EDUCATIONAL DATA

SUPERVISORE DI TESI
Chiar.mo Prof. Cristina Davino
Chiar.mo Prof. Domenico Vistocco

DOTTORANDO
Dott. Pasquale Sannino

COORDINATORE
Chiar.mo Prof. Margherita Scoppola

ANNO 2025-2026

SDA
scuola di dottorato

Index

Introduction

In the scientific literature, the debate on inequality has long been a central theme—not only because of its ethical and theoretical relevance, but also due to its numerous social and economic implications. One of the most widely studied forms of inequality concerns education. Education plays a crucial role, as it significantly affects various dimensions of individuals' lives—including employment, health, and income—which are precisely the channels through which major social inequalities emerge.

A key determinant of educational outcomes is socioeconomic status. The debate on how socioeconomic status affects educational opportunities dates back to the 1960s, when this issue was first brought to light by what can be considered a seminal work: the Coleman Report (Coleman et al., 1966). This report showed, for the first time, that a student's academic performance is influenced not only by the family background, but also by the socioeconomic environment of the peers.

The literature on peer effects—which moves beyond a strictly individual view of student characteristics and focuses on the role of the social context and classroom or school composition—has been growing steadily. The underlying idea is that peers influence academic achievement through mechanisms such as shared norms of effort, mutual expectations, behavioral models, and information exchange (Manski, 1993; Sacerdote, 2001; Epple and Romano, 2011). However, rigorously demonstrating a causal peer effect is extremely challenging.

Given these considerations, it becomes clear that assessing and measuring individuals' educational outcomes is essential to identify potential issues within the school system and to understand where and how policy interventions may be needed. In Italy, this is achieved through the INVALSI (National Institute for the Evaluation of the Education and Training System) tests, which provide standardized assessments of core competencies in Italian, Mathematics, and English. These tests are administered to all students in specific grades of primary and secondary education. The INVALSI assessments also include an anonymous questionnaire, which collects information on students' characteristics and family background. This enables researchers to reconstruct each student's socioeconomic profile. The indicator used to summarize this profile is the so-called *ESCS* (Economic, Social, and Cultural Status). The ESCS is a standardized index derived from a principal component analysis of three indicators proposed by INVALSI: i) the parents' occupational status, ii) the parents' level of education, iii) the possession of certain material goods. The ESCS is measured at the individual level and then aggregated into the class and school

ESCS by taking the average. The Ministry of Education and Merit introduced, via Ministerial Decree No. 90 of May 19, 2023, an official threshold for the school-level ESCS indicator. For upper secondary schools, this value is set at -0.31243, meaning that, all schools with an ESCS value below the threshold should receive special attention, as they host students living in at-risk family and economic conditions. In general, the INVALSI data represent an important instrument for studying educational inequalities in Italy, as they allow the joint analysis of individual performance and contextual factors such as school environment and territorial disparities. Numerous empirical studies have highlighted the persistent performance gap between the northern and southern regions and the strong association between the' socioeconomic background of students and their academic achievements (Agasisti, 2014; Agasisti and Vittadini, 2012).

In such complex contexts, it becomes evident that there is a need for tools capable of isolating the causal effect of certain factors or phenomena on others. This is made possible through causal inference, whose goal is to evaluate the causal effect of a treatment on an outcome of interest, where "treatment" refers to the implementation of a policy, program, or, more generally, an intervention (Pearl, 2000; Spirtes et al., 2000; Elster, 1983). Causal analysis is straightforward when the treatment is assigned randomly, since randomization guarantees comparability between treated and control units: belonging to either group is analogous to a coin flip. However, for ethical or practical reasons, random assignment is often not feasible.

In the absence of randomization, the *Regression Discontinuity (RD)* design offers a viable alternative. Introduced by Thistlethwaite and Campbell (1960), RD estimates treatment effects in non-experimental settings. It is considered one of the most credible quasi-experimental methods for causal inference, and its use has expanded considerably in recent decades. It is now widely adopted in Economics, Sociology, Education, Political Science, Epidemiology, Criminology, and other fields (Cattaneo et al., 2020; Cattaneo et al., 2024; Imbens and Lemieux, 2008).

In a classic Regression Discontinuity design, each unit is associated with a *score* (also called running variable, forcing variable, or index), and treatment is assigned if this score exceeds a known threshold or *cutoff*. The intuition behind the design is that individuals with scores just below the cutoff (who do not receive the treatment) provide a valid comparison group for those just above it (who do receive the treatment). The discontinuity in outcomes at the threshold then identifies the causal effect of the treatment. Thus, in RD, identification of the causal effect relies on the existence of a known threshold determining treatment assignment. Once the general functioning of this approach is understood, it can be applied to the empirical setting

of this thesis.

Since one of the objectives of this thesis is to analyze the role of the school's socio-economic context in determining Italian students' academic performance, the Regression Discontinuity design is particularly well suited for this purpose. It allows for a credible estimation of the causal effect of the socio-economic environment by exploiting the ministerial threshold of the ESCS indicator as the treatment assignment criterion, thereby isolating the actual impact of the school background from other individual or family-related components. As mentioned earlier, for the 2023–2024 school year, the Ministry of Education and Merit established a threshold to classify schools. This threshold serves as the cutoff used to identify schools with a "low" socio-economic status. Consequently, all schools with an ESCS below this threshold are considered treated, while those above constitute the control group. Treatment assignment—attending a school with low ESCS—is therefore not determined by individual choice, but by a clear, observable ministerial rule. Using Regression Discontinuity, it is possible to estimate, in a causal framework, the peer effect of attending a low-ESCS school by comparing the Mathematics outcomes of students located just above and just below the threshold. As variable of outcome, we use the Mathematics score, as it represents one of the most reliable and commonly employed measures for assessing basic cognitive skills, as well as an important predictor of future success in academic and professional contexts (Hanushek and Woessmann, 2012; OECD, 2019).

However, applying RD to the INVALSI data presents an important methodological issue: the running variable, ESCS, is not continuous, but discrete[1], taking on values commonly referred to in the literature as *mass points*. The presence of mass points reduces the effective sample size. Indeed, RD requires sufficient variation in the running variable to construct local approximations on either side of the cutoff. What matters, therefore, is not the total number of observations, but the number of distinct values assumed by the running variable. This results in the so-called *sample size inflation* problem. A common solution in the literature is to aggregate the data at the level of mass points by taking the mean outcome of all observations sharing the same running variable value (Cattaneo et al., 2024). This avoids sample size inflation by producing a new dataset in which the number of observations equals the number of mass points, rather than the total number of raw observations.

However, when multiple individuals share the same value of the running vari-

[1]The ESCS indicator is defined at the school level, meaning that all students attending the same school share the same value. Although, in theory, ESCS values can take any real number, in practice they are finite and correspond to the number of schools in the sample, effectively making the variable discrete.

able, summarizing them exclusively via the mean can be reductive and may fail to capture the actual distribution of the outcome. In the presence of skewness, heavy tails, or outliers, the mean is particularly sensitive to extreme values, potentially leading to biased or unstable estimates. As a consequence, part of the within-group heterogeneity may be concealed or misrepresented, compromising the accuracy of the analysis.

For this reason, two alternatives to the simple mean that are more robust in the presence of outliers within mass points are proposed: the median and the medoid (Sannino et al., 2025a; Sannino et al., 2025c; Sannino et al., 2025b). The use of robust measures such as the median or medoid allows for a more faithful representation of the underlying distribution, especially when mass points exhibit asymmetric or heavy-tailed structures.

To evaluate the impact of mass points in Regression Discontinuity designs with discrete running variables, a Monte Carlo simulation study is also proposed in this thesis. The aim is not only to compare different collapsing strategies—aggregation using the mean, the median, or the medoid—but, more importantly, to provide guidance on choosing the most appropriate approach given the structure of the data.

Nonetheless, RD alone is not sufficient to address another key characteristic of the data: their hierarchical structure. Students are nested within schools, which are in turn nested within regions–or any other territorial aggregate. To account for this additional level of complexity, Multilevel Models are introduced. The goal of multilevel regression is to estimate predictive models while taking into account the hierarchical organization of the data. If on one hand, RD allows for estimating the Local Average Treatment Effect (LATE) at the cutoff; on the other hand, multilevel models allow us to study how such an effect varies across higher-level units. Combining the two approaches makes it possible to estimate both the treatment effect and how it varies depending on the grouping structure.

The approach proposed in this thesis aims to go beyond standard practices for implementing Multilevel Regression Discontinuity in educational settings—such as those described by Luyten (2006) and Steinmann and Olsen (2006)—by integrating both traditional and more advanced RD procedures, including optimal bandwidth selection (the window of observations considered around the cutoff) and kernel weighting of observations. To this end, three different methods are proposed: the *General Bandwidth Method*, the *Average Bandwidth Method*, and the *Weighted Average Bandwidth Method*. The three methods differ in how the optimal bandwidth is selected and how this bandwidth is weighted across groups. To assess the strengths

and weaknesses of these methods, a second simulation study was conducted to compare their performance under various conditions.

Summarizing the thesis, the present work aims to address several methodological issues related to Regression Discontinuity designs by proposing new approaches and applying them to real data. Thus, it has a twofold objective: on the one hand, the main goal is to contribute to the methodological debate on the extension of RD in the presence of discrete running variables and heterogeneous contexts; on the other hand, a second goal is to provide empirical evidence on the role of the school socio-economic context in determining Italian students' Mathematics performance, with particular attention to territorial differences between the North, Centre, and South of the country. From this perspective, the research seeks to offer useful tools both to scholars interested in causal evaluation and to policymakers committed to reducing educational inequalities.

The thesis is structured in 5 chapters. Chapter 1, titled "Introduction to causal inference", introduces the theoretical framework of causal inference. It discusses the goals of causal analysis and the methods used to achieve them, with a focus on Randomized Controlled Trials, considered the gold standard for causal identification. Their functioning and limitations are discussed. The chapter then presents the core technique of this thesis—the Regression Discontinuity design—through a literature review. It traces its origins, evolution, and rapid diffusion across various fields, highlighting its methodological flexibility. The main extensions of RD are outlined, including Sharp and Fuzzy RD, Geographic RD, Regression Discontinuity in Time, and others.

Chapter 2, titled "Estimation procedure in Regression Discontinuity design", focuses on the estimation process in RD. It provides a detailed explanation of the steps involved in implementing the design, with particular emphasis on the Local Polynomial Approach, the most commonly used method in RD. Special attention is given to the optimal bandwidth selection algorithm—a crucial phase in the procedure—the underlying assumptions of the model, and the associated diagnostic tests. The chapter concludes with an introduction to Multi-Dimensional RD settings, including multiple cutoffs, cumulative cutoffs, and multiple scores.

Chapter 3, titled "Investigating the effect of economic, social, and cultural conditions on students' performance: insights from INVALSI Tests", applies the RD framework to the empirical case study of INVALSI tests. The aim is to assess the role of school socio-economic status in shaping Mathematics outcomes among Italian students. The analysis uses INVALSI test data administered to grade 13 (final-year high school) students between March and May of the 2023/2024 school year. The

running variable for treatment assignment is the school's *Economic, Social, and Cultural Status (ESCS)*.

Chapter 4, titled "Regression Discontinuity design with discrete running variable", examines RD settings where the running variable is discrete, with values typically referred to as *mass points*. The chapter discusses the challenges associated with this particular case and the solutions proposed in the literature. It then introduces an original methodological proposal designed to handle situations in which the distribution of the outcome within each mass point is asymmetric—an issue that can compromise the accuracy of the estimates. To evaluate its effectiveness and compare it with traditional approaches, a simulation study is carried out, allowing the methodology to be tested in controlled settings and its performance to be assessed in terms of bias, variance, and mean squared error. Finally, because the INVALSI application involves a discrete running variable, the chapter concludes by extending the proposed method to the case study.

Chapter 5, titled "Handling the hierarchical structure of the data", introduces the Multilevel approach, whose purpose is to account for the nested (hierarchical) structure of the data. The chapter develops a methodological extension of RD to multilevel models, enabling the estimation of whether and how a treatment effect varies across different hierarchical levels. Three strategies for integrating RD with multilevel models are introduced, each defined by a different approach to selecting the bandwidth of observations around the cutoff. To assess the effectiveness of the proposed approaches, a simulation study is conducted, with the aim of identifying the most appropriate method under different empirical conditions. The chapter concludes with the application of the selected method to INVALSI data, exploiting the regional heterogeneity that characterises the Italian context.

Chapter 1

Introduction to causal inference

The goal of causal inference is to evaluate the causal effect of a treatment, where treatment refers to the implementation of a policy, a program, or, more generally, an intervention. The term "causality" has evolved considerably over the years. Until the 1990s, the explicit use of the term causality was relatively rare and often discouraged (Imbens, 2022). Causal inference could, for example, be related to evaluating the effect of a drug or medication on health; or, we could want to evaluate the effect of education on wages.

One of the greatest threats to causal inference is represented by the so-called *problem of confounders*, which Elster (1983) defined as a spurious relationship: a correlation between two variables that does not stem from a direct causal link between them, but from their common relation to some third variable. This situation makes the danger of confusing correlation with causation a constant problem in the statistical explanation of phenomena. In recent years, there have been significant methodological developments in causal inference from observational data (Pearl, 2000; Spirtes et al., 2000). In particular, Pearl (2000) provided a rigorous theoretical framework based on causal graphs and the *do-calculus*, which makes it possible to formalize the problem of confounders and to establish precise conditions for identifying causal relationships beyond mere statistical associations.

Thus, the first question to answer is how a causal effect can be defined, i.e. what we mean when we say that something causes something else. In our framework, causal effects will be defined using the term *"counterfactuals"*, the so-called *"what if?"* questions. As an example, if we refer to a medical treatment, the counterfactual would be: would the patient suffer in the absence of a certain treatment? If we had not administered a certain treatment, would the patient be in the same situation? Instead, if we refer to the effect of education on wages, our counterfactual would be: what would be the wage of graduates if they had never gone to university? Thus, the idea in causal inference is to compare the scenario we are observing with this *"counterfactual scenario"*. The problem is evident, and it arises from the fact that, by definition, the counterfactual is never observed, because it is clear that if we measure the graduate's wage, we could never know what their wage would be if

they had not graduated. For this reason, the main challenge in causal inference is to "construct" the counterfactual scenario, i.e., to deduce what would have happened in the opposite scenario (Chernozhukov et al., 2013). Let's try to construct the counterfactual scenario in a very naïve way. For example, let's say that if a patient takes a medicine for a headache and, after a while, it goes away, it means that the medicine has worked. Or we could say that graduates generally earn more, so education increases wages. All these naïve statements refer to a causal effect that is implicitly based on different types of counterfactuals. In the case of the patient with a headache, we are constructing the counterfactual based on the same unit (the same patient) but before the intervention. In the case of graduates, we are constructing the counterfactual based on another group of units that is subjected to a different intervention (the non-graduates). This is one way to deduce a causal effect.

A natural question arises: is this a reasonable approach to constructing our counterfactual? The answer is negative. In the case of the patient with a headache, we use the same unit at a different point in time. But, very often, the pain simply goes away because its effect fades over time (a phenomenon in statistics called *"mean reversion"*: if a variable deviates significantly from its mean, it is likely that, over time, it will return to its average or expected value). Perhaps the pain would have gone away even if the patient had not taken the medicine. This means that using the same patient to construct our counterfactual might not be optimal; we need many more assumptions if we want to do it, and these assumptions could be very strong. In the second example, we compare people who attend college with those who do not. However, individuals who do not attend college may abstain either because they are unable to or because they choose not to, and thus they differ in many respects from those who do (motivation factor, income preventing them from enrolling in college, different social networks, etc.). When we compare these two groups, we do not know whether we are actually measuring the causal effect of education or if we are confusing it with many other factors that differ between the two groups. Therefore, these "naïve comparisons" are invalid and represent a classic example of *"correlation fallacy"*: the mere fact that two things are associated (correlated) does not mean that one causes the other. If we want to talk about causal effects, we must be sure that the one and only difference between the two groups or units is the treatment and that these units are identical in every other aspect (a condition that is defined in economics by the Latin phrase "Ceteris paribus," which literally means "if everything else remains the same").

Now, after making these premises, let us introduce a model to define and analyze

causal effects. We will start with a very simple model and define the causal effect as *"Potential outcome"*. Suppose that there are n units, indexed by $i = 1, 2, ..., n$ from a certain population at a certain point in time t. At each time period t, unit i can be in one of two scenarios: treated, untreated. Implicitly, we are assuming that our treatment is binary (for simplicity). Now let's introduce the potential outcomes for this unit i:

- $Y_{it}(1)$: the potential outcome when i is treated
- $Y_{it}(0)$: the potential outcome when i is untreated

In this way, we can define the treatment effect for unit i at time t as:

$$\tau_{it} = Y_{it}(1) - Y_{it}(0) \tag{1.1}$$

Thus, the treatment effect is simply the difference between the two outcomes (for example, the wage of a unit if they attend college minus the wage of a unit if they do not attend college). Clearly, the effect of a policy or treatment can vary across units and over time, and this scenario is what we define as *"Heterogeneous Treatment Effect"*.

An important question then arises: what insights can the observed data provide about the phenomenon under investigation?

To start simply, consider only a binary treatment indicator:

$$D_{it} = \begin{cases} 1 & \text{if } i \text{ is treated at time } t \\ 0 & \text{if } i \text{ is untreated at time } t \end{cases} \tag{1.2}$$

And the observed outcome will be:

$$Y_{it} = \begin{cases} Y_{it}(1) & \text{if } D_{it} = 1 \\ Y_{it}(0) & \text{if } D_{it} = 0 \end{cases} \tag{1.3}$$

Thus, we will have:

$$Y_{it} = Y_{it}(1)D_{it} + Y_{it}(0)(1 - D_{it}) \tag{1.4}$$

Or, equivalently:

$$Y_{it} = Y_{it}(0) + \tau_{it}D_{it} \tag{1.5}$$

Now, we're interested in the treatment effect for each unit in the population, which is very useful to know but also very difficult. Obviously, knowing the treatment effect of each unit is very challenging and requires many strong assumptions. Instead,

we can focus on what we call the *"Average Treatment Effect (ATE)"*, that is, the average treatment effect for the entire population. We can define the ATE at time t as:

$$ATE_t = \mathbb{E}[\tau_{it}] = \mathbb{E}[Y_{it}(1) - Y_{it}(0)] \tag{1.6}$$

We are saying that the ATE is given by the average of all treatment effects for all units in the population, which, in turn, is the average of the difference between the potential outcomes of the treated and untreated units. Alternatively, we might want to focus our attention on the ATE for a specific group at time t:

$$ATE_{group,t} = \mathbb{E}[\tau_{it}|group] \tag{1.7}$$

If we could observe the entire population, the ATE for a specific group would be the average of the treatment effect but only for that specific group, rather than the treatment effect for the entire population. We can thus define what we call the *"Average Treatment Effect on the Treated (ATT)"* at time t as:

$$ATT_t = \mathbb{E}[\tau_{it}|D_{it} = 1] \tag{1.8}$$

which is the average treatment effect for units that are currently receiving the treatment. Another specific treatment effect for a group that we can obtain is what we call the *"Conditional Average Treatment Effect (CATE)"* at time t:

$$CATE_t(x) = \mathbb{E}[\tau_{it}|X_{it} = x] \tag{1.9}$$

This is, similarly, the average treatment effect for a specific group in a population (for example, the average treatment effect for men or for women).

Now, the key question is: what can the observed data tell us about these parameters? This is what is called in statistics or econometrics the *"identification problem"*, which has received considerable attention in econometrics (Pearl, 2010). Can we learn something about these parameters of interest based on the data we can observe? Unfortunately, at this point, we face a problem, which is something that the statistician Holland (1986) called *"The Fundamental Problem of Causal Inference"*. Holland says that it is impossible to observe the value of $Y_{it}(1)$ and $Y_{it}(0)$ for the same unit i at the same time t simultaneously and therefore it is impossible to observe the treatment effect on any unit i. This is because, for each unit, we can only observe $Y_{it}(1)$ or $Y_{it}(0)$, but never both. This means that there is always a *counterfactual scenario* that is unobserved: the *counterfactual* will be given by $Y_{it}(0)$ if the unit receives the treatment and by $Y_{it}(1)$ if the unit does not receive

the treatment.

In light of this, all the causal parameters we defined earlier are unobservable. Now, we need to understand how we can deduce the counterfactual to learn something about the treatment effect. As mentioned earlier, we will focus on the average effect, simplifying by ignoring the time dimension t. We stated that in this very simple scenario, our observed data consist of a treatment indicator D_i that tells us whether a unit receives the treatment or not, and an outcome Y_i. A natural approach would simply be to compare the averages of the outcome for the treatment and control units. For example, returning to the example of graduates and non-graduates, we simply compare the average income of students who graduated with those who did not. We can define this difference as:

$$\Delta = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \tag{1.10}$$

This difference seems to be the best we can achieve with the available data. The question is how to link this comparison of observed results with actual and potential results and treatment effects to see if this is a useful and suitable measure of the treatment effect or under what conditions it may be a useful and suitable measure:

$$\begin{aligned}
\Delta &= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \\
&= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\
&= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1] \\
&\quad + \mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\
&= \text{ATT} + \text{SB}
\end{aligned} \tag{1.11}$$

The parameter Δ is given by the difference between the average observed outcome for the treated minus the average observed outcome for the untreated. The first term, $E[Y_i|D_i = 1]$, is the observed outcome for the treated, which we can write as the potential outcome for the treated $E[Y_i(1)|D_i = 1]$; while the second term is the observed outcome for the untreated, $E[Y_i|D_i = 0]$, which we can write as the potential outcome for the untreated $E[Y_i(0)|D_i = 0]$. In the third line, we add and subtract the same term $E[Y_i(0)|D_i = 1]$, that is, the average outcome $Y_i(0)$ for the treated in case they hadn't received the treatment (which we cannot observe). The term $(E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1])$ is the average difference between $Y_i(1)$ and $Y_i(0)$ for the treated, representing the ATT:

- $ATT = (E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]) = E[\tau_i|D_i = 1]$, the difference in the outcome between receiving and not receiving the treatment for the units

that are currently receiving it.

The extra term represented by $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$ is the difference in $Y_i(0)$, i.e. the outcome for the untreated, between the people who are treated and those who are not. This term is generally not equal to zero because, for example, returning to the college case, we might think that people who go to college could have better family resources and could use these resources to find a better job and, as a result, might have a higher salary than those who didn't attend college. This is what we call ***Selection Bias (SB)***:

- $SB = (E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0])$.

This term captures the fact that there is selection in college enrollment, meaning that units are not randomly assigned to college but rather they choose to enroll themselves, which, as mentioned earlier, can depend on many factors such as motivation, social networks, family resources, making the units that attend college different from those who do not.

So, to summarize, we have this interesting decomposition of the average difference Δ in the ATT parameter (which we want to study) and the SB parameter, which is something we need to eliminate. However, this is something we cannot do with the data available because it is a source of bias. ATT and SB cannot be separated because the ATT alone cannot be measured, as it contains the counterfactual related to the potential outcome $Y_i(0)$ for the treated that we cannot observe. Similarly, in the other term related to SB, namely $(E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0])$, we cannot truly measure the SB based on our data, as it also contains the potential outcome $Y_i(0)$ for the treated. Thus, since we cannot measure these two terms separately, we take their sum. However, the sum is not very useful because it contains the bias that we want to eliminate. In fact, considering the sum, if the treatment effect were 10, for example, we wouldn't know the real contribution of each term to the value: ATT could be 10 and SB equal to 0, or ATT could be 100 and SB equal to -90, or even ATT equal to 0 and SB 10.

Since the inability to separate ATT from SB stems from the way treatment is assigned, the real challenge becomes understanding and, ideally, controlling the *'treatment assignment mechanism'*: what determines who receives the treatment and who does not.

1.1 Randomized Controlled Trials (RCTs)

The best way to determine who is treated and who is not is to control the assignment mechanism to the treatment. If we could decide who to assign to the treatment and

who not, we would solve the selection bias problem because, by construction, we would be sure there is no selection. This can be done by implementing what is called ***Randomized Controlled Trials (RCTs)***. The history of RCT dates back to approximately 600 BC with the first clinical trials, and then became a very popular and famous technique in the 19th and 20th centuries (Stolberg et al., 2004). In an RCT, researchers know the treatment assignment mechanism, and they choose it. It is a very popular mechanism in medicine and, recently, also in social sciences. In the simplest case of an RCT, we have a sample of interest and simply flip a coin to decide who will be treated and who will not (*Bernoulli trial*), so to avoid the selection issue. The advantage of flipping a coin is that it guarantees random assignment, and in this case, the potential outcome will be statistically independent of the treatment:

$$(Y_i(1), Y_i(0)) \perp D_i \tag{1.12}$$

In other terms, we mean that the treatment will not be associated with any of the characteristics of the units (for example, it will not be associated with whether a unit is male or female, or whether a unit has a high income or not). When we have random assignment, we can be sure that the treated and untreated units are comparable because, on average, they will be identical except for the fact that some will be treated and others will not. For random assignment, for $d = 0, 1$, we have:

$$E[Y_i|D_i = d] = E[Y_i(d)|D_i = d] = E[Y_i(d)] \tag{1.13}$$

It follows that:

$$SB = E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] = 0 \tag{1.14}$$

Thus, with random assignment, the selection bias is zero. Mathematically, this happens because, since the outcome and treatment are statistically independent, we can eliminate the terms $D_i = 1$ and $D_i = 0$. Intuitively, SB will be zero because we eliminate the selection bias by flipping a coin: we are sure that treated individuals will not be more motivated, will not have higher income, etc.

As a result, the simple comparison of the outcome of the treated with the outcome of the untreated will give us the average treatment effect:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1) - Y_i(0)] = \tau_{ATE} = \tau_{ATT} \tag{1.15}$$

In an RCT, the average treatment effect for the treated is equal to the average treatment effect for the entire population.

Therefore, it can certainly be stated that the main advantage of an RCT is due to the fact that the distortion from selection bias is eliminated, thus allowing the average treatment effect to be derived simply by comparing the outcome of the treated with the outcome of the untreated. Therefore, the estimate of the average treatment effect will be very simple. Let us take a closer look at how estimation works. Suppose we have a random sample consisting of n independent and identically distributed observations and we randomly allocate the treatment by flipping a coin:

$$(Y_i, D_i)_{i=1}^n \tag{1.16}$$

As mentioned above, we can estimate the average ATE treatment effect as a difference on average:

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0 = \frac{\sum_{i=1}^n Y_i D_i}{n_1} - \frac{\sum_{i=1}^n Y_i(1 - D_i)}{n_0} \tag{1.17}$$

where:

$$n_1 = \sum_{i=1}^{n} D_i; n_0 = \sum_{i=1}^{n}(1 - D_i) \tag{1.18}$$

This estimator will enjoy several properties:

- **Unbiased and consistent**: First, the variance of the estimator $\hat{\tau}$ will have a very simple form given by the sum of the variance of the treated units divided by the number of treated units, and the variance of the untreated units divided by the number of untreated units:

$$Var[\hat{\tau} \mid D_1, D_2, \ldots, D_n] = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}, \quad \sigma_d^2 = Var[Y_i(d)] \tag{1.19}$$

- **Asymptotically normal**:

$$\frac{\tau - \hat{\tau}_{ATE}}{se(\hat{\tau})} \to \mathcal{N}(0, 1) \tag{1.20}$$

Thanks to these properties, it is possible to construct the usual inferential tools to assess the statistical significance of the estimated treatment effect:

- **Test hypothesis**: for $H_0 : \tau_{ATE} = \tau_0$ vs $H_A : \tau_{ATE} \neq \tau_0$, reject H_0 if

$$|\mathrm{T}| = \left| \frac{\hat{\tau} - \tau_0}{se(\hat{\tau})} \right| \geq z_{1-\frac{\alpha}{2}} \tag{1.21}$$

- **Confidence Interval**:

$$CI_{1-\alpha}(\tau_0) = [\hat{\tau} - z_{1-\alpha/2}se(\hat{\tau}), \hat{\tau} + z_{1-\alpha/2}se(\hat{\tau})] \tag{1.22}$$

In an RCT, we can estimate the treatment effect not only by simply taking the difference between the means but also by implementing an OLS linear regression:

$$Y_i = \alpha + \tau D_i + \varepsilon \tag{1.23}$$

Where the coefficient τ will give us an estimate of the ATE. Clearly, choosing either method makes no difference. However, estimating the ATE using the OLS method may have some advantages such as:

- Ability to control for covariates
- Estimate heterogeneous treatment effects
- Estimate multiple treatment effects
- Take into account the possible correlation between observations (clustering).

Alternatively, to conduct inference, we could turn to *Fisher's approach*. This is an alternative procedure for conducting inference in an experiment. The difference between this approach and the classical one shown earlier is that, while classical inference works by exploiting the central limit theorem and therefore requires a sufficiently large sample size, this approach can be used even when the sample size is small.

1.2 Threats to the validity of experiments

In order to fully exploit the potential of experiments, it is necessary to conduct a study on their validity. In this regard, it is important to make a distinction between:

- *Internal validity*: an experiment is internally valid if the statistical inference made about causal effects is valid for the population under analysis.
- *External validity*: an experiment is externally valid if the results and inferences can be generalized to other populations.

1.2.1 Internal validity

There are several causes that make an experiment internally invalid:

- **Failure of randomization**: A threat to the internal validity of experiments is the lack of proper randomization. When assignment to treatment is not random but influenced by the characteristics or preferences of participants, the results of the experiment may reflect both the effect of the treatment and the effect of the non-random assignment. For example, in a study on a vocational training program, if participants are assigned to the treatment group based on the first letter of their last name (first or second half of the alphabet), systematic differences between groups could be observed. Due to ethnic disparities in last names, the racial composition of the groups could vary. Since race may be correlated with factors such as work experience, education level, and other labor market characteristics, these omitted variables could influence the results, creating systematic differences between the groups. In these cases, non-random assignment generates a correlation between the treatment and the error term.

- **Violation of the experimental protocol**: This is the case when units voluntarily choose to refuse the treatment despite being assigned to the treatment, or conversely, receive the treatment even though they were not assigned to it. In this case, due to the element of choice, a correlation between treatment and error would be generated, leading to biased estimates. However, in this case, a non-biased estimate of the causal effect could still be obtained by using instrumental variable estimation of the treatment effect if information on both the treatment actually received and the initial assignment is available.

- **Attrition**: This condition occurs when one or more units, after being assigned to a treatment or control group, drop out of the experiment.

- **Experimental effects**: This threat to internal validity arises because some units, knowing they are part of an experiment, may change their behavior (*Hawthorne Effect*).

- **Small sample sizes**: Another source of threat is related to small sample sizes, which can cause problems in conducting statistical inference.

1.2.2 External validity

The causes of threats to the external validity of experiments must also be considered. Among these, we have:

- **Unrepresentative sample**: This occurs when the studied population and the population of interest are not sufficiently similar to generalize the results.
- **Unrepresentative program**: Just as with the population, the similarity issue also applies to the program or policy studied and the one of interest.
- **General equilibrium effects**: This threat, as defined by economists, may arise because the actual program is permanent, while the experimental program has the main characteristic of temporariness. This transition between two such different scenarios could cause the economic environment to change in a way that makes the results non-generalizable.

1.3 Regression Discontinuity design

As previously mentioned, the most effective way to estimate causal effects is to conduct an RCT. However, in practice, it is very difficult to create the conditions to carry out it, both for ethical reasons and the high computational costs. Sometimes, in reality, due to external events, the treatment of some individuals happens *"as if it were"* random. This condition of near-randomness produces what is called a *"quasi-experiment"* (or *"natural experiment"*). These external variations that lead to the birth of a quasi-experiment can arise, for example, from the implementation of policies of various types, changes in institutions, or geographical factors. One of the most credible methods for analyzing causal effects in these conditions is *"Regression Discontinuity (RD) design"*. RD was introduced by Thistlethwaite and Campbell in 1960 (Thistlethwaite and Campbell, 1960) as a method to estimate treatment effects in non-experimental contexts, where treatment assignment depends on exceeding a known threshold (called *cutoff* or *threshold*) of an observed variable (called *forcing variable*, *running variable*, or *score*). To better understand RD, consider the following example: suppose that a school introduces a test to assign financial support (scholarships) to all students whose score exceeds a certain threshold c on the entrance test. The goal is to estimate the causal effect of the scholarships (treatment, D_i) on academic outcomes (outcome, Y_i). The outcome could, for instance, be represented by the GPA—a grading system commonly used in schools and universities, particularly in the United States, to summarize a student's overall academic

performance. Thus, the assignment mechanism for the treatment is as follows:

- Students take an exam for which they will receive a score X_i
- Scholarships will be assigned to all those who score greater than or equal to the threshold $X \geq c$

Unlike an RCT, this time it is not the researcher who decides who will receive the treatment and who will not, but there is a very clear rule that determines this ($X_i \geq c$) (Angrist and Pischke, 2009). The goal now is to capture the causal effect of the scholarships. One might think of using a naïve approach, simply comparing those who receive the scholarship with those who do not. However, this approach would lead to a series of confounders because, by construction, students who receive the scholarship have better academic performance, and these results can be correlated with many different factors such as IQ, study hours, motivation, etc. Therefore, if we simply compare students with and without a scholarship, we obtain:

$$\Delta = ATT + SB \tag{1.24}$$

Where SB represents the selection bias, which in this case is caused by the confounders mentioned earlier (IQ, study hours, motivation).

In general, many of these confounders are unobservable, which is a significant problem. RD takes these factors into account, both the confounders and the fact that, by construction, students will be different. Suppose that:

- Students do not have exact control over their score. Certainly, students can have some control because they can decide how many hours to dedicate to studying or, for example, their motivation. However, this control cannot be exact (e.g., if the cutoff were 0.5, students could not decide to get exactly 0.51).
- The confounders will not have a *"jump"* at the cutoff.

These are assumptions that, in many cases, are more than reasonable. Through RD, we will compare students just above the cutoff with those just below it. The basic idea is that we cannot compare all students above the cutoff with all students below the cutoff because these will differ from each other in many ways. Instead, the closer we get to the cutoff, the more we can compare students who are similar in many aspects: they probably studied the same amount of hours, have similar IQs, or have very similar motivation. However, while some were perhaps just lucky to end up just above the cutoff, others were not as lucky and ended up below. For example,

two students might have answered a question entirely by chance, but while the first one was lucky enough to guess it and get a score of 100.1, the second was unlucky and got a score of 99.9. Thus, the closer we get to the cutoff, the more comparable the students will be, and this is the fundamental idea behind RD.

From this point onward, the following notation will be used:

- **Potential outcome**: $(Y_i(1), Y_i(0))$, with $\tau_i = Y_i(1) - Y_i(0)$
- **Running variable**: X_i. For the moment, it is assumed that this variable is continuous (there is also the case of a discrete running variable, which will be discussed later, and which can generate several problems in the analysis).
- **Treatment indicator**: $D_i = D_i(X_i) = 1$ if treated, 0 otherwise. There will be a deterministic rule such that if the value of the running variable is above the cutoff, the unit will receive the treatment, otherwise, it will not.
- **Observed outcome**: $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$

One of the main advantages of RD is that it is very easy to represent graphically and is very intuitive. In fact, sometimes it is possible to spot the presence of discontinuities simply by plotting a scatterplot between the outcome and the running variable.

In Figure 1.1, we represent a typical RD graph. On the X axis, we have the running variable, which, returning to the previous example, could be the entrance test score, and on the Y axis, we have the outcome, namely the GPA. Two different curves are shown:

– The black curve represents the average outcome without the scholarship as a function of the score. What would the GPA be if the entrance test score were

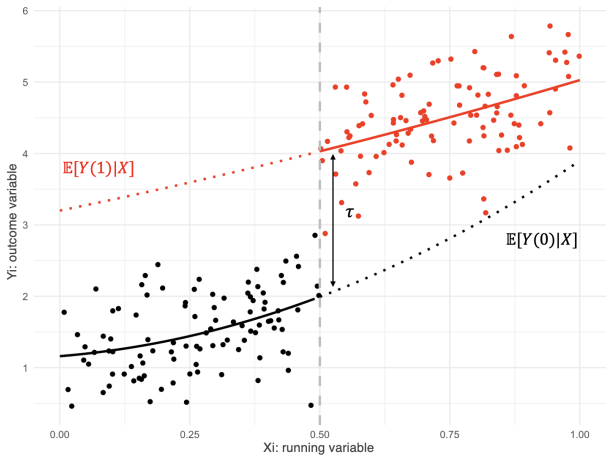


Figure 1.1: Treatment effect in RD

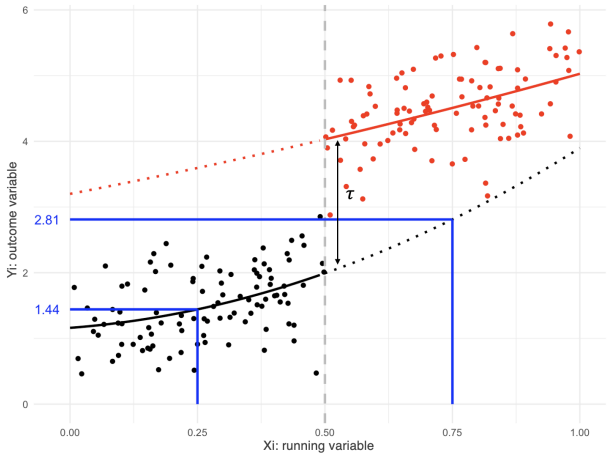


Figure 1.2: The value of the Outcome if the Running Variable did not exist

0.25 without the scholarship? Moving for a while at the Figure 1.2, the Y for the unit that scored 0.1 is about 0.58. What would the GPA have been if the entrance test score were 0.75 without the scholarship? By projecting it (blue line), the answer would be about 2.34. This is how it would work if the scholarship did not exist. Since the scholarship exists, we can also plot the red curve.

– The red curve: represents the outcome when the scholarship is awarded. Since the red curve is above the blue one, it means that the scholarship has a positive effect on the GPA.

The treatment effect is represented by the vertical distance between these two curves, i.e., the parameter τ. However, there is a problem because the two curves black and red will be observed only for those who do not receive the treatment and for those who do (the contiuous part of the red and black curves). This means that the dashed black curve, as well as the dashed red curve, are not observed, as they represent the outcome of the units if the treatment had not been assigned and the outcome of the units if the treatment had been assigned to everyone. In summary, **we will never observe $Y(1)$ and $Y(0)$ at the same time.**

A very naïve approach would be to simply calculate the average outcome for those who receive the treatment and compare it with the average outcome for the units that do not receive the treatment (the two horizontal dashed blue lines in Figure 1.3). However, in this way, we would obtain an estimate that is quite far from the true treatment effect, because we confuse the effect of the scholarship with confounders. Furthermore, using this naïve approach, we are comparing units whose score is very high in the entrance test (for example, a score of 0.9 or 1.0) with units whose score is very low (for example, a score of 0.0 or 0.1). Obviously, these two

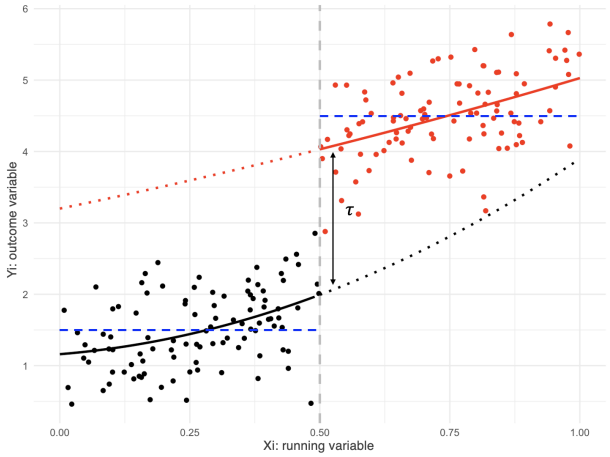


Figure 1.3: Treatment effect in RD: naïve approach

groups are not comparable. Since these units are so different from each other, one might consider excluding them from the analysis and only taking those units whose score is between 0.25 and 0.75 (Figure 1.4). From Figure 1.5, we can see how we can approach the estimate of the true treatment effect τ, reducing the bias. To get closer to the true treatment effect, one might consider continuing to exclude units that are still very different from each other. This leads to focus the comparison only on those units that are very close to the cutoff (Figure 1.5). In this way, we get as close as possible to the true parameter τ. Therefore, the idea behind RD is to localize the analysis around the cutoff, looking only at a small window around it and comparing only those units that are just above the cutoff with those just below it. Under some assumptions that will be discussed later, this process will eliminate the selection bias.

In summary, only units with scores of 0.49 and 0.51 (for example) will be selected,

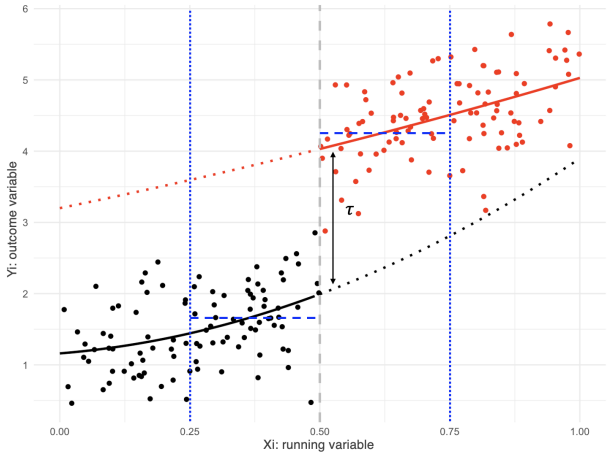


Figure 1.4: Treatment effect in RD taking the window between 0.25 and 0.75

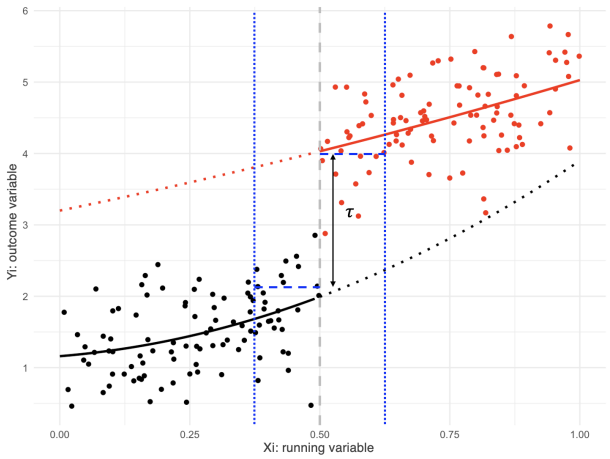


Figure 1.5: Treatment effect in RD taking the window between 0.375 and 0.625

because these units have probably dedicated the same amount of study hours, will be similar in terms of IQ or motivation, but, perhaps, due to a purely lucky factor related to a single wrong or correct answer, one ended up in the treatment group and the other in the control group. By getting very close to the cutoff, we will have something very similar to a randomized experiment. Thus, with the use of RD, it is possible to say something about the treatment effect only at the cutoff because it is the only point where we will have overlap, the only point where we will have treated and untreated units. In Figure 1.4, it would not be possible to measure the treatment effect if the cutoff were, for example, 0.25 or 0.75, because at these points there are only untreated or treated units, respectively. This means that the treatment effect cannot be studied at other points in the score distribution: the treatment effect is identified in a non-parametric way only at the cutoff. By identification, we mean that by observing the data, we can say something about the treatment effect; by non-parametric we mean that no functional form or assumptions about the model are made. In fact, one of the advantages of RD is that no assumptions need to be made about how the functions (the red and black ones) behave, whether they are linear, quadratic, or cubic: the choice is guided by the data. Thus, this is a huge advantage, as non-parametric methods help avoid the so-called *"misspecification bias"*, which occurs, for example, when assuming that a function is linear when it is actually not, leading to incorrect estimates. For the estimation, we need to understand how to calculate the two points whose distance represents the treatment effect (the two points at the intersection of the red curve and the black curve with the dashed vertical line representing the cutoff). Thus, in order to obtain the so-called ***"Nonparametric identification"*** (Hahn et al., 2001), we will need very mild assumptions:

1. Sharp design: $D_i = 1(X_i \geq c)$

2. Smoothness: $\mathbb{E}[Y_i(0)|X_i = x]$, $\mathbb{E}[Y_i(0)|X_i = x]$ continuous at $x = c$

The first assumption involves a *Sharp* design, where crossing the cutoff means necessarily receiving the treatment, and not crossing it means not receiving it (the differences between the two most famous RD design, the *Sharp* and the *Fuzzy* ones, will be discussed in Chapter 2). With the second assumption, which cannot be verified in practice, it is intended that the two curves (the black and red ones) are continuous at the cutoff, showing no jump. The presence of a jump in either of these two functions, could related to another type of treatment different from the one of interest. And, if there is another treatment far from the cutoff, we would not be able to distinguish between the two treatments. The bad news is that this is something that cannot be controlled within the data, since we will only observe the solid functions in figures 1.1, 1.2, 1.3, 1.4 and 1.5 previously shown: we will not observe, of course, the treatment function if the unit had not received the treatment or the non-treated function if it had received the treatment. Once this assumption is made, we will have:

$$\mathbb{E}[\tau_i|X_i = c] = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] \tag{1.25}$$

Thus, under these two assumptions, it is very simple to identify the average treatment effect at the cutoff.

1.3.1 The evolution of Regression Discontinuity: from origins to widespread adoption

Regression Discontinuity Design (RDD) was first introduced by Thistlethwaite and Campbell in 1960 (Thistlethwaite and Campbell, 1960) as a method to estimate treatment effects in non-experimental contexts, where treatment assignment depends on exceeding a known threshold (referred to as the *cutoff* or *threshold*) of an observed variable (also known as the *forcing variable*, *running variable*, or *score* in the literature). In their initial application of the method, Thistlethwaite and Campbell analysed the impact of merit-based awards on subsequent academic outcomes, leveraging the fact that the awards were assigned based on a test score. The core idea of the research design was that individuals with scores just below the threshold (who did not receive the award) represented suitable comparisons to those with scores just above the threshold (who did receive the award). The discontinuity between the two groups at the threshold provided a reliable estimate of the average treatment effect.

Following the work of Thistlethwaite and Campbell, Van der Klaauw (2002) and Angrist and Lavy (1999) also applied the RD design in the education sector to estimate the discontinuities generated by, respectively, a financial aid programme and a policy targeting class size reduction on educational outcomes. By employing RD, Van der Klaauw estimated how financial aid increased the probability of university enrollment for students with financial need. Angrist and Lavy, on the other hand, studied the treatment effect of the "Maimonides' rule", a policy implemented in Israeli public schools stating that classes could not be composed of more than 40 students. Authors showed that reducing class sizes led to significant and substantial improvements in test scores for third and fourth-grade students, but not for those in third secondary grade.

Despite being introduced over fifty years ago, this evaluation strategy has garnered significant attention from economists only relatively recently, becoming one of the most widely used non-experimental techniques. One reason for this rise in popularity lies in the relatively mild assumptions required by RD compared to other non-experimental methods (Hahn et al., 2001; Lee and Lemieux, 2010). Numerous studies in the literature clarified the methodological aspects of RD, providing practical guides for its application across diverse analytical contexts (see, e.g., Imbens and Lemieux (2008); Cattaneo et al. (2020); Skovron and Titiunik (2015); Hahn et al. (2001); Cattaneo et al. (2024); R. Jacob et al. (2012); Cattaneo and Titiunik (2022)). In many cases, the recent popularity of RD designs appears to be well-justified. Black et al. (2007) and Buddelmeyer and Skoufias (2004) compared RD with randomised experiments, showing that results obtained from the two methods were very similar (see Cook and Wong (2008), for a synthesis of studies comparing RD and experiments).

Cook and Wong (2008) conducted a review tracing the history of RD, describing its origins, the lack of interest shown by statisticians, its limited acceptance in economics from 1972 to approximately 1995, and its subsequent revival and increasingly common usage. Lee and Lemieux (2010) aimed to clarify the growing body of RD-based studies, identifying 77 applications in economics, categorised by sector. These applications span education, labour markets, economic policy, health, crime, environment, and other fields. Obviously, since 2010, the number of studies employing RD has clearly increased further.

1.3.2 Extensions and applications of Regression Discontinuity designs

Building on its historical development and methodological consolidation, the RD framework has been applied in a wide range of contexts and has also been extended into several variants. These applications illustrate both the flexibility of the design and its ability to address diverse research questions beyond the original educational setting. In what follows, we review some of the most influential studies and introduce the main extensions of RD—such as Sharp, Fuzzy, Geographic, and Time-based designs—highlighting their distinctive features and empirical relevance.

An innovative study by Meyersson (2014) examined the impact of Islamic political governance on women's empowerment using a *Sharp RD* design in Turkish municipalities. In a Sharp RD, the probability of receiving treatment is equal to 1 if the value of the running variable exceeds the cutoff, 0 otherwise. Thus, there is an abrupt jump in the probability of treatment at the cutoff from 0 to 1. Analyzing the 1994 municipal elections, in which an Islamic party narrowly won in some areas, the study compared municipalities where the party barely won or lost, using the party's margin of victory as the running variable to generate treatment and control groups. Despite a crude negative correlation between Islamic governance and women's rights, RD results showed a significant increase in secondary education among young women in the six years following the election. This finding contradicted the correlation analysis, which, unlike RD, did not focus on units just above or below the threshold, but instead analyzed the entire sample.

Remaining within the realm of Sharp RD, another notable study was conducted by Ludwig and Miller (2007). This research analyzed the effects of the "Head Start" program on health and education, exploiting a funding discontinuity in 1965, when the Office of Economic Opportunity (OEO) provided technical assistance to the 300 poorest districts in the United States. Access to treatment was reserved only to the 300 poorest districts. The results show a significant reduction in child mortality due to program implementation and an increase in educational attainment. Despite some limitations in the available data, the results suggest that Head Start has produced benefits in excess of costs in the most disadvantaged districts, reducing educational inequalities and improving opportunities for poor children.

A design distinct from Sharp RD is the *Fuzzy RD*. Unlike the Sharp design, where the probability of receiving treatment jumps from 0 to 1 at the cutoff, in a Fuzzy RD, the probability of treatment increases when the running variable exceeds the cutoff but does not jump abruptly from 0 to 1. A recent study by Londoño-Vélez

et al. (2020) examined the impact of financial aid on higher education enrollment, university choice, and student composition in Colombia using a large-scale program targeting high-achieving, low-income students. The results of the RD design indicated that the eligibility for financial aid increased the immediate enrollment rates between 56.5% and 86.5%, narrowing the socioeconomic gap in enrollment among the best performing students.

Given its flexibility, RD criteria can also be applied in geographic frameworks, where the running variable is represented by a boundary (geographic or administrative) dividing regions or territories into treatment and control groups. This approach, known as *Geographic Regression Discontinuity (GRD)*, is increasingly popular in political science and has been applied to various topics. One of the first studies using geographic discontinuity was conducted by Card and Krueger (1994). This study analysed the effects of a 1992 policy raising the minimum wage in New Jersey by comparing employment in 410 fast-food restaurants in New Jersey and Pennsylvania, where the minimum wage remained unchanged. Contrary to predictions from traditional economic models, the study found that the increase in the minimum wage did not reduce employment in New Jersey fast-food establishments. Instead, the results suggested a relative increase in employment compared to Pennsylvania. Prices in New Jersey fast-food restaurants increased relative to those in Pennsylvania, indicating that the cost of the minimum wage increase was partially passed on to consumers.

Keele and Titiunik (2015) take up the studies conducted by Huber and Arceneaux (2007) and (Krasno and Green, 2008) to provide an excellent guide for the use of the geographic discontinuities. These authors exploited differences in the distribution of presidential campaign television advertisements within the same state to study their effects on voter turnout and political attitudes. The authors compared voters residing in adjacent counties belonging to different media markets (Designated Market Areas, or DMAs), where some markets received a large number of advertisements while others received few or none. Both studies focused on the 2000 presidential election, hypothesising that voters near the border between two DMAs were effectively randomised in their exposure to campaign ads. This quasi-experimental situation arose because, due to their proximity to Philadelphia, residents of eight counties in southern New Jersey were exposed to 2,247 presidential advertisements during the final three weeks of the 2000 campaign. In contrast, voters in 12 counties in northern New Jersey, near New York City, received only 16 advertisements. This stark difference between counties was due to the designation of southern New Jersey as part of the Philadelphia DMA, whereas northern New Jersey was part of

the New York DMA. In both studies, the authors found no significant evidence that exposure to presidential campaign advertisements during the 2000 election increased voter turnout.

Another framework in which RD is applied is *Regression Discontinuity in Time (RDiT)*, where the running variable is time, and the cutoff corresponds to a specific date. Most applications of RDiT focus on estimating the treatment effects of environmental or energy policies (see, e.g., Anderson (2014), Auffhammer and Kellogg (2011), Burger et al. (2014), Chen et al. (2009), Grainger and Costello (2014), Lang and Siler (2013)). According to Hausman and Rapson (2018), the proliferation of this technique in energy and environmental economics is tied to the availability of high-frequency pollution data. Anderson (2014) utilised a natural experiment in Los Angeles in 2003, when a 35-day strike halted the city's public transport system. The author leveraged this abrupt disruption to estimate the effects of public transport services on traffic congestion, assuming that most public transport users commute along routes most affected by delays. By limiting observations to a specific window around the cutoff (known as *bandwidth*), set at 28 days on either side of the strike's start date (14 October 2003, with the bandwidth spanning 16 September to 10 November 2003), the results showed a 47% increase in delays when public transport services ceased.

When analysing the effectiveness of a programme, policy makers may be interested in more than just the overall average effect. For example, there may be situations where the average net effect is negative, but the programme is still adopted because of its distributional implications. Consider a vocational training programme: if the programme increases wages at the upper end of the wage distribution, it is likely to be rejected by policymakers. Conversely, if the programme raises wages at the lower end, it may be approved. Traditional methods that account for treatment heterogeneity but focus solely on estimating the average treatment effect are insufficient for distinguishing between programmes with these differing distributive characteristics. For these reasons, another framework explored within RD is the quantile approach, through the so-called *Quantile Regression Discontinuity (QRD)*. A significant contribution in the literature is the work of Frandsen et al. (2012), who introduced a nonparametric estimator for the local quantile treatment effect within the regression discontinuity design using data from B. A. Jacob and Lefgren (2004) for the application. These authors examined a policy implemented in 1996 in Chicago public schools, which required students to pass mathematics and reading tests in June. Students who failed to meet the standards in either test were mandated to attend a six-week summer school programme. The goal was to exploit

this discontinuity to identify the causal effect of summer school attendance on test performance one and two years later. Results showed the heterogeneity of treatment effects: while the impact was negligible at the lower end of the distribution, there were significantly positive effects at the upper end. The authors interpreted unobserved heterogeneity as a measure of motivation, suggesting that motivated students benefited from the additional summer instruction, whereas unmotivated students gained nothing, arguing that students who dislike school do not improve when required to attend additional hours.

In the context of RD with heterogeneous effects, Becker et al. (2013) analysed the distribution of Objective 1 Structural Funds from the European Commission to regions in EU member states below a certain income threshold. Employing an RD design with systematically heterogeneous treatment effects, the study examined how the treatment effect varied according to the quality of governance and the level of human capital in different regions. Using regional per-capita income levels prior to a programming period as the running variable and interacting Objective 1 treatment with institutional quality and education levels, the authors found that only regions with sufficient human capital and good institutions were able to translate programme transfers into faster growth in per-capita income and investments.

Chapter 2

Estimation procedure in Regression Discontinuity design

The estimation procedure in RD requires to estimate two different regression functions, one to the right and one to the left of the cutoff. There are different approaches that can be exploited to thus end. The first (which is the method used in the 1990s and early 2000s) is the so-called ***"Global parametric approach"***, which consists of a global estimation, as the name suggests. This approach involves choosing a degree p for the polynomial and then estimating a polynomial to the right and one to the left and comparing the value of the intercept. It is a very simple and intuitive approach but it comes with some problems. The first issue is the choice of the polynomial: we are dealing with a parametric estimation problem where a bias will be introduced if the degree p of the polynomial is wrong. For example, one might think of choosing a one-degree polynomial, but if the observations do not behave linearly because the true model is actually cubic, it will simply produce bias. On the other hand, one might try to avoid bias by choosing a very high degree for the polynomial, making it more flexible to the data, but once again, this could introduce bias due to overfitting, a problem known in statistics as *"Runge's phenomenon"*. It is known that, due to this phenomenon, higher-degree polynomials behave poorly at the boundary points, which is a major problem since we are interested in estimating precisely at those points. A second approach is the so-called ***"Flexible parametric approach"***: it involves trying to estimate a polynomial within an ad hoc bandwidth. This approach is much more in line with the basic idea of RD, where we can say something about the causal effect only at the cutoff. However, the problem with this approach is that the bandwidth is chosen ad hoc, which creates a problem due to arbitrariness in its selection. Therefore, we need to find a data-driven approach to determine how close we can get to the cutoff. For this reason, the third approach, which is the most recommended, is the so-called ***"Local polynomial approach"***. The advantage of this approach is that it is completely non-parametric. Therefore, no assumptions need to be made about whether the true data are linear, quadratic, or cubic. Another advantage is that it offers a data-

driven method for selecting the bandwidth. In fact, the bandwidth will be chosen based on the nature of the available data, which eliminates the arbitrariness that characterizes the other approaches.

2.1 Global parametric approach

As mentioned earlier, using the global parametric approach to estimate the average treatment effect means considering all the observations to the left and to the right of the cutoff, regardless of their proximity to it. Thus, it involves making a parametric assumption about the two global regressions to estimate. For example, using the notation proposed in the Chapters, let us assume that these functions are linear:

$$\mathbb{E}[Y_i(d)|X_i] = \alpha_d + \beta_d(X_i - c)$$

If these two functions were linear, it would mean that the treatment effect could be estimated by implementing a simple regression of the outcome (Y_i) using an intercept (α_0), the indicator for treatment (D_i), the running variable ($X_i - c$), and an interaction between the two ($(X_i - c)D_i$):

$$\mathbb{E}[Y_i|X_i] = \alpha_0 + (\alpha_1 - \alpha_0)D_i + \beta_0(X_i - c) + (\beta_1 - \beta_0)(X_i - c)D_i + u_i \qquad (2.1)$$

where the coefficient $\alpha_0 - \alpha_1$ will represent the treatment effect. One thing to note is that when including the running variable, it will be denoted as $X_i - c$, so that it is centered in such a way that the new cutoff becomes 0. Thus, the study is reduced to an RD (Regression Discontinuity) design where the cutoff is 0.

As mentioned earlier, an important choice when implementing the RD is the order of the polynomial p. If the simplest polynomial, i.e., of order 0, is chosen, as shown in Figure 1.3 in Chapter 1, it would be like estimating a flat line. This means comparing the average of the outcome above and below the cutoff. Clearly, this will not provide a reliable estimate of the treatment effect. For this reason, one might consider using a polynomial of order 1. As shown in Figure 2.1, this could be an excellent estimate of the average treatment effect, since the distance between the two dashed lines is very similar to the true unknown parameter τ. Alternatively, one might consider fitting an even more flexible polynomial, such as a quadratic polynomial with $p = 2$, as shown in Figure 2.2. By comparing Figures 2.1 and 2.2, we can observe that a quadratic polynomial provides a worse estimate: by trying to give more flexibility to the polynomial, we only move further away from the actual value of the parameter τ.

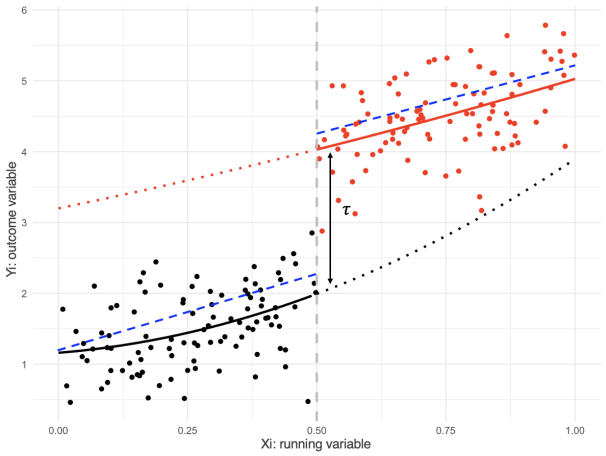


Figure 2.1: Global parametric approach with $p = 1$

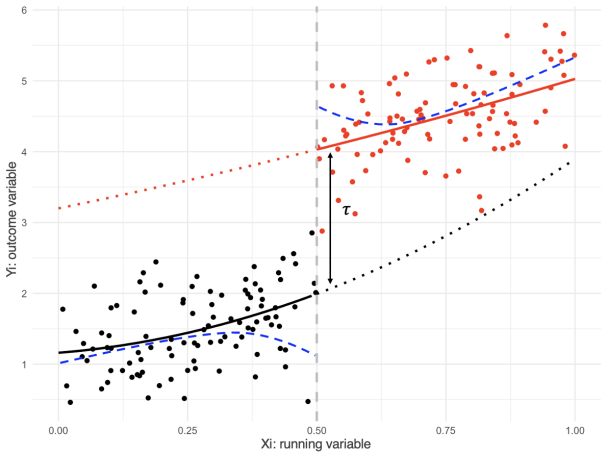


Figure 2.2: Global parametric approach with $p = 2$

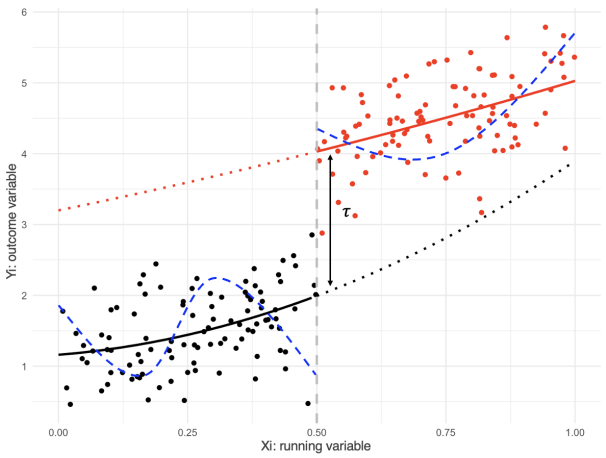


Figure 2.3: Global parametric approach with $p = 3$

The more flexible the polynomial, the more likely it is that the treatment effect will be overestimated. Indeed, by continuing to increase the order of p until a cubic polynomial is reached, as shown in Figure 2.3, we can appreciate how the situation worsens. Since we can never know the true model, the Global parametric approach seems not to be the most suitable method. All this can be explained again by the problem highlighted earlier, known as Runge's phenomenon: observations with scores of 0.0 or 0.1 will influence the estimator at the boundary points, affecting the shape of the polynomials to the right and left of the cutoff. For this reason, an alternative approach do not use all the available data, but only those within a certain window around the cutoff point (the bandwidth).

2.2 Local polynomial approach

The Local polynomial approach introduces another important choice in the context of RD, namely the choice of the bandwidth to consider. Once the bandwidth around the cutoff is chosen, observations outside this bandwidth will automatically be excluded. This result can be achieved through the use of the classic Ordinary Least Squares (OLS) approach. To do this, suppose we have a cutoff $c = 0$. Choose a bandwidth $h > 0$ and a kernel function $K(\cdot)$ that assigns a certain weight to the observations:

$$(\hat{\alpha}^+, \hat{\beta}^+) = \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 K\left(\frac{X_i}{h}\right) \mathbb{1}(X_i \geq 0) \tag{2.2}$$

$$(\hat{\alpha}^-, \hat{\beta}^-) = \underset{(\alpha,\beta)}{\arg\min} \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2 K\left(\frac{X_i}{h}\right) \mathbb{1}(X_i < 0) \tag{2.3}$$

The treatment effect at the cutoff will be given by:

$$\hat{\tau} = \hat{\alpha}^+ - \hat{\alpha}^- \tag{2.4}$$

where the superscripts $+$ and $-$ indicate, respectively, the treatment group (to the right of c) and the control group (to the left of c). The only difference with OLS is that the procedure is performed separately for the observations above ($X_i \geq 0$) and below the cutoff ($X_i < 0$), which in this case is assumed to be 0 for simplicity. The key term in the two equations 2.2 and 2.3 is the term $K(X_i/h)$, which is the weight assigned to the observations depending on their distance from the cutoff. Therefore, a choice is introduced in the implementation of RD, namely the choice of the *Kernel function*, which essentially represents the weight to be assigned to each observation. Generally, a weighting function is chosen that assigns a higher weight the closer the

observations are to the cutoff, as they are the most influential for the comparison.

In summary, the steps to follow with this approach, which is the most popular, are:

1. Choose a polynomial order p and a kernel function $K(\cdot)$.
2. Choose a bandwidth h.
3. Assign a weight of 0 to observations outside of h and a certain weight to observations within h.
4. Implement a weighted linear OLS regression separately for the treatment group and the control group.
5. Calculate the difference between the intercepts of the two linear regressions, which will represent the treatment effect.

2.2.1 The Choice of the Bandwidth

A crucial topic concerns the choice of the bandwidth: how many observations should be involved in the analysis? The goal is to estimate two different regressions, one to the right and one to the left of the cutoff, the difference between the intercepts providing the treatment effect measure. Thus, from the theory about OLS, we will have:

$$y = \alpha + \beta x + \varepsilon \tag{2.5}$$

$$\hat{\beta} = \frac{\hat{\text{cov}}(x, y)}{\hat{V}(x)} \tag{2.6}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{2.7}$$

Translated into the RD framework where weights, bandwidth, and cutoff are added, becomes:

$$\hat{\beta}^{+} = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X}h)K_h^{+}(X_i)}{\sum i = 1^n X_i(X_i - \bar{X}_h)K_h^{+}(X_i)} \tag{2.8}$$

$$\hat{\alpha}^{+} = \bar{Y}_h - \bar{X}_h\hat{\beta}^{+} \tag{2.9}$$

The coefficient α is the one of most interest. Now, we need to understand the functioning of these two estimators, $\hat{\beta}^{+}$ and $\hat{\alpha}^{+}$. To this end, we turn to the non-parametric approach. The parametric approach assumes a given form for the true model, e.g. linear, and then select the corresponding estimator without exploiting

the data to check the validity of the assumptions. In the non-parametric approach, on the other hand, these linear regressions are viewed as an approximation of the true model, acknowledging that it might not be correct. For this reason, in the non-parametric approach, the goal is to understand the form of the bias, its size, so that we can try to make it as small as possible and model it. The expected value of the two estimators, $\hat{\beta}^+$ and $\hat{\alpha}^+$, is:

$$\mathbb{E}[\hat{\beta}^+ \mid \mathbf{X}] = \dot{\mu}_1(0) + o_{\mathbb{P}}(h) \tag{2.10}$$

$$\mathbb{E}[\hat{\alpha}^+ \mid \mathbf{X}] = \mu_1(0) + h^2\mathcal{B} + + o_{\mathbb{P}}(h^2) \tag{2.11}$$

Where $\mu_1(x)$ is given by:

$$\mu_1(x) = \mathbb{E}[Y_i(1)|X_i = x], \quad \dot{\mu}_1(x) = d\mathbb{E}[Y_i(1)|X_i = x]/dx \tag{2.12}$$

Such term represents the expected value of the potential outcome as a function of x, which graphically represents the two true curves that we do not know in Figure 2.4 (the red and black curves). Therefore, $\mu_1(0)$ is the value of the function at the cutoff.

On average, the slope in Equation 2.10 would estimate the first derivative of Equation 2.12, since the first derivative corresponds to the slope of the tangent, and here we have a function that could be non-linear being approximated by a linear function. It's as if we were estimating the slope of the function at the cutoff. Therefore, the coefficient of the slope ($\hat{\beta}^+$) will estimate the first derivative of this function $\mu_1(x)$. The term $o_{\mathbb{P}}(h)$ is simply a residual that will disappear if the sample is large enough.

Instead, $\hat{\alpha}^+$, which is the most relevant term, will estimate the exact value of the function at the cutoff, which is precisely what we want to obtain. In Equation 2.11,

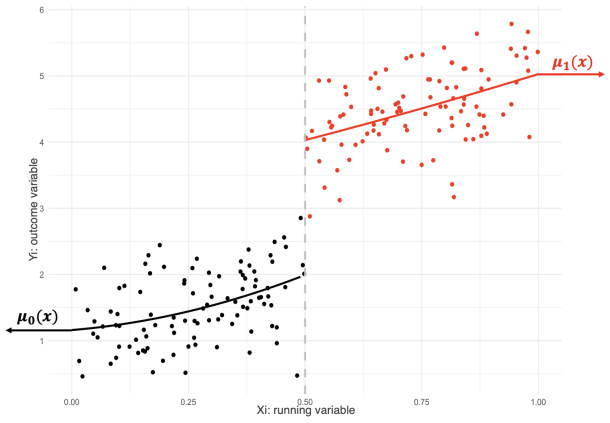


Figure 2.4: The potential outcome as a function of the running variable

there is also the term $h^2\mathcal{B}_+$ which indicates an approximation of the bias of the estimator (this bias exists because we are using the non-parametric approach). This additional term depends on two things:

1. The term $\mathcal{B}_+$, which is a constant that depends on the second derivative of the function. The second derivative represents the curvature, allowing us to consider the amount of non-linearity. Therefore, there is a bias term, and we attempt to use a linear function to approximate a function that is actually non-linear and has a certain curvature. The bias will depend on how non-linear the true function is. Since the second derivative of a linear function is 0, the more linear the true function is, the more the constant $\mathcal{B}_+$ will tend to 0, and therefore the bias will be reduced. Conversely, if the true function is far from being linear, the second derivative will be far from 0, and the bias will increase. An example of this is shown in figures 2.5 and 2.6. In Figure 2.5, if we tried to estimate the true functions with linear functions, we would get an excellent result, since near the cutoff these functions are almost linear, and the bias will be very low. On the other hand, in Figure 2.6, the opposite would happen, and we would have a very high bias because the true functions are very non-linear.

2. The term h^2, which represents the bandwidth squared. Clearly, the smaller the bandwidth, the smaller the bias, and the closer it will be to 0. In Figure 2.7, if we take the bandwidth h_1 (which is smaller than h_2), we are reducing the bias compared to what we would get with h_2, because in the first bandwidth, we are approximating the true function with a linear function, getting very close to a good solution. However, if we took h_2, we would increasingly move away from linearity, causing the bias to increase.

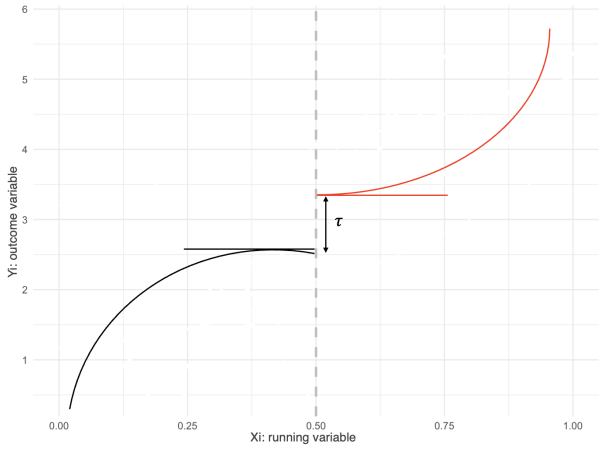


Figure 2.5: The bias problem: a good estimate of τ

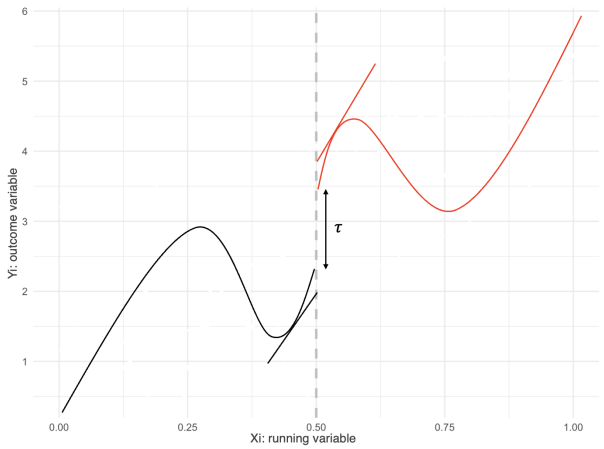


Figure 2.6: The bias problem: a bad estimate of τ

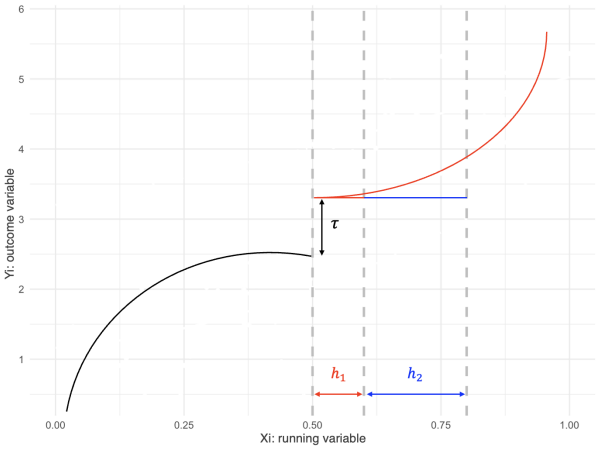


Figure 2.7: Bias reduction by reducing the bandwidth

Thus, to summarize, the bias will depend both on the amount of non-linearity in the true model and on the bandwidth we choose. On one hand, the amount of non-linearity will depend on how the true model behaves, something that cannot actually be controlled. On the other hand, it is possible to control and choose the bandwidth. Therefore, we can use the following formula to choose the bandwidth and reduce the bias as much as possible:

$$\mathbb{E}[\hat{\alpha}^+|X] - \mu_1(0) = h^2\mathcal{B} + o_{\mathbb{P}}(h^2) \tag{2.13}$$

If h were equal to 0, the term $h^2\mathcal{B}$ would be 0, and the estimator would be unbiased. However, this is not possible because h^2 cannot be 0, as in that case, there would be no observations within the bandwidth. Another issue arises from the variance: if we choose a very small bandwidth, there will be few observations within it, and the variance of the estimator will increase. Therefore, we need to look at the variance of the estimator at the cutoff, which will be given by:

$$\text{Var}[\hat{\alpha}^+|X] = \frac{\mathcal{V}+}{nh} + o_{\mathbb{P}}\left(\frac{1}{nh}\right) \tag{2.14}$$

In Equation 2.14, we again have the term $o_{\mathbb{P}}$, which is still something that will disappear once the sample is sufficiently large. The term $\mathcal{V}_+$ (the subscript + indicates that we are on the right side of the cutoff) is a constant that indicates the variability of the outcome around the cutoff. This term, which depends on the true model that cannot be controlled or known, is divided by nh, following the same principle as for the sample mean, where the variability of the estimator is given by its variance divided by the number of observations n. This time, the sample size is multiplied by the bandwidth h, resulting in the term nh, which is referred to as the *"effective sample size"*. Such name is motivated by the fact that we do not use all observations for our estimates, but only those within the bandwidth. Again, the bandwidth is something we can control. If desired, we can also control the sample size, but this is very rare (for example, in surveys, it can be done by increasing the number of interviewees). All of this leads to the so-called ***"bias-variance trade-off"***: the closer we are to the cutoff, the closer the bias will be to 0, but the variance will increase because, as we go back to the term $\frac{\mathcal{V}_+}{nh}$, the denominator will become smaller and the variance will increase. If, on the other hand, we choose a very large bandwidth, the variance will be small because the estimator will be very precise, but the bias will be very high (see equation 2.14). Therefore, the question arises: how can we summarize this trade-off, and how can we use it to choose the bandwidth?

The above concerns the coefficient $\hat{\alpha}^+$, but, clearly, it holds also for the left side of

the cutoff, i.e. for the coefficient $\hat{\alpha}^-$. It is worth to recall that we are interested in the treatment effect, see Equation 2.14, given by the difference between the intercepts of the treated and untreated groups at the cutoff. Therefore, once we obtain more information about the coefficients $\hat{\alpha}^+$ and $\hat{\alpha}^-$, we can gather more information about the treatment effect $\hat{\tau}$:

$$\mathbb{E}[\hat{\tau}|X] - \tau = h^2\mathcal{B} + o_{\mathbb{P}}(h^2), \quad \mathcal{B} = \mathcal{B}^+ - \mathcal{B}^- \tag{2.15}$$

$$\text{Var}[\hat{\tau}|X] = \frac{\mathcal{V}}{nh} + o_{\mathbb{P}}\left(\frac{1}{nh}\right), \quad \mathcal{V} = \mathcal{V}^+ - \mathcal{V}^- \tag{2.16}$$

The bias of the treatment effect estimator will again depend on the term $h^2\mathcal{B}$. However, its variance will depend on the term nh (effective sample size). At this point, the Mean-Squared Error (MSE) comes into play, which is given by the sum of the variance and the square of the bias:

$$MSE(\hat{\tau}) = Bias(\hat{\tau})^2 + \text{Var}[\hat{\tau}] \tag{2.17}$$

Thus, the MSE measures how far the estimator is from the true parameter. By substituting the terms mentioned earlier, we can approximate the MSE with the following equation:

$$AMSE(\hat{\tau}) \approx h^4\mathcal{B}^2 + \frac{\mathcal{V}}{nh} \tag{2.18}$$

These substitutions represent a significant advantage, since the equation will depend on the constant $\mathcal{B}$ and the bandwidth h that is chosen, as well as the variance. However, one must always consider the trade-off: if a small bandwidth is chosen, the term $h^4\mathcal{B}^2$ tends to 0, but $\frac{\mathcal{V}}{nh}$ increases.

It is possible to balance this trade-off by choosing the bandwidth that minimizes the MSE, solving a function that depends on h and minimizing it. In this way, we obtain the formula for the ***MSE-optimal bandwidth***:

$$h^*_{\text{MSE}} = \arg\min_h \ \text{AMSE}(\hat{\tau}) \tag{2.19}$$

$$= \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2}\right)^{1/(2p+3)} n^{-1/(2p+3)} \tag{2.20}$$

where p indicates the order of the polynomial, and the optimal bandwidth depends on three elements:

1. $\mathcal{V}$: the variability of the outcome at the cutoff. It positively depends on the variability of the outcome, because, if the outcome is very noisy, more

observations will be needed to try to reduce this noise;

2. $\mathcal{B}$: the bias of the estimator. It negatively depends on the bias because, if it is very large, the function will be very nonlinear, and the bandwidth will need to be narrowed to move closer to the cutoff, giving a better chance that the function will be approximately linear at least there. On the other hand, if the bias is small, it means the function will be more linear even away from the cutoff, allowing for a larger bandwidth;

3. n: the sample size. This term is raised to the power of $-1/5$, meaning that the larger the sample size, the smaller the optimal bandwidth tends to be, because with more observations available, a sufficient number will lie close to the cutoff, allowing the estimator to focus on a narrower neighborhood around the threshold.

Since the treatment effect is given by the difference between two different functions at the cutoff, in some applications, an asymmetric bandwidth is chosen, meaning different widths on the right and left of the cutoff. In these cases, the formula for the MSE-optimal bandwidth becomes:

$$h_{MSE,-} = \left(\frac{\mathcal{V}_-}{2(p+1)\mathcal{B}_-^2} \right)^{1/(2p+3)} n_-^{-1/(2p+3)} \tag{2.21}$$

$$h_{MSE,+} = \left(\frac{\mathcal{V}_+}{2(p+1)\mathcal{B}_+^2} \right)^{1/(2p+3)} n_+^{-1/(2p+3)} \tag{2.22}$$

where $h_{MSE,-}$ and $h_{MSE,+}$ represent the bandwidths for the control and treatment groups, respectively.

2.2.2 Inference procedure

A final step concerns inference: we need to construct confidence intervals and hypothesis tests to determine whether this effect is truly significant or not. When we want to apply the central limit theorem (if we have a sufficiently large sample, our estimator will be approximately normal), which is typical in inference, our estimator will be approximately normal but characterized by bias:

$$\sqrt{nh*}(\hat{\tau} - \tau) \xrightarrow{D} N(B, \Omega) \tag{2.23}$$

Thus, it will no longer be the usual approximately normal distribution with mean 0 and some variance, but rather an approximately normal distribution with bias and

some variance. The explanation is that the optimal bandwidth width selected to balance bias and variance will never be such that it reduces the bias to 0. Therefore, it is essential to consider this because if we ignored the bias, the p-values and confidence intervals would be incorrect.

To ensure the estimates are correct, one of the best approaches is called *"bias correction"*. The idea of this method is to recognize the presence of bias and subtract it from the estimator to correct for it:

$$\sqrt{nh*}(\hat{\tau} - \tau - B) \xrightarrow{D} N(B, \Omega) \tag{2.24}$$

However, we need to estimate the bias since it is unknown. Therefore, instead of having the term B, we will have:

$$\sqrt{nh*}(\hat{\tau} - \tau - \hat{B}_n) \xrightarrow{D} N(B, \Omega + \Sigma) \tag{2.25}$$

This time, there is an additional term, Σ, because every time something is estimated (in this case, the bias), it affects the variance of the estimator. Confidence interval construction can also be done in different ways. In particular, it is possible to construct confidence intervals by erroneously ignoring the presence of bias:

$$CI_{US} = [\hat{\tau} \pm 1.96 \cdot \sqrt{\mathcal{V}}] \tag{2.26}$$

where the subscript US stands for "Undersmoothing."

A more robust strategy consists of constructing confidence intervals accounting for bias through the "bias correction" approach:

$$CI_{bc} = [(\hat{\tau} - \hat{\mathcal{B}}) \pm 1.96 \cdot \sqrt{\mathcal{V}}] \tag{2.27}$$

where the subscript bc stands for bias correction, and an estimator $\hat{\mathcal{B}}$ for the bias $\mathcal{B}$ is included.

Finally, an even more robust strategy, known as "robust bias correction," can be employed. This strategy, which performs better both theoretically and in finite samples, ensures smaller coverage errors and a shorter average length compared to the previous two. Moreover, robust bias correction is valid even when using the MSE-optimal bandwidth in point estimation, and it allows the same data to be used for both point estimation and inference. Confidence intervals generated with robust bias correction are based on eliminating the estimated bias term $\hat{\mathcal{B}}$ from the RD point estimator. However, unlike CI_{bc}, this approach allows the estimated bias term to converge in distribution to a random variable, thereby contributing to

the approximation of the RD estimator's distribution. This setup leads to a new asymptotic variance $\mathcal{V}$ that includes the variability introduced by the bias correction process. Consequently, $\mathcal{V}$ is larger than the conventional OLS variance when using the same bandwidth, thus incorporating the additional uncertainty arising from bias estimation:

$$CI_{rbc} = [(\hat{\tau}_{SRD} - \hat{\mathcal{B}}) \pm 1.96 \cdot \sqrt{\mathcal{V}bc}] \tag{2.28}$$

2.2.3 Assessing the validity of Regression Discontinuity

As previously mentioned, RD requires compliance with some fundamental assumptions to ensure the validity of causal estimates. In particular, we do not want units that have exact control over their score, otherwise, selection bias would be generated, and the estimates would be invalid. Therefore, it is important to provide some evidence regarding the validity of the technique and its assumptions. Specifically, the assumptions to be respected are:

- **No sorting around the cutoff**. RD is invalid if individuals can manipulate the score. Returning to the example of students mentioned in the Chapters, the technique will not be valid if students can decide exactly what score to obtain on the exam in such a way as to enter the treatment group to benefit from the treatment itself. If this were not the case—if students could choose whether to be in the treatment or control group near the cutoff—selection bias would be generated. In Figure 2.8, an example of running variable manipulation is shown. Specifically, in the left part of the graph, the assumption is respected as there is no jump at the cutoff in the distribution of the running variable. On the other hand, in the right part of the graph, there is a clear jump at the cutoff, meaning that, probably, the units were aware of the implementation of a certain policy/program for which they were able to manipulate the running variable by deciding whether to be in the treatment or control group. One way to check for this issue is to test for discontinuities in the density function of the running variable using the so-called *"McCrary density test"* (McCrary, 2008), which will give to us an objective evidence regarding the sorting.

- **Continuity away from the cutoff**. It is important to remember that identification relies on the continuity of $E[Y_i(d)|X_i]$. In reality, it is not possible to directly check the continuity of the functions of the treated and untreated at the cutoff, because we only observe the functions to the right or left of the cutoff, never both, as it would be impossible to observe the function of the treated if they had never received the treatment or vice versa. However, it is

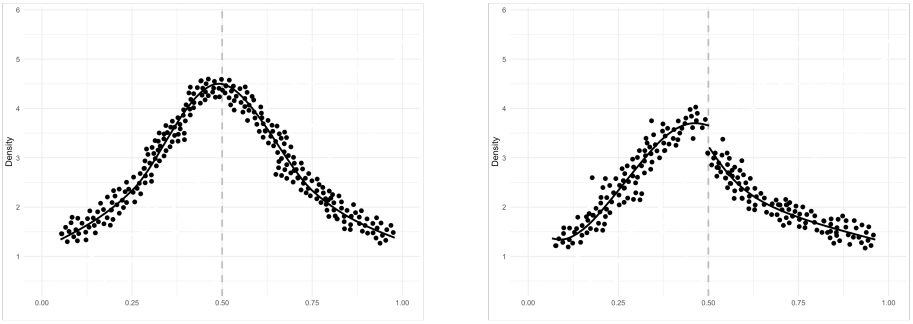


Figure 2.8: No sorting around the cutoff assumption

possible to check for continuity at other points away from the cutoff. If there were other discontinuities far from the cutoff, one could no longer be sure that $E[Y_i(d)|X_i]$ is continuous—there could be other treatments at play beyond the one being estimated. Testing this assumption is quite simple: one only needs to change the cutoff value and check whether a significant treatment effect appears.

– **No discontinuities in covariates and placebo outcomes**. In many cases, it is well known that there are variables for which no treatment effect should be present. One example is provided by variables that, in RD terminology and in experimental studies in general, are called "placebo outcomes," meaning outcomes that should not be affected by the treatment. In reality, a distinction should be made between predetermined covariates and placebo outcomes. Predetermined covariates are variables determined before treatment assignment and therefore cannot be influenced by the treatment itself (e.g., gender or age). Placebo outcomes, on the other hand, are variables determined after treatment assignment that, based on causal logic, should not be influenced by the treatment. For example, if the treatment concerns the implementation of safety measures for certain cars, one would expect a treatment effect on mortality due to car accidents but not on mortality from other causes, such as cancer deaths. Thus, cancer mortality would be a good placebo outcome in this case. The reasoning here is quite simple. As mentioned earlier, RD aims to simulate a sort of randomness in the assignment mechanism to the treatment or control group to ensure that the treatment effect is isolated. To achieve this, units very close to the cutoff are compared, units that are similar in every respect except for the outcome and the fact that, perhaps due to pure luck, they ended up in different groups. Therefore, these two groups will differ only because some received the treatment and others did not, which is why a discontinuity in the outcome is generated. For this reason, to validate the

design, the two groups must be identical in all other aspects, meaning in their covariates. Thus, covariates should not show any discontinuity at the cutoff. Otherwise, if they did, the two groups would differ in other ways as well, and it would no longer be possible to isolate the treatment effect on a given intervention or policy outcome. This assumption is also very easy to test since one only needs to estimate the regression discontinuity using covariates as the outcome and check that there is no discontinuity at the cutoff. In general, covariates are not essential in RD. Sometimes they are included only to reduce the variance of the estimator and increase its precision (Cattaneo et al., 2023; Calonico et al., 2019; Frölich and Martin, 2019).

2.3 Sharp and Fuzzy Regression Discontinuity

The two most well-known RD designs are the *Sharp* design and the *Fuzzy* design. As mentioned in Chapter 1, in a Sharp RD design, crossing the cutoff necessarily means receiving the treatment, while not crossing the cutoff means not receiving the treatment. In many cases, being assigned to the treatment does not necessarily imply receiving it. Understanding compliance and non-compliance is crucial for interpreting RD results. In a Sharp RD design, all units whose score value exceeds the cutoff point enter the treatment group

- If $X_i \geq c$, the unit i will be assigned to the treatment group.
- If $X_i < c$, the unit i will be assigned to the control group.

Fuzzy RD, on the other hand, is characterized by "imperfect compliance," which means that some units above the cutoff point (eligible for treatment) decide not to join to the treatment, while some units below the cutoff (not eligible for treatment) manage to obtain it. To have an RD design in this case, discontinuity is still needed, meaning that there must be a jump in the probability of receiving treatment at the cutoff. However, unlike in the Sharp RD design, this probability will no longer be strictly 0 or 1. Thus, there will be a jump in the probability of treatment at the cutoff, but this jump will not necessarily be equal to 1. This is illustrated in Figure 2.9, where we can see how the probability of receiving treatment at the cutoff differs depending on whether we are in a Sharp (left side) or a Fuzzy design (right side). In practice, many examples of Fuzzy RD can be found. One of the most common cases of Fuzzy RD is represented by a program aimed at combating poverty. In this case, all individuals with an income below a certain threshold benefit from the treatment. However, given the importance of the policy, administrators might

decide to extend the practice to some families very close to the cutoff, even if they are not eligible to be treated. In many cases, it may also happen that some units refuse treatment.

It can be stated that Sharp RD is a particular case of Fuzzy RD in which the probability of receiving treatment jumps from 0 to 1 at the cutoff. In Fuzzy RD, treatment assignment is defined as:

$$Z_i = 1(X_i \geq c) \tag{2.29}$$

This equation indicates whether a unit is above or below the cutoff, and, hence, whether the unit is assigned to treatment or not. The treatment status is given by:

$$D_i = D(X_i) \tag{2.30}$$

where D is a function of the score but is endogenous due to self-selection, meaning that even if a unit is above the cutoff, it may decide not to be treated, and vise versa. In the Sharp design, we have $D_i = Z_i$. In contrast, in the Fuzzy RD, D_i could be different from Z_i.

In this scenario, it is no longer sufficient to compare the treated and untreated groups, as receiving treatment is now based on a decision and is endogenous, making the observations non-comparable. The fundamental idea remains the same: comparing units just above and just below the cutoff but in a different way, borrowing some concepts from the Instrumental Variable (IV) literature (Pearl, 2000; Angrist and Krueger, 2001).

In Fuzzy RD, multiple scenarios can arise since it is no longer simply the case that units above c are treated and those below are not. Thus, the treatment status of unit i just above/below the cutoff is:

$$D_{1i} = \lim_{x \downarrow c} D_i(x), \quad D_{0i} = \lim_{x \uparrow c} D_i(x) \tag{2.31}$$

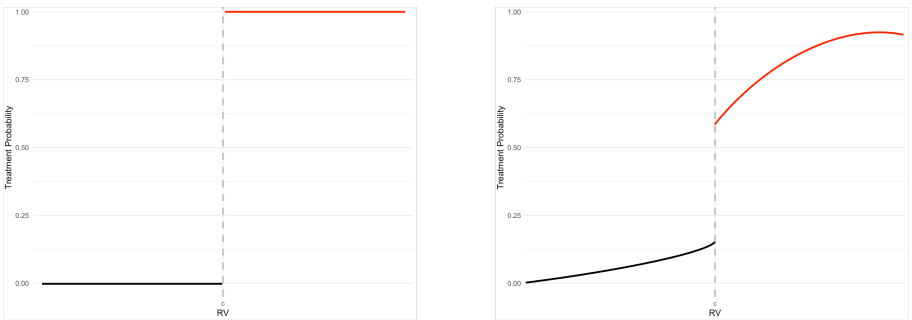


Figure 2.9: Treatment probability in Sharp (left side) and Fuzzy design (right side)

Now, four possibilities arise:

1. **Always-takers**: $D_{1i} = D_{0i} = 1$. These are units that, regardless of their position, always manage to access and receive the treatment. Thus, if they are above c, they will receive the treatment, and if they are below, they will manipulate the system to obtain it.

2. **Never-takers**: $D_{1i} = D_{0i} = 0$. These are units that, regardless of their position, never receive the treatment.

3. **Compliers**: $D_{1i} = 1, D_{0i} = 0$. These are units that do not receive treatment when they are below c, but receive it when they are above c.

4. **Defiers**: $D_{1i} = 0, D_{0i} = 1$. These are units that should not receive treatment when they are below c, yet they still manage to obtain it. Conversely, when they are above c and should receive the treatment, they do not.

In a Sharp RD scenario, we have only one type of unit described above: the Compliers. In a Fuzzy RD scenario, however, all four possibilities described above can be present. To better explain the Fuzzy RD, we must start with a key concept. We know that in the Sharp RD, it is sufficient to compare treated and untreated units, and the vertical difference at the cutoff will provide the average treatment effect (ATE). One might think of extending the same procedure to the Fuzzy RD and then observing what happens. This is what is called the *Intention To Treat (ITT)* parameter. Essentially, this approach ignores the fact that we are in a Fuzzy scenario and simply replicates what would be done in a Sharp scenario. Thus, we compare the outcome of units just above the cutoff with the outcome of units just below it:

$$\begin{aligned} \tau_{ITT} &= \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x] \\ &= \mathbb{E}[(Y_i(1) - Y_i(0))(D_{1i} - D_{0i}) | X_i = c] \end{aligned} \tag{2.32}$$

where:

$$Y_i(1) - Y_i(0) = \tau_i$$

$$D_{1i} - D_{0i} \begin{cases} 1 & \text{for Compliers} \\ -1 & \text{for Defiers} \\ 0 & \text{for Always Takers/Never Takers} \end{cases}$$

In Equation 2.32, there is an expectation at the cutoff ($\mathbb{E}|X_i = c$) and two terms. The first one is simply the difference between the potential outcomes of treated and untreated units, $Y_i(1) - Y_i(0) = \tau_i$ (in a Sharp RD design, this would be the end of the analysis). Since there is imperfect compliance, there is an additional term, $D_{1i} - D_{0i}$, which represents the decision of the units. Thus, we are computing the average treatment effect at the cutoff and then multiplying it by this extra term. In particular, regarding this extra term, if there were only Compliers, it would be equal to 1 since $D_{1i} = 1$ and $D_{0i} = 0$. Instead, in case there are only Always Takers, the units would receive the treatment regardless of their position relative to the cutoff, so the extra term would always be equal to 0 since $D_{1i} = 1$ and $D_{0i} = 1$. If there were only Never Takers, the result would be the same: the extra term would be equal to 0 since $D_{1i} = 0$ and $D_{0i} = 0$. This means that both Always Takers and Never Takers would be excluded from the calculation of the average treatment effect because the extra term is always 0. For Defiers, the opposite of Compliers happens. In this case, those above the cutoff refuse to receive the treatment, while those below still manage to obtain it. The extra term will always be equal to -1 since $D_{1i} = 0$ and $D_{0i} = 1$.

To summarize, the average treatment effect receives a positive weight of 1 for Compliers, a negative weight of -1 for Defiers, and a weight of 0 for Always Takers and Never Takers. This presents a problem because it means that we are averaging while assigning a negative weight to some units in the population (Defiers). This implies that the effects will have opposite signs for Compliers and Defiers. The issue is that we might estimate an effect of 0 in the parameter τ_{ITT} even if the actual treatment effect is not 0, simply because Defiers would counterbalance Compliers. Thus, one might mistakenly conclude that the effect is 0 just because the Compliers could be "canceled out" by the Defiers. This is the so-called *"Identification Problem"* that arises when there is imperfect compliance.

The problem is that we can only observe the effect for those who are treated or those who are not treated; we cannot know what would have happened to a treated unit if it had not received the treatment, just as we cannot know what would have happened to an untreated unit if it had received the treatment. The only possible solution is to assume that there are no Defiers. This assumption is called *"monotonicity assumption"*:

$$D_{1i} \geq D_{0i}, \quad \forall i \tag{2.33}$$

With this assumption, we are essentially saying that D_{1i} cannot be smaller than D_{0i}. Intuitively, the monotonicity assumption implies that units just above the cutoff

cannot reduce their probability of receiving the treatment, but they also cannot increase it. Based on this assumption, the difference $D_{1i} - D_{0i}$ can only be 1 (for Compliers) or 0 (for Always Takers or Never Takers). Now, the parameter τ_{ITT}, which, to recap, is calculated by simply performing a standard RD comparison of the outcome of units just above the cutoff with those just below, can be written as:

$$\tau_{ITT} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c, D_{1i} > D_{0i}] \cdot \mathbb{P}[D_{1i} > D_{0i}|X_i = c] \tag{2.34}$$

where:

- $\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c, D_{1i} > D_{0i}]$ is the average treatment effect at the cutoff for Compliers ($D_{1i} > D_{0i}$), which is called $LATE$ because it represents the Local Average Treatment Effect;
- $\mathbb{P}[D_{1i} > D_{0i}|X_i = c]$ is the proportion of Compliers.

Multiplying by the proportion of Compliers acts as a scaling factor on the effect. There could be cases where τ_{ITT} is close to 0 even if $LATE$ is large because there are very few Compliers. However, the τ_{ITT} parameter remains an interesting metric for policymakers because it measures the effect of "offering the treatment" to units. In fact, this is crucial before implementing a certain policy.

At this point, it becomes important to explain why, in this case, we are able to identify only the effect on Compliers. The fundamental idea in an RD design is to exploit this discontinuity at the cutoff, leveraging the fact that if a unit is just above the cutoff, its behavior will be slightly different from another unit that is just below. However, in a Fuzzy RD, this is true only for the Compliers because if a unit is an Always Taker or a Never Taker, its behavior will not change, regardless of whether it is above or below the cutoff point. Therefore, an RD strategy will say nothing about the behavior of these two subpopulations because their behavior will not be affected by crossing the cutoff. Thus, in a Fuzzy scenario, the best that can be done is to say something about those units that change their behavior when they cross the cutoff, and these units are represented by the Compliers.

For the moment, let us ignore the fact that we are in a Fuzzy scenario and use the notions introduced so far, simply comparing the outcome of units just above the cutoff with that of units just below, multiplying the $LATE$ by the proportion of Compliers. The next step is to leverage the so-called *"first stage"*, a terminology that comes from the literature on Instrumental Variables (IV). The first stage simply concerns the effect of treatment assignment on the treatment status, which, in this case, simply refers to the proportion of treated units just above the cutoff compared

to the proportion of treated units just below:

$$\tau_{FS} = \lim_{x \downarrow c} \mathbb{E}[D_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i | X_i = x] \tag{2.35}$$

where:

- $\lim_{x \downarrow c} \mathbb{E}[D_i | X_i = x]$ represents the proportion of units currently treated just above the cutoff;
- $\lim_{x \uparrow c} \mathbb{E}[D_i | X_i = x]$ represents the proportion of units currently treated just below the cutoff.

If this were a Sharp RD scenario, the first term would be equal to 1 and the second equal to 0, since the jump at the cutoff goes from 0 to 1. In the Fuzzy scenario, however, this condition is not necessarily fulfilled, as it is possible to have treated units below the cutoff and untreated units above. Nevertheless, we can still compute this difference and show that, under the monotonicity assumption, the first stage coefficient τ_{FS} is equal to the proportion of Compliers:

$$\tau_{FS} = \mathbb{P}[D_{1i} > D_{0i} | X_i = c] = \mathbb{P}[complier | X_i = c] \tag{2.36}$$

This is because we know that above the cutoff we can find either Compliers or Always Takers; therefore, the term $\lim_{x \downarrow c} \mathbb{E}[D_i | X_i = x]$ is given by the proportion of Compliers (P_C) plus the proportion of Always Takers (P_{AT}). Conversely, we know that below the cutoff we can find only the proportion of Always Takers; thus, $\lim_{x \uparrow c} \mathbb{E}[D_i | X_i = x]$ is given by the proportion of Always Takers (P_{AT}). However, by computing the difference between these two, we obtain exactly the proportion of Compliers:

$$(P_C + P_{AT}) - P_{AT} = P_C \tag{2.37}$$

Thus, the first stage coefficient identifies the proportion of Compliers at the cutoff. This is a very positive result because, looking at the parameter τ_{ITT} in the Equation 2.34, it is something that can be easily estimated from the data, as it depends only on observable variables. It represents the average outcome of units just above and just below the cutoff. Once computed, we obtain the $LATE$ for the proportion of Compliers.

Next, we need to consider the proportion of treated units above and below the cutoff to recover the proportion of Compliers. Finally, we simply divide the param-

eter τ_{ITT} by τ_{SF} to recover the $LATE$:

$$\frac{\lim_{x\downarrow c}\mathbb{E}[Y_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[Y_i|X_i=x]}{\lim_{x\downarrow c}\mathbb{E}[D_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[D_i|X_i=x]}=\tau_{FRD} \tag{2.38}$$

where the numerator represents the parameter τ_{ITT} and the denominator the parameter τ_{FS}. By dividing the two terms, we obtain the Fuzzy RD parameter, which is the true $LATE$ at the cutoff:

$$\tau_{FRD}=\mathbb{E}[Y_i(1)-Y_i(0)|X_i=c,D_{1i}>D_{0i}] \tag{2.39}$$

It is important to keep in mind that this is a parameter only for the Compliers, which is the subpopulation that changes its behavior depending on whether the units are above or below the cutoff. This is precisely what we are interested in when using RD. In fact, the behavior of Never Takers or Always Takers, is not of interest since it remains the same regardless of their score and whether they are above or below the cutoff. Each term in Equation 2.38 is simply an RD estimate. The numerator is an RD using Y_i as the outcome and X_i as the score, while the denominator is an RD using D_i as the outcome and X_i as the running variable. Therefore, it is possible to estimate two separate RDs and then construct the ratio of these two estimators using all the tools presented so far (bandwidth selection, local polynomial estimation, robust bias-corrected inference, etc.). It is interesting that using local constant regression (estimating two flat functions), it is numerically equivalent to implementing a Two Stage Least Squares (2SLS) regression.

2.4 Local randomization approach

The fundamental logic of RD is that, near the cutoff, being treated or not is almost as if it were random. For example, returning to the example of scholarships, we do not compare students who have scores of 0.80 and 0.20, as they will differ in many aspects; rather, considering a cutoff value equal to 0.5, we compare students who received 0.51 and 0.49, where the first ones may have been slightly lucky on the test and guessed a question that allowed them to surpass the threshold. Thus, near the cutoff, receiving the treatment or not is something very close to randomness. However, the framework presented so far does not formalize this interpretation. The only aspect discussed concerns the need for regression functions to be continuous at the cutoff, which is a good practice.

At this point, a natural question arises: in which cases can we think of RD as a local experiment? Every experiment can be seen as a specific type of RD in which:

– The score is a uniformly distributed random variable.

– The cutoff is chosen to ensure a given probability of treatment.

Just think about how units are assigned to treatment and control in an RCT. In such a case, one simply takes the list of units and assigns them entirely at random to the treatment and control groups, as flipping a coin. The fundamental difference between the two techniques is that, in an RCT, the score is uncorrelated with the potential outcome by design. In fact, the uniform random variable will have no effect on the outcome in any way. Instead, in the classic RD design, the mechanism is different because, as an example, it is known that the infant mortality rate strongly depends on individuals poverty level, or that salary somehow depends on whether one has attended university or not, and so on. Graphically, the main difference between the two is that randomized experiments can be seen as an RD design where the regression functions are flat (Figure 2.10, left side), whereas in the classical RD, the functions may change depending on the score (Figure 2.10, right side). It is well known that functions in RD are not flat, but one can assume that close to the cutoff point, the idea of flat regression functions is a fairly good approximation. In Figure 2.11, if we look only at the bandwidth between $c - w_0$ and $c + w_0$, the assumption that these functions are flat is likely to be much more reasonable. This is the main idea of the *Local randomization approach* (Cattaneo et al., 2024; Cattaneo et al., 2016; Cattaneo et al., 2018). Thus, the idea is that there exists a window $W_0 = [c-w, c+w]$ around the cutoff in which a randomized experiment exists. The following two conditions must be fulfilled for a randomized experiment:

– The probability distribution of X is not correlated with individuals' characteristics. This means that covariates, in the sense of characteristics, are balanced, particularly for the running variable.

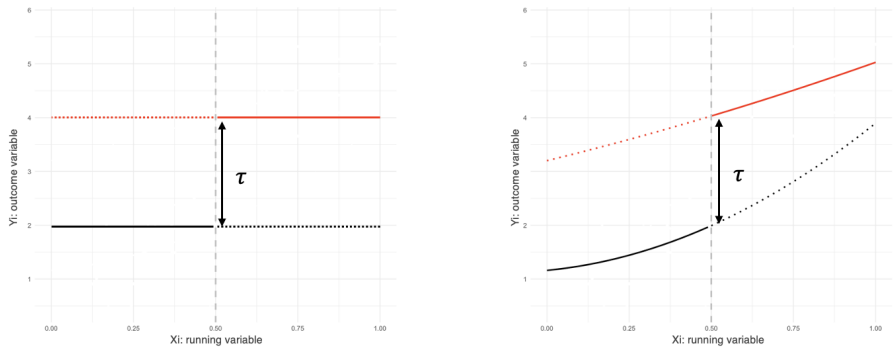


Figure 2.10: Randomized experiment (left side) vs Regression Discontinuity Design (right side)

– The regression functions are flat. Thus, the outcome is not affected by the score value: $Y_i(d, x) = Y_i(d)$.

The first condition states that everyone has the same probability of being treated or not at a point very close to the cutoff. Returning to the scholarship example, it is like saying that those who scored 0.51 and 0.49 differ simply because they randomly guessed or missed a question. The second condition states that these two functions are locally flat. It is clear that these conditions are much stronger than the previous ones, as we are assuming that in a given window, the functions take on a specific form (flat). Thus, this approach will be defined as non-parametric, and it is precisely this feature that allows RD to be interpreted as a randomized experiment.

At this point, it is possible to leverage this idea in RD to determine the appropriate bandwidth to consider. If we assume that there exists a window where RD and randomized experiments coincide, we "only" need to identify it. Thus, we look at the covariates and select only the bandwidth in which they are balanced. This is a completely different approach from the classical bandwidth selection method that uses algorithms minimizing MSE or other criteria. To achieve this, an iterative procedure composed of multiple steps is used:

– Step 1: Choose a statistic to measure the balance of covariates, for example, the mean difference, comparing the proportion of males and females, the average age, or the average income between the two groups; or using something more precise like the so-called *"Kolmogorov-Smirnov"* method, which compares the distributions of covariates rather than just the means to check if they are balanced.

– Step 2: Consider an initially very small window $W_0^{(1)} = [c - w_{(1)}, c + w_{(1)}]$.

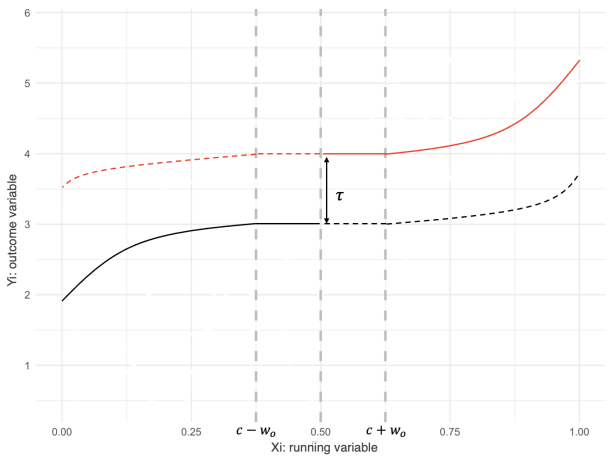


Figure 2.11: The Local randomization approach idea in RD

– Step 3: If there is covariate balance in the first selected bandwidth, then it is possible to slightly increase the initial window.

The procedure continues iteratively until a situation of imbalance between covariates is found. This procedure is illustrated graphically in Figure 2.12. Once the bandwidth is identified, we can proceed as in a randomized experiment, where the treatment effect can be deduced simply by comparing the mean outcomes of the treated and untreated individuals:

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 \tag{2.40}$$

The main limitation of this approach is that no information are available on unobserved covariates. For this reason, it is generally used as a complement to the classical RD design.

2.5 Multi-Dimensional Regression Discontinuity

So far, we focused on the standard RD, where units are characterized by a score that allows them to receive a specific treatment if it exceeds the cutoff point. In this type of design, both the score and the cutoff were scalars. However, it is very common to encounter designs characterized by more than one score, more than one cutoff, or both. These designs are known as Multi-Dimensional RD since the score and/or cutoff are no longer scalar but multidimensional, implying that the treatment threshold depends on multiple dimensions rather than a single scalar value. This adds complexity to the design but allows for the modeling of more realistic scenarios where multiple factors simultaneously influence treatment assignment.

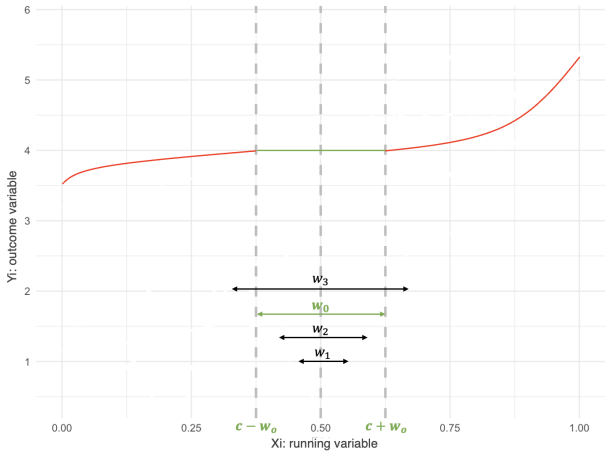


Figure 2.12: Bandwidth choice in the Local randomization approach

A typical example of Multi-Dimensional RD, could be the case a scholarship program for students that may be awarded based on both an economic score (e.g., students with a family income below 10,000 euros) and a merit-based score (e.g., students with an average exam grade above 26). It is essential to distinguish between three different scenarios: Multiple cutoffs, Cumulative cutoffs, and Multiple scores, as detailed in the following subsections.

2.5.1 Multiple cutoffs

In a multiple cutoffs scenario, the cutoff changes depending on the region or over time. However, the treatment remains the same in the sense that all units will receive the same treatment once they exceed their respective cutoff. For example, suppose two programs are implemented, one in Italy and one in France, to combat poverty. A subsidy is offered to all individuals whose annual income is below €10,000 in Italy and below €15,000 in France: for an Italian citizen, the running variable is centered at 0 by subtracting 10,000, while for a French citizen, it is centered at 0 by subtracting 15,000. By doing so, the problem is reduced to a single cutoff, which is 0. In this case, it is very common to solve the problem using the "normalizing and pooling" method, which essentially consists of recentering the running variable according to the cutoff each unit faces, meaning that a different cutoff value is subtracted depending on the observations.

To better understand the normalizing and pooling method, suppose there is a cutoff. Each unit has its own value C_i, and consider the case where there is a finite and discrete set of cutoffs:

$$\mathcal{C} = c_0, c_1, ..., c_j$$

At this point, we simply recenter the running variable based on the cutoff value:

$$\tilde{X}_i = X_i - C_i$$

After this transformation, the cutoff will be the same for everyone and equal to zero. Once this problem is reduced to a classic RD design with a single cutoff, estimation can proceed in the usual way. Thus, to obtain the treatment effect, units just above the cutoff are compared with those just below it based on the newly recentered value of the running variable:

$$\tau^p = \lim_{x \downarrow 0} \mathbb{E}[Y_i | \tilde{X} i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i | \tilde{X}_i = x] \tag{2.41}$$

If there were only one cutoff, the average treatment effect at that cutoff would be

estimated. However, in this case, there is no longer a single cutoff but multiple cutoffs for different units. Thus, the question is about what type of parameter is being identified when using this normalizing and pooling approach, given that the parameter τ defines, by definition, the treatment effect at the cutoff but, in this case, we have multiple cutoffs.

Under certain continuity assumptions discussed earlier, the parameter τ^p, obtained after recentering the running variable, becomes:

$$\tau^p = \sum_{c \in \mathcal{C}} \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c, C_i = c] \omega(c) \tag{2.42}$$

Where:

- $Y_i(1) - Y_i(0)$: represents the treatment effect for unit i
- $X_i = c$: indicates that the treatment effect is measured at the cutoff
- $C_i = c$: indicates that the treatment effect is measured for each cutoff within the data
- $\omega(c)$: represents weights.

Thus, Equation 2.42 aims to calculate all treatment effects for different cutoffs and then compute an average using weights. In this way, a weighted average of the treatment effect at different cutoffs for different subpopulations is obtained. The weights used are given by:

$$\omega(c) = \frac{f_{X|C}(c|c)\mathbb{P}[C_i = c]}{\sum_{c \in \mathcal{C}} f_{X|C}\mathbb{P}[C_i = c]} \tag{2.43}$$

These weights depend on:

- $f_{X|C}(c|c)$: represents the density of the running variable at each cutoff
- $\mathbb{P}[C_i = c]$: represents the proportion of the sample at each cutoff.

However, there are disadvantages associated with the normalizing and pooling approach. The first is that policy relevance may be unclear because, with this approach, different treatment effects for different populations—characterized by different cutoffs for specific reasons—are combined, and as a result, they may not be comparable with one another (as if the heterogeneity characterizing them is eliminated). A second disadvantage is that using this approach might flatten variations that could actually be used to identify some other parameter of interest.

To better understand how RD works in a multiple cutoff scenario, let us introduce an example. Suppose we have two different schools where an education policy is implemented to assign scholarships to students. These schools are characterized by two different cutoffs: in the first school, students can access the treatment if their grade point average is above 50, while in the second school, the grade point average must be above 60. At this point, the regression function will be given by the expected potential outcome as a function of the score and the different cutoffs:

$$\mu_d(x, c) = \mathbb{E}[Y_i(d)|X_i = x, C_i = c], \quad d \in 0, 1 \tag{2.44}$$

For example, the expected potential outcome as a function of the grade point average for first school could, in principle, be a completely different function from the expected potential outcome as a function of the grade point average for second school, since these schools could be highly heterogeneous. Therefore, there is no reason to assume that the functions will be the same for the different schools. It is possible to define the Conditional ATE, which is the average treatment effect at the cutoff for each region separately:

$$\tau(x, c) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x, C_i = c] = \mu_1(x, c) - \mu_0(x, c) \tag{2.45}$$

Now, consider when or how one can explore this variation in the cutoff to understand something interesting about the treatment effect and its heterogeneity. In Figure 2.13, the previously described example is graphically represented. The first school, represented by the black functions, is characterized by the cutoff c_0. On the right there is the function for the treatment group, and on the left there is the function for the control group, with the treatment effect at the cutoff indicated in the figure as $\tau(c_0, c_0)$. The second school, represented by the red functions, faces a higher cutoff indicated as c_1. For the second school as well, the average treatment effect at the cutoff is indicated as $\tau(c_1, c_1)$. One interesting aspect would be to try to estimate the parameter $\tau(c_1, c_0)$, which represents the treatment effect at the higher cutoff for the first school. However, this parameter could only be measured if one were able to change the academic merit threshold for that school and increase it. If it could be estimated, some interesting comparisons could be made between the two schools. Alternatively, one could compare the parameter $\tau(c_1, c_0)$ with $\tau(c_0, c_0)$ to investigate whether the treatment effect is heterogeneous across the distribution of the running variable. Another option is to compare the parameter $\tau(c_1, c_0)$ with $\tau(c_1, c_1)$ to determine whether, given a certain level of merit, the effect of the scholarships is greater in the first or second school. In principle, in a standard RD design,

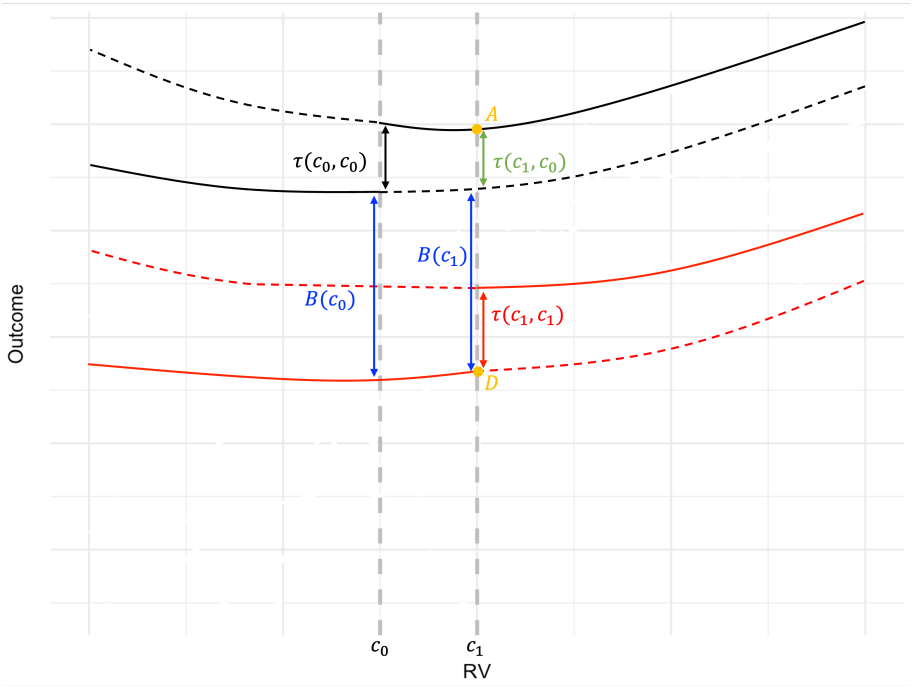


Figure 2.13: Treatment effect with multiple cutoffs

this would not be possible because, near the parameter $\tau(c_1, c_0)$, there would only be units that received the treatment and none belonging to the control group (Figure 2.13, point A). However, it is possible to observe some untreated units in the second school at the same level of the cutoff one is interested in (Figure 2.13, point D). Therefore, one could try to compare the treated units from the first school with the untreated units from the second school. This solution appears to address the overlap problem in RD but still does not provide a true measure of the treatment effect of interest. In fact, comparing the treated units in the first school with the untreated units in the second school would yield the treatment effect in the first school with a higher cutoff $\tau(c_1, c_0)$ plus a difference resulting from the fact that these regions could be very heterogeneous, indicated in the figure as $B(c_1)$. Even in the absence of treatment, the outcome would be different between the two schools because, for example, one might be wealthier than the other, or one might consist of students who dedicate more hours to studying. This difference can also be seen in the fact that the potential outcome of the untreated units—the dashed black and red lines—are at different levels across all points of the running variable.

Thus, comparing the treated units in the first school with the untreated ones in the second school would yield the treatment effect plus the bias, which cannot be estimated because it depends on the untreated units in the first school that will never be observed. However, it is possible to observe, near c_0, the untreated units from both regions. For this reason, one could estimate the bias $B(c_0)$ by shifting to

the lower cutoff level c_0, which can be observed because there are both untreated units from the first school and untreated units from the second school.

The question then becomes whether this bias is equal to the bias that cannot be estimated and is of interest for the calculation of $\tau(c_1, c_0)$, i.e., it is $B(c_0) = B(c_1)$. If one assumes that they are equal, then the bias term $B(c_1)$ would also be known and could be subtracted from the previously mentioned difference. This approach closely follows the logic of the *"parallel trend assumption"* typical of the *Difference-in-Differences (DID)* approach, except that instead of time, the x-axis represents the running variable, which is continuous.

Now, one must determine whether the equality assumption of the two biases at the different cutoffs makes sense. Technically, this is something that can never be known for certain because, as already stated, the bias at $B(c_1)$ will never be observed. However, one can examine the shape of the two functions from the starting point up to c_0 and check whether they are parallel, as shown in Figure 2.14 with the vertical blue lines. If they appear parallel at all points below c_0, then the assumption that they are equal is reasonable, and then, to obtain the treatment effect for the first school at c_1, given by $\tau(c_1, c_0)$, it is sufficient to subtract $B(c_1)$ instead of $B(c_0)$. This would mean that the bias is constant as a function of the score.

It is important to recall that one of the limitations of RD is that it is only possible to infer the treatment effect at the cutoff. In this multiple cutoff scenario, as long as one can sustain this assumption about the constancy of the bias, it becomes possible to extrapolate the treatment effect even away from the cutoff.

2.5.2 Cumulative cutoffs

In the case of Cumulative Cutoffs, the treatment is multivalued. Here different doses of treatment depend on the value of the score. For example, suppose there are two different cutoffs: if an observation is below the first cutoff, it will not receive any treatment; if it is above the first but below the second cutoff, it will receive a certain amount of treatment; if it is above the second cutoff, it will receive a different amount of treatment. Therefore, we have a treatment that assumes multiple values (different levels of treatment) depending on where the observation falls on the running variable. Thus, in addition to changing the cutoff, the treatment dose also changes. Instead, in the Non-Cumulative design presented earlier, a unit with a score $X_i = x$ can be exposed to any cutoff $c \in \mathcal{C}$, and the treatment dose remains the same, regardless of the cutoff it faces. In this scenario, we have a multivalued treatment $D_i \in d_1, d_2, \ldots, d_j$. In this case, a parameter of interest measures the

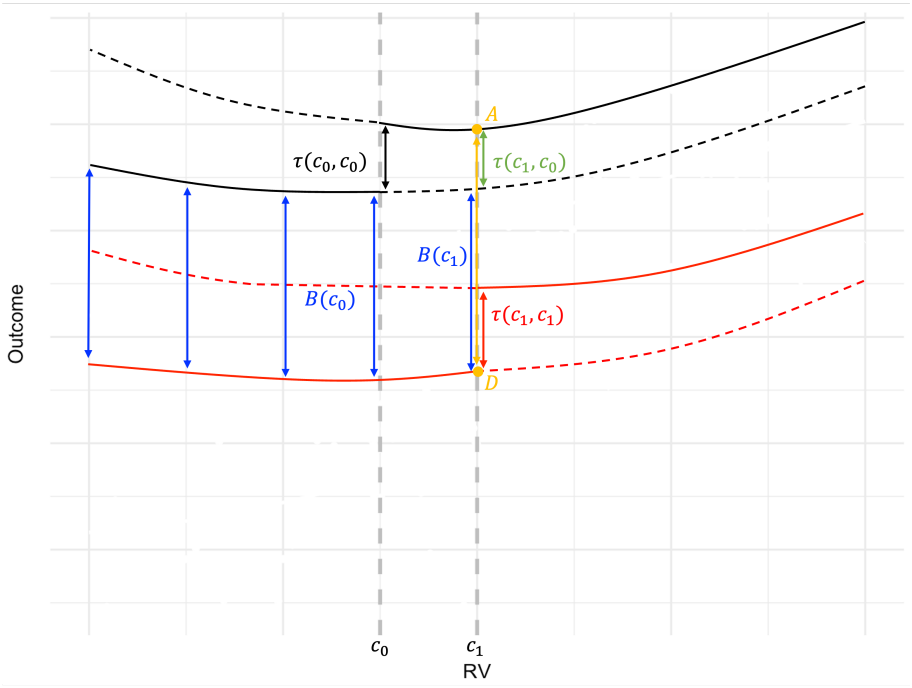


Figure 2.14: Testing the parallel trend in the bias assumption

switch from one treatment dose to another:

$$\tau_j = \mathbb{E}[Y_i(d_j) - Y_i(d_{j-1})|X_i = c_j] \tag{2.46}$$

where the difference $Y_i(d_j) - Y_i(d_{j-1})$ precisely measures the switch from treatment d_{j-1} to treatment d_j.

2.5.3 Multiple scores

In the Multiple scores scenario, treatment is assigned based on multiple running variables. For example, a scholarship might be awarded based on merit and economic status (above a certain merit threshold and below a certain poverty threshold).

The distinguishing characteristic of a Multiple Scores Design is that treatment is assigned based on multiple running variables. The most common case is when there are two running variables, $X_i = (X_{1i}, X_{2i})$. For instance, suppose that this time, scholarships are awarded based on entrance test scores evaluating skills in both Italian and Mathematics, such that students who score above 50 in Italian and above 60 in Mathematics receive the funding. Thus, assuming that b_1 and b_2 are the cutoff points for each dimension, the treatment is assigned if a unit has both scores above the respective cutoffs:

$$D_i = \mathbb{1}(X_{1i} \geq b_1)\mathbb{1}(X_{2i} \geq b_2) \tag{2.47}$$

The goal is to identify the multidimensional RD parameter:

$$\tau(b) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = b], \quad b \in \mathcal{B} \tag{2.48}$$

It is called multidimensional because, this time, instead of having a simple cutoff, there is a region in which each point can have a different treatment effect. This concept is illustrated graphically in Figure 2.15. In this example, a unit is treated if it scores above 50 in Italian and 60 in Mathematics on the entrance test. Graphically, this corresponds to the red area labeled "Treated area." The remaining blue areas correspond to the control area, labeled "Control area." The figure also shows how this design differs from the classic one presented so far, since, instead of a single cutoff, there are *boundaries* that define treatment group membership. The underlying logic remains the same: compare units just below the treatment area with those just above:

$$\tau(b) = \lim_{\substack{d(x,b)\to 0 \\ x\in B_t}} \mathbb{E}[Y_i \mid X_i = x] - \lim_{\substack{d(x,b)\to 0 \\ x\in B_c}} \mathbb{E}[Y_i \mid X_i = x] \tag{2.49}$$

where B_t and B_c are the treatment and control regions, respectively. The challenge lies in defining the distance from the new cutoff and measuring the treatment effect. To do so, various methods can be used to reduce the problem to a univariate RD estimation.

The first method is pooling, where a distance measure is defined for each unit by normalizing the running variable based on the shortest distance to the nearest boundary point, leveraging the approach explained in the Multiple Cutoffs section (see Section 2.8.1). An alternative approach, considered intermediate, is used when one wants to assess treatment effect heterogeneity along the boundary. The starting point consists of defining a set of boundary points of interest, which in the example shown in Figure 2.15 might include points such as $(50, 70), (75, 60)$. The second step involves defining a method to calculate distance (e.g., Euclidean distance) and then estimating a pooled RD at each specified point.

A special case of Multiple-Score RD is the Geographic RD (GRD). In a typical GRD design, the treatment and control areas correspond to adjacent regions or neighboring municipalities. The boundary set B corresponds to geographic borders that separate the two regions, and the score X_i corresponds to geographic coordinates (latitude and longitude). The cutoff is represented by a geographic boundary that separates two regions or an administrative/political border. An example of GRD is shown in Figure 2.16. The two polygons represent two different Italian regions (Lombardia in blue, Emilia-Romagna in red). The treatment region corre-

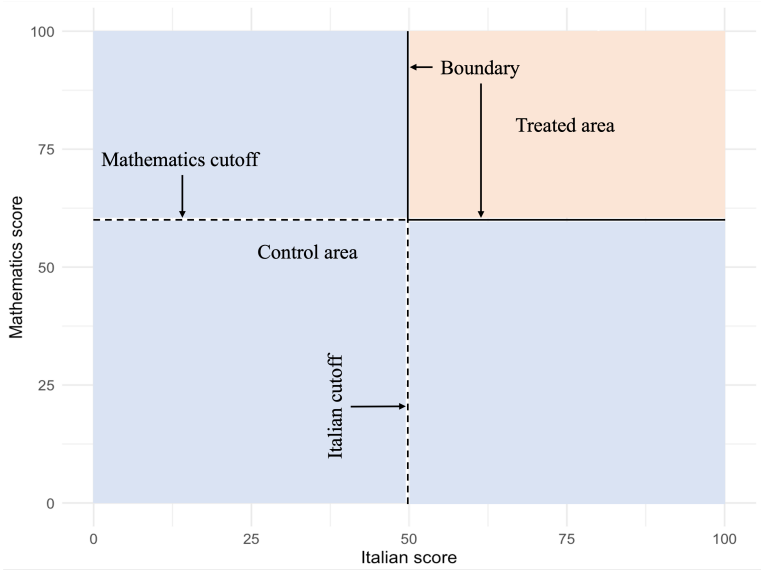


Figure 2.15: Multi-score assignment mechanism

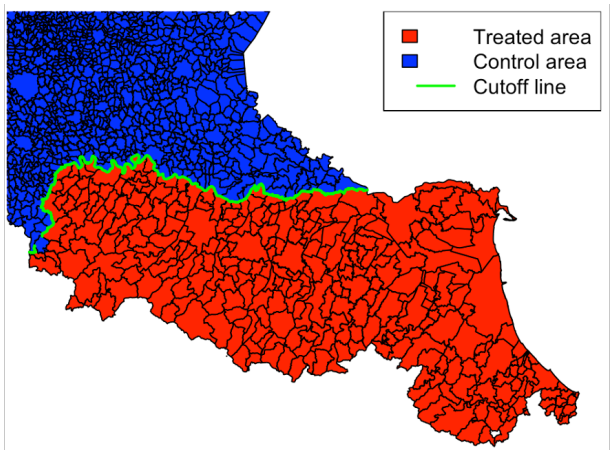


Figure 2.16: Geographic RD: an example with Italian municipalities

sponds to municipalities within the red area, while the control region corresponds to municipalities within the blue area. The cutoff is represented by the administrative boundary between the two regions, marked by the green line. Also, with GRD, the same logic described so far applies: reducing the design to a classic RD with a single score and cutoff by normalizing the running variable or defining specific points where the treatment effect is estimated to assess its heterogeneity. However, when dealing with a GRD design, several challenges may arise that researchers must consider.

The first issue deals with compound treatments: when using a boundary represented, for example, by a state or county line, one must be cautious, as changing regions may involve multiple simultaneous policy changes that could inflate the estimated treatment effect. A second issue concerns the manipulation of the running variable, as in many cases, units may be able to move from the control area to the treatment area or vice versa. For example, consider two regions with very different tax regimes; clearly, the region with lower taxation levels could incentivize people to move to the treatment area. A final issue is related to spillovers/interference: when units are very close to each other, they may influence each other indirectly (Angrist and Lavy, 1999; Forastiere et al., 2006; Papadogeorgou et al., 2019).

Chapter 3

Investigating the effect of economic, social, and cultural conditions on student's performance: insights from INVALSI Tests

Policymakers face a crucial challenge in reducing inequalities across different segments of the population. Education plays a fundamental role in this process, as it has a significant impact on various aspects of individuals' lives, including employment, health, and income. Abdullah et al. (2015) highlight that education can help mitigate income inequalities by redistributing wealth from the rich to the poor. Similarly, Furnée et al. (2008) show a positive relationship between education and health, both directly and indirectly, through factors such as employment, economic conditions, psychosocial resources, and lifestyle choices.

A key determinant of educational outcomes is socioeconomic status (SES). The 1966 Coleman Report (Coleman et al., 1966)—one of the most influential education reports of the last 60 years (Hanushek, 2016)—emphasized for the first time that a student's academic performance is shaped not only by family background but also by the socioeconomic environment of their peers. Monitoring students' competencies is therefore essential. The link between school performance and socioeconomic background has been the subject of investigation for decades. The first major synthesis was the meta-analysis by White (1982), which reviewed nearly 200 studies published before 1980 and showed that the strength of the association depended heavily on how both SES and achievement were measured: under the "standard" definitions of the time, the relationship appeared on average rather weak, although it varied depending on the indicators and metrics used.

Following White, the 1990s produced multiple empirical studies with results that did not always converge. In some cases, the association appeared strong: for instance, in the work of Lamdin (1996) on standardized test performance, where SES entered as a significant factor alongside other school variables, or in Sutton

and Soderstrom (1999), which showed that average school outcomes closely tracked the demographic and socioeconomic characteristics of the catchment area. In other contexts—especially with restricted samples—the effect of SES could weaken to the point of being negligible or non-significant: such as in the studies on "inner-city" adolescents by Ripple and Luthar (2000), where psychosocial and school adaptation factors were more predictive, or in research highlighting the influence of teachers' perceptions on performance, such as Seyfried (1998).

This heterogeneity motivated a second influential synthesis: the meta-analysis by Sirin (2005), which reexamined studies published between 1990 and 2000. The picture that emerges is clearer: the relationship between SES and achievement varies, on average, from moderate to strong. However, its magnitude depends on the unit of analysis (student, school, district), on the scope and source of the SES measure, and on the type of outcome used—reinforcing the importance of measurement choices in shaping conclusions.

A recent study by Tan et al. (2025) provides a third meta-analysis that, unlike the previous two, focuses on the effects of SES at the school level (School SES). It synthesizes the results of 97 studies published between 2000 and 2020, analyzing a total of 480 effect sizes. The findings show that students in high-SES schools benefit from access to high-quality human resources (skilled and motivated principals and teachers) that foster a positive school climate, underscoring that material resources alone are not the key factor.

At the same time, recent research has updated the way disadvantage itself is measured. Today it is more common to use a diverse set of indicators—family income, mother's education, and family structure. By contrast, school SES is typically measured using several key indicators, including parents' education, household income, parents' occupational status, and access to material and cultural resources within the home. School SES is defined as the average SES of its students. In other words, to measure a school's SES, researchers calculate the mean of the socioeconomic data of all its students. Student achievement is more closely related to school SES than individual SES. In other words, the overall well-being and social environment of the school appear to exert a stronger influence on learning outcomes than a student's individual socioeconomic background (OECD, 2016). A particularly telling example is the study by Borman and Dowling (2010), which reexamined Coleman's historical data and estimated that the socio-demographic composition of the school had 1.75 times the impact of students' individual characteristics in explaining differences in academic performance.

It is in this context that the literature on *peer effects* emerges, shifting attention

from individual resource endowment to the role of class and school composition. The central idea is that peers influence outcomes through norms of effort, mutual expectations, behavioral models, and information channels. Formally, the distinction proposed by Manski (1993) separates *endogenous effects* (influence of peers' own outcomes), *contextual effects* (influence of peers' exogenous characteristics, such as parents' education), and *correlated effects* (reflecting shared factors). Sacerdote (2011), instead, defines peer effects as almost any externality in which peers' background, current behavior, or outcomes affect an individual's outcome. However, the peer effects literature in education is rather contradictory and controversial, with studies reporting positive, negative, or, in some cases, null effects (Sacerdote, 2011; Epple and Romano, 2011). Typically, the literature investigates interactions among classmates (De Paola and Scoppa, 2010; Granger-Serrano and Villarraga-Orjuela, 2021; Gu, 2023; Vardardottir, 2013), roommates (Sacerdote, 2001; Zimmerman, 2003), or schoolmates (Schneeweis and Winter-Ebmer, 2007; Zimmer and Toma, 2000).

This third chapter follows this approach and also focuses on interactions between schoolmates to estimates the peer effect of attending a school classified as low-ESCS according to the ministerial threshold, comparing Mathematics outcomes among students.

3.1 The INVALSI survey

Given these premises, it is necessary to assess and measure students' educational achievement to identify potential problems and critical issues at the school level and to understand where and how intervention may be needed. In Italy, this is done through the INVALSI (acronym for "National Institute for the Evaluation of the Education and Training System") tests, which measure in a standardized way the core skills in Italian, Mathematics, and English, and involve all Italian students attending specific grades in primary and secondary school.

INVALSI, established in 1999, is a public research body under Italian law. It initially conducted surveys, collected information, and developed tools and software for analyzing exam results and evaluating staff. Today, under its mandate from public institutions, it carries out multiple functions. Since the 2005–2006 school year, its key task has been to conduct periodic and systematic assessments of Italian students' knowledge and competencies, as well as of the overall quality of the educational provision of schools and vocational training institutions, including in the context of lifelong learning. These assessments are based on national tests administered to

students across various grades, producing results of crucial importance. They are also used to investigate the causes of school failure and dropout, with reference to social context and types of educational provision. Alongside the Italian, Mathematics, and English tests, students also complete an anonymous questionnaire, which collects information on their personal and family background.

3.1.1 A focus on the running variable and the outcome

This chapter analyzes data from the INVALSI tests administered to students in grade 13 (the final year of high school) between March and May of the 2023/2024 school year.

The running variable for treatment assignment is the school's *Economic, Social, and Cultural Status (ESCS)* index[1]. Thanks to the anonymous questionnaire administered to students during the INVALSI tests, it is measured at the individual level and then aggregated to the class and school levels. This index is derived from a principal component analysis based on three factors:

- HISEI: parents' occupational status
- PARED: parents' level of education expressed in years of formal schooling, standardized according to international criteria
- HOMEPOS: ownership of specific material goods serving as proxies for an economically and culturally favorable learning environment

The HISEI factor is composed of two indicators: father's occupational status and mother's occupational status. Initially, parental occupations were classified into 12 specific employment categories. These raw data were then aggregated into 6 distinct groups, ordered along an ascending occupational status scale. This operation produced two ordinal categorical variables (father's occupational level and mother's occupational level), whose combined maximum value was used to define the student's family occupational status indicator (HISEI).

The PARED factor is determined by recording the educational attainment of both parents, which was initially classified into 6 ordered categories. The final family indicator was established by selecting the highest level of education between the two parents. This qualification was then converted into an estimate of years of schooling using the International Standard Classification of Education (ISCED 97) (OECD, 1999). The ISCED 97, adopted by the OECD, UNESCO, and EUROSTAT,

[1] https://www.invalsiopen.it/indicatore-escs-valutazione-equa/

divides educational pathways into seven levels and serves as the basis for creating internationally comparable statistical indicators in the field of education.

Attention is also given to the presence or absence of certain household items. Particular emphasis is placed on HOMEPOS indicator, as the possession of specific material goods is considered by several researchers to be one of the main proxies for measuring family socioeconomic status (OECD, 2007; OECD, 2008). For its construction, the following dichotomous items were considered:

- At home, do you have a quiet place to study?
- At home, do you have a computer you can use for studying?
- At home, do you have a desk for doing homework?
- At home, do you have educational software, e.g., GeoGebra, Matlab, Google Earth, MindMapper, etc.?
- At home, do you have a fixed Internet connection e.g., modem?
- At home, do you have a room of your own?
- At home, do you have books of classical literature, e.g., Dante, Manzoni, Tolstoy, etc.?
- At home, do you have works of art, e.g., paintings, sculptures, etc.?
- At home, do you have technical manuals, e.g., software user guides, etc.?
- At home, do you have a dictionary, in Italian or other languages?
- Approximately how many books are there in your home excluding school textbooks?

These items were then summarized through a scaling procedure based on the Rasch methodology (Campodifiori et al., 2010; Bond and Fox, 2007).

As previously described, the *ESCS* represents the first principal component of a PCA conducted on the variables HISEI, PARED, and HOMEPOS. The evolution of the *ESCS* is monitored by the Italian Ministry of Education to guide policies aimed at reducing socio-economic disparities. In this regard, in 2023 the Ministry of Education and Merit issued Ministerial Decree No. 90 of May 19, 2023, which identifies as socioeconomically disadvantaged those secondary schools with an *ESCS*

below -0.31243[2] (see Section 3.2.2 for details on the decree). This value is used as the external threshold of the running variable that determines treatment assignment: starting from this cutoff, all schools with an *ESCS* below the threshold are classified as *"low" ESCS* schools. Consequently, the treatment group consists of students attending schools identified as low-*ESCS* according to the threshold established by the Ministry of Education and Merit.

The outcome variable is the students' Mathematics test score obtained in the INVALSI tests (*Math score*), estimated using a Rasch model (Rasch, 1966; Rasch, 1980). The Rasch model, which provides so-called "objective measurements," is used to estimate student ability independently of the difficulty of the items administered. In theory, this model eliminates distortions related to variability in subjects or items. In practice, however, a student's response is still influenced by item difficulty, which in turn depends on the student's level of preparation.

Figure 3.1 shows the distribution of *ESCS* by region using violin plots, colored by Geographic area, with boxplots embedded inside. The regions are ordered in ascending order according to the median value. A clear negative trend can be observed for the southern regions, which all cluster at the lower end of the distribution accompanied by greater variability in *ESCS* values—evidence of substantial differences in socioeconomic and cultural background.

Figure 3.2, in turn, presents a scatterplot of average *ESCS* against average *Math score* at the regional level. Regions are color-coded by geographical area. The red vertical and horizontal lines indicate, respectively, the national mean of *ESCS* and the national mean of the *Math score*. Almost all southern regions—except Abruzzo and Molise—fall in the third quadrant[3], with values below the mean on both dimensions. Conversely, the first quadrant contains the remaining regions, all of which have a score above average on both variables. Lazio is a special case, as it is the only central region located in the fourth quadrant, with an average *ESCS* above the mean but an average Mathematics score below the mean. Once again, a clear gradient emerges between the South and Islands and the rest of the country.

3.1.2 The INVALSI data

The original dataset contained 473,990 observations (students) and 36 variables. After an initial phase of variable selection—where those irrelevant to the analysis

[2] https://www.miur.gov.it/documents/20182/7414469/m_pi.AOOGABMI.Registro+Decreti%28R%29.0000090.19-05-2023.pdf/44e239ff-2151-8c6a-3d45-851a38834ded?version=1.0&t=1690386661238

[3] The quadrants are numbered counterclockwise, starting from the top-right one

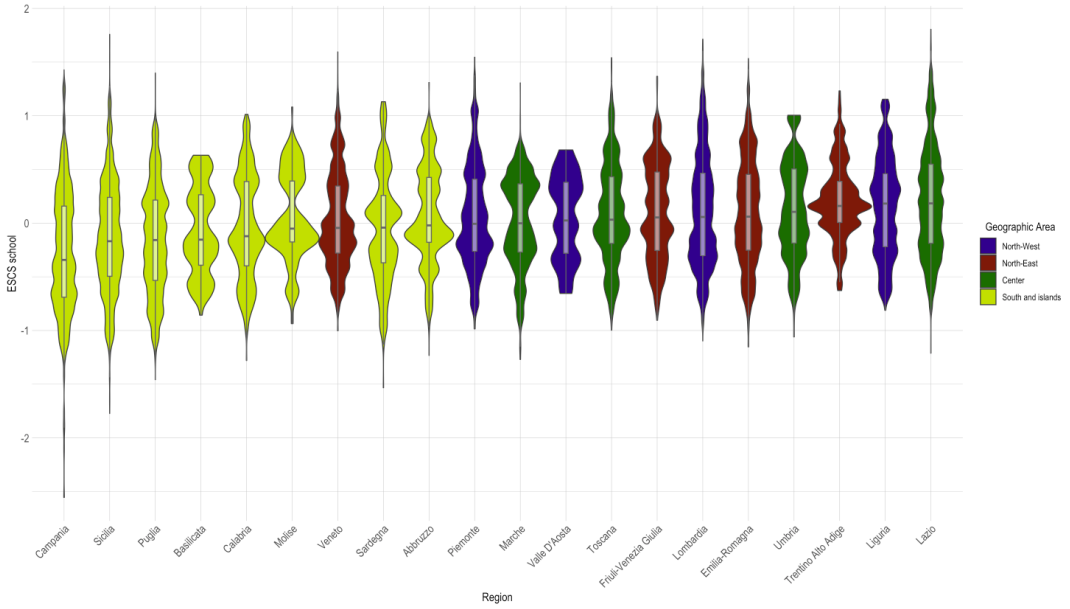


Figure 3.1: Distribution of the ESCS school by Region and Geographic area

Figure 3.2: Average regional ESCS and average Math score: comparison across regions

were removed[4], we retained 13 variables, 10 quantitative and 4 qualitative, which are reported in the first two columns of Tables 3.1 (quantitative variables) and 3.2 (qualitative variables). The variable *N student class was* later created to capture the class size, meaning the number of student in the class attended by each student.

With regard to missing values in the quantitative variables, our approach aimed to minimize information loss. The first step was to remove 784 individuals with *NA* in the treatment assignment variable, namely the school *ESCS*. The decision to exclude these cases stems from the assignment mechanism: since the exact value of the running variable was unknown, it was not possible to determine with certainty whether these students belonged to the treatment or the control group. Next, a further 235 individuals were removed because they had *NA* in all variables.

Before proceeding with imputation, we decided to drop the variables corresponding to written grades in Italian, Mathematics, and English for two reasons: first, because they contained very high proportions of missing values—76.81%, 85.12%, and 81.91% respectively; second, because in the absence of written grades, students' first-term assessment is based only on their oral grade[5]. The missing values for oral grades were then imputed using the mean at the province level, differentiated by school type. After this preprocessing phase, the dataset consists of 472,971 observations and there are no missing values.

Table 3.1 reports the main position indexes of quantitative variables. The table shows that oral grades display similar, though not identical, averages: *Italian* has a mean of 6.90 and a median of 7.00 (CV = 0.16), *English* 6.94 and 7.00 (CV = 0.18), while *Mathematics* is slightly lower, with a mean of 6.47 and a median of 6.29 (CV = 0.21). The fact that the medians for *Italian* and *English* are slightly higher than their means suggests a mild left-skew (a few low grades pulling the mean downward), while in *Mathematics* the greater heterogeneity (higher CV) is consistent with a more dispersed distribution of results.

[4]The variables removed were: i) Grade: school grade; ii) Sidi Invalsi: student's SIDI INVALSI code; iii) Timetable code: student's timetable code; iv) Number of connected PCs: number of PCs available for test administration; v) Number of computer labs: number of computer labs in the test location; vi) ISTAT province code: code of the province to which the school belongs; vii) Month: month of birth; viii) Year: year of birth; ix) Building code: school building code; x) Level: student's performance level in the mathematics test; xi) Written grade Italian: first-term written grade in Italian; xii) Written grade Math: first-term written grade in Mathematics; xiii) Written grade English: first-term written grade in English; xiv) School unit code: code of the school branch; xv) ESCS class: Economic, Social and Cultural background of the class; xvi) ESCS student: Economic, Social and Cultural background of the student; xvii) Place: student's place of birth; xviii) Father's place: father's place of birth; xix) Mother's place: mother's place of birth; xx) Regularity: indicator of regularity with respect to school progression; xxi) Origin: indicator of the student's origin; xxii) Province: province code

[5]`https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2017-04-13;62!vig=`

Table 3.1: A description of the quantitative variables

Variable	Description	Min	Max	Median	Mean	CV
Student code	Code of the student					
Class code	Code of the class					
School code	Code of the school					
Oral grade italian	Oral grade for the first term in Italian	1.00	10.00	7.00	6.90	0.16
Oral grade math	Oral grade for the first term in Math	1.00	10.00	6.29	6.47	0.21
Oral grade english	Oral grade for the first term in English	1.00	10.00	7.00	6.94	0.18
Age	Age of the student	17.00	22.00	19.00	19.18	0.04
ESCS school	Economic social and cultural background of the school	-2.55	1.81	-0.03	0.00	22.43
N student class	Number of student in the class	1.00	42.00	19.00	19.15	0.28
Math score	Grade obtained in the INVALSI mathematics tests	57.99	306.64	191.34	193.01	0.20

Table 3.2: A description of the qualitative variables

Variable	Categories	% Frequency
School type	Scientific lyceum	24.27%
	Other lyceum	29.69%
	Technical institute	31.65%
	Professional institute	14.38%
Gender	Male	50.02%
	Female	49.98%
Geographic area	North-West	23.54%
	North-East	18.03%
	Center	19.83%
	South and islands	38.61%
Region	Valle D'Aosta	0.18%
	Piemonte	6.34%
	Liguria	2.14%
	Lombardia	14.87%
	Veneto	7.72%
	Friuli-Venezia Giulia	1.77%
	Emilia-Romagna	6.96%
	Toscana	5.81%
	Umbria	1.46%
	Marche	2.70%
	Lazio	9.86%
	Abruzzo	2.12%
	Molise	0.50%
	Campania	14.18%
	Puglia	6.78%
	Basilicata	0.98%
	Calabria	3.28%
	Sicilia	8.53%
	Sardegna	2.24%
	Trentino-Alto Adige	1.57%

Age is highly homogeneous, with a mean of 19.18, a median of 19, a range from 17 to 22 years, and a CV of 0.04, indicating very limited variability around the central value. By contrast, class size shows considerable dispersion: mean 19.15, median 19, range 1–42, and CV = 0.28, reflecting the presence of schools with very different class structures.

The *Math score*, which is the outcome variable of our study, has a mean of 193.01 and a median of 191.34, with CV = 0.20: variability is moderate, and the closeness of the mean and median suggests the distribution is not strongly skewed. The *ESCS*, our running variable, spans a wide range (2.55; 1.81) and has a median of 0.03, with a mean close to 0, as expected for a standardized index. In this case, the CV (22.43) is not informative, since the denominator is very close to zero.

Turning to the qualitative variables reported in Table 3.2, we recoded the categories of the variable *Geographic area*, merging "South" and "South and Islands" into a single category. In terms of distribution, the South and Islands account for the largest share of students (38.61%), followed by the North-West (23.54%), the Center (19.83%), and the North-East (18.03%). *Gender* variable is almost perfectly balanced, with 50.02% male and 49.98% female. Regarding the variable *School type*, most students attend a Technical institute (31.65%), followed by Other lyceums (29.69%), Scientific lyceums (24.27%), and Professional institutes (14.38%). Lombardia is the most represented region, with 14.87% of students, followed by Campania (14.18%) and Lazio (9.86%).

3.2 Empirical analysis

From this point forward, we transformed the running variable, the school *ESCS*. Specifically, it was centered at 0 (with 0 corresponding to the threshold value of 0.31243) and then inverted by multiplying it by 1, so that the treatment group appears on the right-hand side and the control group on the left. This operation was carried out purely as a matter of convention.

Before verifying whether the assumptions required to implement the RD design hold (see Section 2.3.3), we first examine our treatment assignment mechanism. As mentioned earlier, treated students are those attending schools with a low *ESCS* according to the threshold set by the Ministry of Education and Merit. Figure 3.3 illustrates the probability of receiving treatment across individuals. As expected, this probability shifts from 0 to 1 at the cutoff, moving from the control group on the left to the treatment group on the right—evidence that we are dealing with a Sharp RD design (see Section 2.3). The treated sample consists of 132,750 students,

while the control group includes 340,221 students.

3.2.1 Checking the assumptions of the Regression Discontinuity

The first assumption to be tested is the *no sorting around the cutoff* assumption. This assumption states that individuals cannot manipulate the mechanism assignment to the treatment or control group. Under ordinary circumstances, this can be checked by applying the McCrary density test (McCrary, 2008). The idea behind the McCrary test is simple: plot the distribution of the running variable and verify whether a discontinuity appears at the cutoff. However, as Figure 3.4 shows, there is a sharp jump at the cutoff, moving from the control group on the left to the treatment group on the right—something that may raise concerns.

However, we are dealing with a special case of RD, namely a case in which the running variable is discrete. In such contexts, the distribution of the running variable will display continuous jumps not necessarily due to manipulation but rather as a consequence of its discrete nature. Indeed, we observe repeated jumps throughout the distribution, not only at the cutoff.

To address this issue, Frandsen (2017) proposes a test specifically designed for discrete running variables. The basic idea is that, in the absence of manipulation, the number of observations exactly at the cutoff should follow a predictable distribution given the overall frequency of the running variable values nearby. In the presence of manipulation, however, the observed number of units at the cutoff would appear anomalous compared to the expected value. The Frandsen test thus makes it possible to distinguish between irregularities due to the discrete nature of the variable and those attributable to strategic behavior by individuals, offering a more appropriate diagnostic tool than the classical McCrary test. Applying the Frandsen test, we obtain a p-value of 0.582. This result provides no statistical evidence of manipulation of the running variable at the cutoff, suggesting that the observed irregularities are attributable to its discrete nature rather than to strategic behavior.

In such cases, beyond the existence of a specific test for discrete running variables, it is important to assess ex ante whether individuals—in our case students or families—could realistically manipulate the running variable. In our study, the running variable is the *ESCS*, an indicator that is neither disclosed nor easily predictable by students or families at the time of school enrollment. It is constructed ex post on the basis of information collected through questionnaires and socioeconomic data, and thus it is something parents or students have access to in advance. Moreover, the threshold determining treatment is extremely precise and narrowly

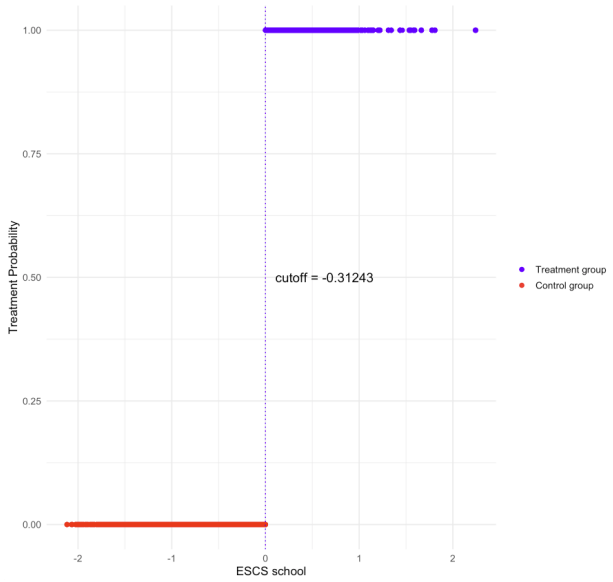


Figure 3.3: Conditional probability of receiving treatment

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted.

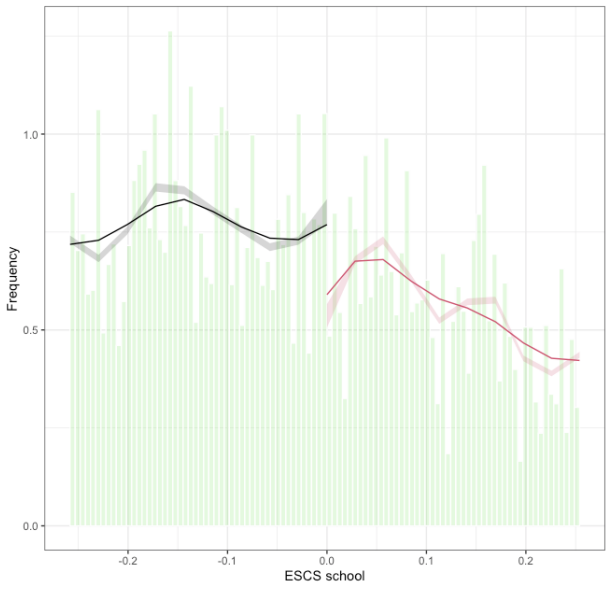


Figure 3.4: McCrary density test: the distribution of the running variable

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted. The black and red curves are local polynomial (smoothed) estimates of the density of the running variable on the left and right of the cutoff. The bands around the lines represent the confidence intervals around the estimated density curves.

defined, making it practically impossible for anyone to "fine-tune" their score. Even if someone were inclined to manipulate their value, the precision of the threshold and the exogenous construction of the index would render such manipulation unfeasible or prohibitively costly. This combination—a non-anticipatory running variable and a sharply defined threshold—strengthens the plausibility of the assumption that systematic manipulation of the running variable is absent. Accordingly, the lack of evidence of mass discontinuity provided by the Frandsen test is consistent with the theoretical argument that the *ESCS* is not subject to strategic behavior.

For the moment, we set aside the discrete nature of the running variable and proceed as in a standard RD design, postponing the discussion and application of the discrete design to Chapter 4.

Another assumption that must be verified concerns *predetermined covariates* and *palcebo outcomes*, i.e., variables that cannot be affected by the treatment and that can be used to test the plausibility of the RD design (see Section 2.2.3 for a discussion of the difference between placebo outcomes and predetermined covariates). It is important to recall that the most straightforward way to conduct an RD analysis is to relate the outcome exclusively to the running variable. Although this basic setup is generally adequate in most cases, some scholars prefer to enrich the specification by including additional covariates beyond the score itself (Cattaneo et al., 2020). The core assumption in a valid RD design is that, at the cutoff, the treatment and control groups should not differ systematically in predetermined covariates, since these cannot have been influenced by treatment assignment. To test this, we examine the null hypothesis that the treatment effect is zero for each predetermined covariate. If significant effects were found for these variables, the credibility of the RD design would be undermined.

The covariates tested for plausibility are reported in Table 3.3. The variables *Age* and *Male* are considered as predetermined covariates. On the other hand, the three oral grades are considered as placebo outcomes. The analyses were carried out in `R` using the package `rdrobust`, which implements local polynomial Regression Discontinuity (RD) estimators with robust bias-corrected confidence intervals and inference procedures developed in Calonico et al. (2014), Calonico et al. (2019), and Calonico et al. (2018). The first column refers to the covariate under consideration. The second column reports the bandwidth obtained using the `mserd` bandwidth selection algorithm implemented in `rdrobust`, which minimizes the MSE and provides a symmetric bandwidth on both sides of the cutoff (see Section 2.2.1 for further details on bandwidth selection in RD). The third column reports the treatment effect estimate obtained from the RD. The fourth column specifies the weighting function

used, namely a triangular kernel that assigns greater weight to observations closer to the cutoff. The fifth and sixth columns report, respectively, the p-value and the robust bias-corrected confidence intervals. Finally, the last two columns show the number of observations to the left and to the right of the cutoff within the selected optimal bandwidth.

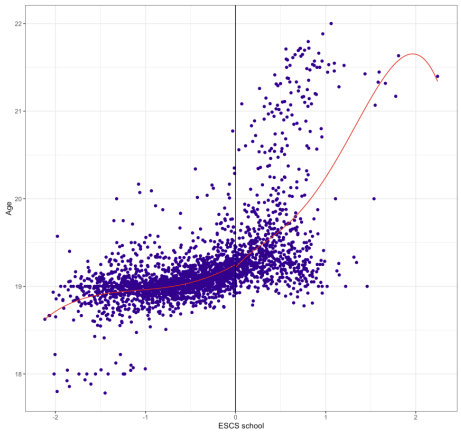
The results suggest good balance for almost all predetermined covariates and placebo outcomes considered. Oral grades in Italian, Mathematics, and English do not show statistically significant discontinuities at the cutoff (p-values of 0.12, 0.11, and 0.35, respectively), and the confidence intervals include zero—consistent with the RD assumption of continuity of potential outcomes with respect to the running variable. Both age and gender show a small negative jump: while statistically significant due to the large sample size, the magnitude is negligible and unlikely to affect the main results. Figure 3.5 displays the typical regression discontinuity plots for each of the considered covariates: on the right is the treatment group (students attending low-*ESCS* schools), while on the left is the control group. Each plot provides a visual sense of the treatment effect: if the two regression curves (shown in red) diverge at the cutoff, a discontinuity is present. The plots show near-linear continuity in almost all cases, with the exception of the variable *Male*, which displays a slight downward discontinuity, consistent with the results already reported in Table 3.3. Both the formal analysis and the graphical analysis indicate that the students right above and below the cutoff are similar in terms of their high school performance.

Variable	**MSE-Optimal Bandwidth**	**RD Estimator**	**Kernel**	**Robust Inference**		**N of Obs**	
				p-value	*95% CI*	*left*	*right*
Age	0.188	−0.024	Triangular	0.01	[−0.042, −0.006]	70,719	54,573
Male	0.084	−0.076	Triangular	0.00	[−0.097, −0.064]	29,667	27,022
Oral grade italian	0.171	−0.021	Triangular	0.12	[−0.048, 0.006]	63,089	50,778
Oral grade math	0.169	−0.031	Triangular	0.11	[−0.055, 0.004]	62,730	49,932
Oral grade english	0.266	0.017	Triangular	0.35	[−0.015, 0.042]	96,565	69,890

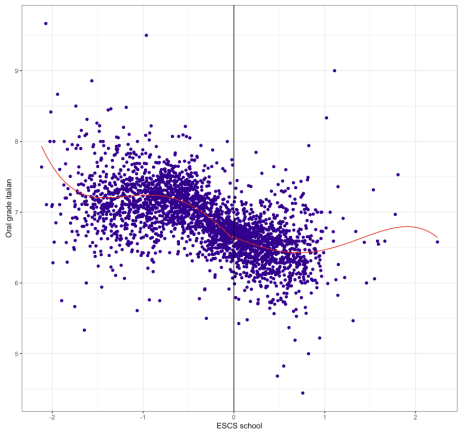
Table 3.3: RD effects on predetermined covariates and placebo outcomes

3.2.2 Results: the causal impact of school ESCS

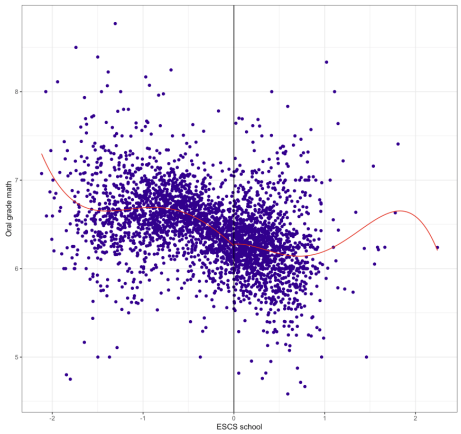
Before presenting the main effect of interest—namely, the impact of attending low-*ESCS* schools on *Math score*—it is important to clarify a key point concerning Ministerial Decree No. 90 of May 19, 2023. The decree establishes that, for upper



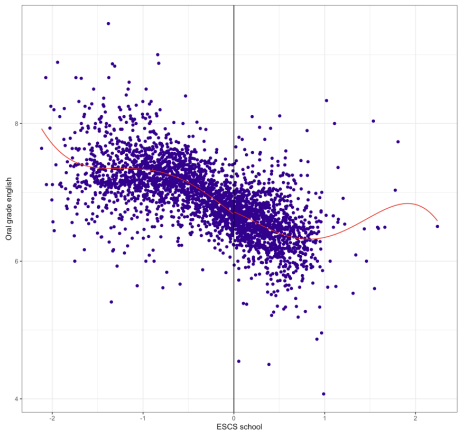
(a) RD Plot Age-ESCS school



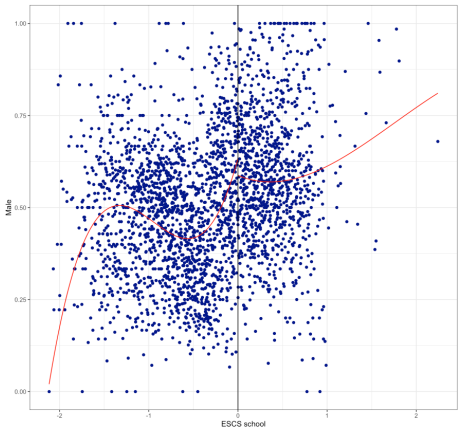
(b) RD Plot Oral grade italian-ESCS school



(c) RD Plot Oral grade math-ESCS school



(d) RD Plot Oral grade english-ESCS school



(e) RD Plot Male-ESCS school

Figure 3.5: RD Plot predetermined covariates-ESCS school

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted.

secondary schools, classes may be formed with no more than 27 students if the following conditions are met:

- The *ESCS* is below the threshold of −0.31243;
- The school dropout indicator[6] is above the threshold of 1.33396;
- The depopulation indicator[7] is below −4.88110.

In our dataset, however, we only have information regarding the *ESCS*. Moreover, we conducted an analysis to show that, on both sides of the *ESCS* threshold, the proportion of classes with fewer than 27 students is essentially the same, indicating that no policy intervention is confounding our estimated effect. To support this claim, Table 3.4 reports an analysis grouping students by the variable *class code*. Overall, there are 27,267 classes, of which 8,554 belong to the treatment group and 18,713 to the control group. As shown in the third and fourth columns, the percentage of classes with more than 27 students and the percentage of classes with 27 students or fewer are very similar across the two groups. To further strengthen our analysis, we use the variable *N student class* as a placebo outcome to test whether there is any discontinuity at the cutoff attributable to policy implementation. Figure 3.6 presents a standard RD plot of the placebo outcome *N student class* against the running variable *school ESCS*. The figure clearly shows continuity on both sides of the cutoff: there is no jump when moving from the control group on the left to the treatment group on the right. This means that, at the threshold, there are no policies in place that could confound the treatment effect of interest.

Once all the assumptions were verified, the RD was then estimated. Figure 3.7 shows the RD plot between the outcome of interest—the *Math score*—and the running variable, *ESCS school*. A clear discontinuity can be observed at the cutoff: moving from the control group on the left (students attending schools with a nonlow ESCS) to the treatment group on the right (students attending schools with a

[6]Measures the share of young people who abandon education or fail to reach expected competencies

[7]Measures the decline of population in a given area and the intensity of demographic outflow

Table 3.4: Placebo outcome: an analysis of the number of student in each class

	Number of classes	Classes with more than 27 students	Classes with 27 students or less
Treatment	8,554 (31.37%)	324 (3.78%)	8,230 (96.22%)
Control	18,713 (68.63%)	469 (2.50%)	18,224 (97.50%)

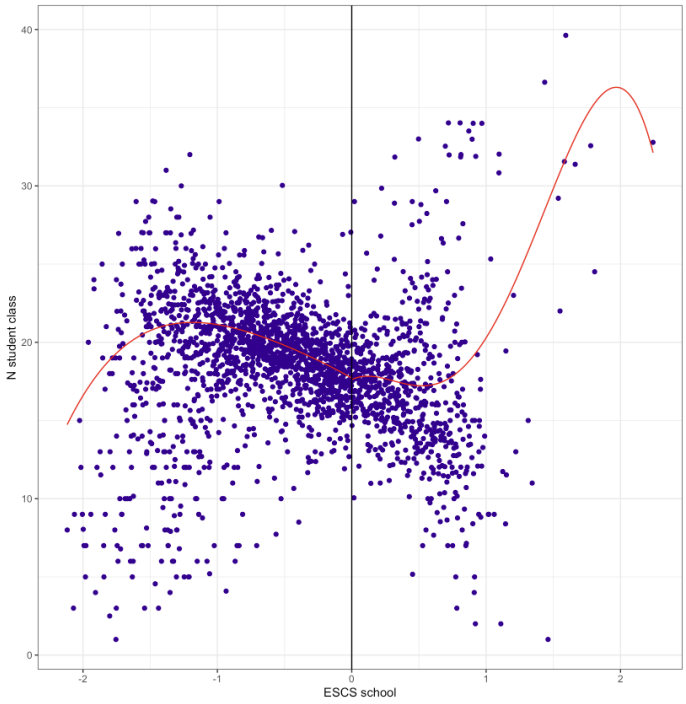


Figure 3.6: RD Plot placebo outcomes: N student class-ESCS school

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted.

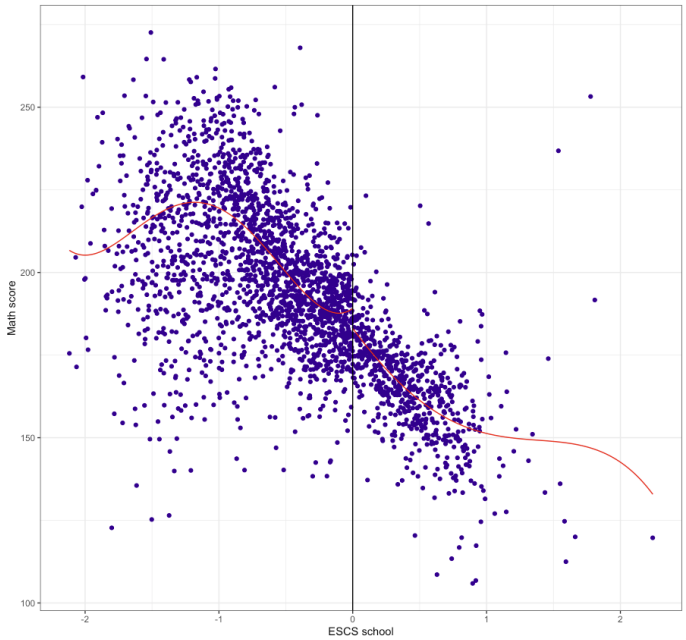


Figure 3.7: RD Plot LATE: Math score-ESCS school

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted.

low ESCS). Table 3.5 reports the magnitude of the LATE. Once again, the optimal bandwidth was selected using the `mserd` function from the `rdrobust` package, which chooses the bandwidth that minimizes the MSE while accounting for the bias–variance trade-off (see Section 2.2.1). As weighting function, we used a triangular kernel, since when combined with the MSE-optimal bandwidth, it yields to a point estimator with optimal properties (Cattaneo et al., 2020). To estimate the causal effect, we employed a local polynomial of order 1, the most commonly used specification in the literature (see Section 2.1 for a discussion of local polynomial order).

With a bandwidth width of only 0.131 on either side of the cutoff and significance at the 1% level, the treatment effect is negative and equal to -1.071, indicating that attending a low-*ESCS* school has adverse consequences for *Math score*. However, considering that the outcome ranges from 57.99 to 306.64, this effect is very small and can be regarded as practically negligible.

Table 3.5: RD effect on the outcome Math score

Variable	**MSE-Optimal Bandwidth**	**RD Estimator**	**Kernel**	**Robust Inference**		**N of Obs**	
				p-value	*95% CI*	*left*	*right*
Math score	0.131	-1.071	Triangular	0.00	$[-2.115, -0.338]$	47,258	38,699

Faced with a negative but modest LATE, the most cautious interpretation is that the socioeconomic composition of the school environment—and thus the potential mechanisms of peer interaction—constitutes a real channel through which educational inequalities are transmitted, but not a dominant one in explaining mathematics scores or students' academic performance more generally. Simply attending a low-*ESCS* school generates a statistically detectable effect, yet one that is economically limited. This implies that the most effective levers for reducing learning gaps in mathematics must combine the management of school composition with targeted instructional interventions and broader school- or system-level policies. In other words, changing the social composition of a school alone is not sufficient. This result is consistent with the literature on peer effects, which suggests that classroom context can influence performance through norms of effort, shared expectations, or behavioural models. However, in the Italian case analysed here, such effects appear to operate in a real but not dominant manner: socioeconomic composition alone is not enough to produce large differences in outcomes.

Although small, the negative result we observe supports the idea that peer influence is exerted primarily through the school climate, expectations, and daily

management of learning activities. In short, school composition matters, but what is crucial is the quality of teaching and the organization of learning in the most vulnerable contexts. To achieve substantial improvements in Mathematics, strategies must therefore move beyond the belief that a slight shift in class composition is enough and instead focus on integrated interventions. These include: intensive, data-driven instructional support, professional training, attention to school climate and prevention of behavioral disruptions, and the allocation of stable, planned additional funding.

It is important, however, to acknowledge that the analysis conducted thus far is subject to two limitations. The first one is that, as previously noted, we are neglecting the discrete nature of the running variable. The RD design with a discrete running variable is applied in Chapter 4. The second limitation concerns the heterogeneity of effects across students, schools, and geographical contexts—something that a LATE at the threshold cannot fully capture. Proper assessment requires subgroup analyses and alternative specifications, which is addressed in Chapter 5 by incorporating the regional component into the analysis.

Chapter 4

Regression Discontinuity design with discrete running variable

In the previous chapters, the RD design was introduced under the assumption that the running variable was continuous. In reality, in many applications, the nature of the running variable is discrete. In these cases, the implementation of RD encounters some obstacles. Discrete random variables take on a finite number of values:

$$X \in \{x_0, x_1, \ldots, x_k\} \quad \text{with} \quad \mathbb{P}[X = x_k] > 0, \forall k$$

These values can be, for example, 1, 2, 3, or not necessarily integers but still finite numbers, where each value has a non-zero probability (there is some probability that x equals 0, some probability that it equals 1.5, and so on). These values from x_0 to x_k are generally defined as *"mass points"* (Cattaneo et al., 2024).

There are two characteristics that distinguish the RD design with a discrete running variable. The first is that many units may have the same score on the running variable; the second is that there may exist units located exactly at the cutoff point:

$$\mathbb{P}[X_i = c] > 0 \quad \text{with} \quad i = 1, ..., n$$

This second characteristic represents a problem, at least within the "continuity-based framework," which is the standard framework with the weakest assumptions. In fact, as the name suggests, it requires continuity of the running variable, something that is absent when the running variable is discrete. This is problematic because, as mentioned, in RD the goal is to get as close as possible to the cutoff point and compare those units. However, if the running variable is discrete and takes values such as $-2, -1, 0, 1, 2$, it is not possible to get arbitrarily close to the cutoff, even with a very large sample. Indeed, in this case, if the cutoff were 0, from the right we could observe units exactly at the cutoff; but from the left, the closest point to the cutoff would be -1, and we could not get any closer.

From a theoretical point of view, in such cases one could not strictly rely on RD. However, from a practical perspective, even in the presence of mass points, RD

still represents a good approximation. In this regard, it is important to distinguish between the 3 possible cases in which mass points may arise:

1. *Heaping*: in this case we have a variable that appears to be continuous but has mass points on top of it. Some examples of this situation are test scores or dates of birth.

2. *Rounding*: this means that there is an underlying continuous running variable that we do not observe, but instead we observe a discretized version of it. A clear example is age, which is in itself a continuous variable but is very often observed in years or months (as if rounded up or down). Another example is income, which is also a continuous variable but is often categorized (income brackets).

3. *Discrete running variable*: a case in which the variables are intrinsically discrete by their nature. Examples include the number of seats in the Senate or the number of employees in a firm (consider, for instance, a program implemented for firms with fewer than 50 employees; this variable will naturally take discrete values such as 1, 2, 3, etc.).

Figure 4.1 presents an example of heaping, created with simulated data, designed to have a discrete running variable and an increasing outcome, with a jump at the cutoff. On the left side of the figure, the histogram of the running variable is shown, while on the right side, a scatterplot of the running variable against the outcome is displayed. The running variable represented is essentially continuous, in the sense that it takes values across the entire range of the histogram, but distinct spikes appear where certain values occur more frequently than others. The right panel of Figure 4.1 shows the scatter plot of the running variable against the outcome, where the presence of repeated observations is evident in the form of stacked bars. The case of rounding is illustrated in Figure 4.2, where the running variable may take only integer values. Here too, as in Figure 4.1, we generated simulated data designed to have a discrete running variable and an increasing outcome, with a jump at the cutoff. In this case, the histogram appears as shown in the left panel, with no observations between adjacent values. The right panel of Figure 4.2 displays a scatter plot of the running variable against the outcome. Conceptually, one might imagine that there are underlying values (represented by the gray points) that could in principle be observed, but in practice are not, because the variable has been rounded up or down (as in the case of age, which is typically reported in years or months). Thus, the gray points are unobserved, while only the blue points appear in

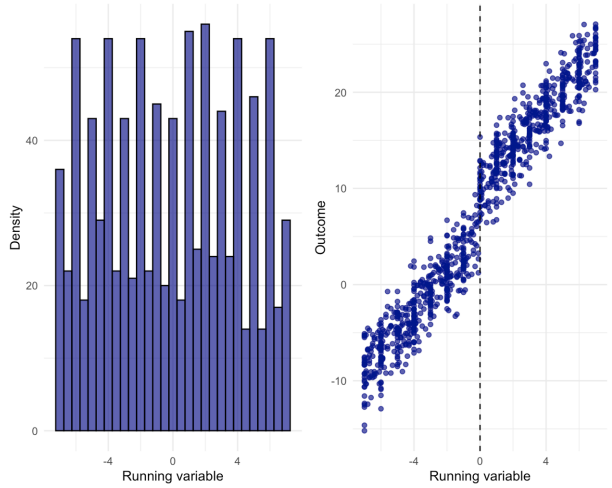


Figure 4.1: Discrete running variable: Heaping case

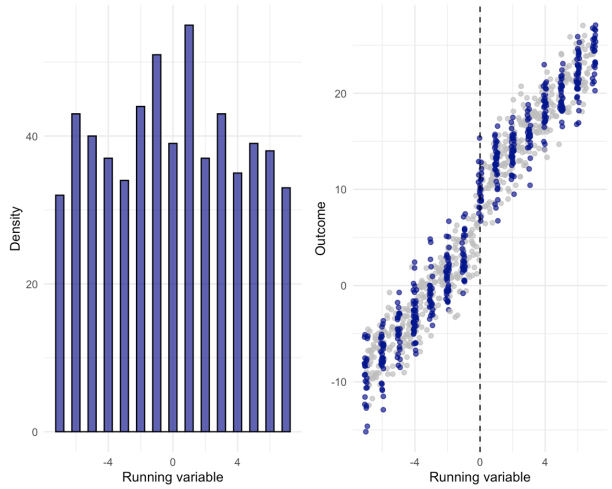


Figure 4.2: Discrete running variable: Rounding case

the data. Focusing on the scatter plot, it is possible on the right side to observe units exactly at the cutoff, providing information on treated units at that point. However, when attempting to compare them with untreated units, the closest observations are at -1, since no units exist between -1 and 0.

The third and final case, that of a discrete running variable, is not shown graphically because it closely resembles the rounding case, with the only difference being that the gray points in the scatter plot are absent altogether, as they do not exist.

In the case of heaping, continuity-based methods may still provide a reasonable approximation. As noted earlier, from a theoretical standpoint RD should not (or could not) be applied, since the running variable is not strictly continuous. Nevertheless, it remains a useful approximation because it still allows researchers to get very close to the cutoff. However, in the other two cases, using RD is much more complex. A good solution could be to use the Local randomization approach (Cattaneo et al., 2024; Cattaneo et al., 2016; Cattaneo et al., 2018) (see Section 2.4 for a detailed discussion on the Local randomization approach).

4.1 The mass points problem

Several issues must be considered when using RD with a discrete running variable. First, the presence of mass points can reduce the effective sample size. In fact, Local polynomial methods require variation in the values of the running variable in order to work: it must be possible to construct functions to the right and to the left of the cutoff, and to do so, the outcome must vary with changes in the running variable. Without such variation, estimation would not be possible. Figure 4.3 provides a graphical illustration of this issue. The figure displays three scatter plots of three different running variables against three different outcomes: the right panel was obtained by generating a single discrete value of the running variable associated with multiple outcome values; the central panel features a running variable that takes on only two discrete values with multiple outcome values different from the previous outcome; the third panel consists of a continuous random running variable and a continuous random outcome. In the right-hand panel, it is shown that one cannot estimate the relationship between the outcome and the running variable because there is no clear way to fit a line. In the center panel, even though it is artificially possible to fit a line, the information provided about how the outcome varies with the running variable is very limited, since there are only two reference points. The ideal scenario is shown in the left-hand panel, where there is a consistent variation in the running variable, making it possible to estimate the line more accurately.

What matters, therefore, is not the total number of observations, but the number of distinct values of the running variable. Adding a very large number of observations that all take only two distinct values of the running variable would contribute little information. This means that Local polynomial methods behave as if the effective sample size were equal to the number of mass points, not the total number of observations, because the mass points provide information about how the outcome changes with the running variable. Thus, in the presence of mass points, the sample size is effectively "artificially inflated," since one may have a very large number of observations but only a small number of distinct running variable values. For example, one might have a sample of 300,000 observations, but if the running variable takes only 1,000 distinct values, then the effective sample size is 1,000, not 300,000. This phenomenon is known as *"sample size inflation"*. One possible solution in such cases is to aggregate the data at the level of the mass points, taking the mean outcome for all observations that share the same value of the running variable (Cattaneo et al., 2024). This approach avoids sample size inflation, producing a new sample whose number of observations equals the number of mass points rather than the total number of raw observations.

In the case of rounding, where the issue is that some units are observed directly at the cutoff point (Figure 4.2), the treatment effect discussed so far can no longer be identified—at least not in a nonparametric way. Recall that global polynomial methods can behave very irregularly, especially near the cutoff. Thus, the effect is no longer nonparametrically identifiable because it is not possible to get arbitrarily close to the cutoff. The best approach in such situations is to rely on parametric assumptions (Dong, 2015). However, this requires specifying assumptions about the true model, which is complicated because even if the correct model were known, the resulting estimates might still be inconsistent.

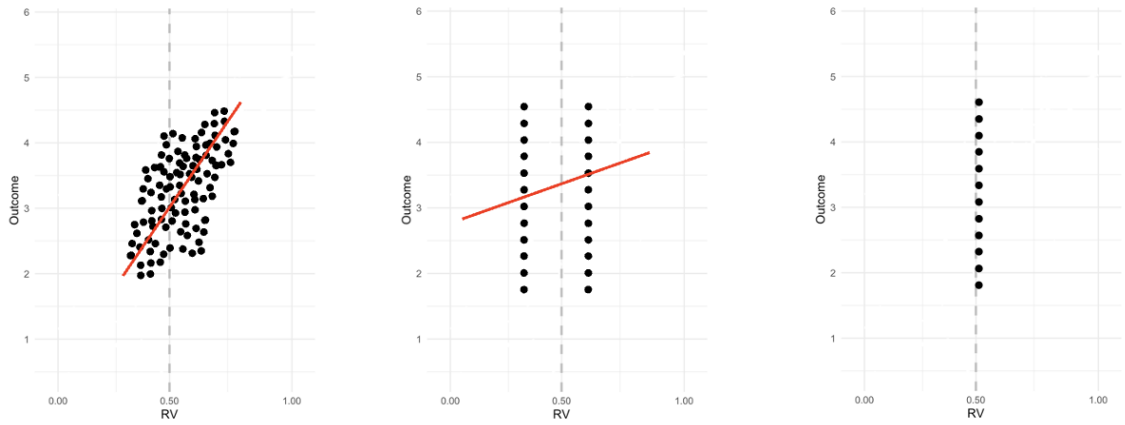


Figure 4.3: Estimation problem with the presence of mass points

In the case of a discrete running variable (the third case), where the running variable is discrete by its very nature, continuity-based methods fail not only because it is impossible to approach the cutoff arbitrarily closely, but also because the approach is conceptually invalid. For example, if the running variable is the number of workers, it makes no sense to consider 49.9 or 50.1 workers. The best that can be done in such settings is to compare units exactly at the cutoff (say, 0) with units at -1 (the closest point to the cutoff from the left), since it is impossible to get any closer. Therefore, Local polynomial methods are meaningless here, as well as iterative bandwidth-selection algorithms that minimize MSE. In such cases, the most appropriate approach is to rely on the Local randomization framework.

In summary, when working with a discrete running variable, it is important to analyze the mass points, assess how many there are, and examine their distribution. If there are many mass points and, in particular, if they are sufficiently close to the cutoff, continuity-based methods still provide a reasonable approximation. Care must be taken in defining the effective sample size because, as discussed, the true sample size is determined not by the total number of observations but by the number of mass points. One possible strategy is to aggregate the data at the level of the mass points and then carry out the analysis on these aggregated data. If there are only a few mass points and they are far apart—especially far from the cutoff—continuity-based methods will not work, either practically or conceptually. In such cases, the local randomization approach may be more appropriate from a conceptual standpoint.

4.2 A solution to the mass points problem

As mentioned in the previous section, a possible solution to the problems arising from the presence of mass points is to collapsing the data using the mean. This means that, within each discrete value of the running variable shared by a certain number of observations, that discrete value is summarized by taking a simple average of the outcome. Table 4.1 illustrates an example of how collapsing to the mean works, assuming there are only two mass points (-2.00 and -1.98) and ten individuals. The first column refers to the individual, the second to the value of the running variable, the third to the corresponding outcome value, and the fourth to the collapsed outcome for the first six individuals who share the same discrete value of the running variable, and for the subsequent four.

However, in many cases, summarizing a group of individuals who share the same value of the running variable using the simple mean may yield an incomplete picture

Table 4.1: Collapsing on the mean: an example with ten individuals and two mass points

Individual	RV	Y	Y collapsed on mean
1	-2.00	2.54	9.42
2	-2.00	4.87	
3	-2.00	6.12	
4	-2.00	8.91	
5	-2.00	1.34	
6	-2.00	32.76	
7	-1.98	3.18	5.09
8	-1.98	5.72	
9	-1.98	8.45	
10	-1.98	2.99	

of the complex structure of the outcome, masking the presence of outliers, skewed distributions, or heavy tails. In such circumstances, the mean tends to be strongly influenced by a few extreme values, producing biased or overly variable estimates and reducing the ability to capture the true heterogeneity of the group. This limitation becomes particularly relevant when the sample size is small, as the efficiency of the mean comes at the cost of marked vulnerability to the shape of the outcome distribution.

For this reason, we propose two alternatives to the simple mean that aim to be more robust in the presence of outliers within mass points: the median and the medoids (Kaufman and Rousseeuw, 1990). The use of robust measures such as the median or the medoid allows for a more faithful representation of the underlying data structure, especially in the presence of skewed distributions or anomalous values. The median, being insensitive to outliers, provides a stable representation of the central tendency of the group, reducing the distortion that the mean would suffer under strong skewness. The medoid, in turn, offers an additional advantage: it identifies an actually observed and representative unit of the group, thereby preserving the discrete nature of the data and limiting distortions caused by extreme values. Both methods therefore yield more compact estimates that are less sensitive to the shape of the distribution.

The *Partitioning Around Medoids* (PAM), also known as *K-Medoids*, was first introduced by Kaufman and Rousseeuw (1990). Like K-Means, it is a partitioning clustering method, but instead of focusing on centroids, it focuses on medoids. The method seeks k representative objects, called medoids, that minimize the average dissimilarity of all objects in the dataset relative to their nearest medoid. As Kauf-

man and Rousseeuw themselves noted, the idea of using a representative object for cluster analysis had previously been discussed by other researchers in related contexts, such as Vinod (1969), Rao (1971), and Mulvey and Crowder (1979). Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n points in a space with a distance function d. A medoid is defined as:

$$x_{\text{medoid}} = \arg\min_{y \in \mathcal{X}} \sum_{i=1}^{n} d(y, x_i).$$

Thus, medoids represent those objects within a cluster for which the average dissimilarity with respect to all other objects is minimal, and they are always constrained to belong to the dataset itself. Our idea is to treat each group of mass points as a separate cluster and to extract from it the individual that best represents the group through the use of medoids. One of the main advantages of the medoids is its low susceptibility to outliers (even lower than that of the median). On the other hand, one of its main disadvantages—typical of non-hierarchical clustering methods—is the selection of the initial clusters. However, by considering each mass point as a distinct cluster, we overcome this drawback of the medoid approach.

Table 4.2 includes two additional columns that highlight the differences between the use of the mean, medoids, and the median. We can observe that for the first

Table 4.2: Collapsing on the mean: an example with ten individuals and two mass points

Individual	RV	Y	Y collapsed on mean	Y collapsed on medoid	Y collapsed on median
1	-2.00	2.54	9.42	6.12	5.50
2	-2.00	4.87			
3	-2.00	6.12			
4	-2.00	8.91			
5	-2.00	1.34			
6	-2.00	32.76			
7	-1.98	3.18	5.09	5.72	4.45
8	-1.98	5.72			
9	-1.98	8.45			
10	-1.98	2.99			

group of observations (running variable/mass point value $= -2.00$), where an outlier is present (32.76, individual 6), the mean provides a much more distorted summary measure, as it is less stable in the presence of skewness. By contrast, the summaries provided by the medoid and the median remain more faithful and representative

of the group's characteristics. The difference between the two lies in the fact that, for the medoid, the value corresponds to an observation already present in the data distribution (individual 3).

Turning to the other group of individuals (running variable/mass point value $= -1.98$), the three methods yield very similar summaries due to the absence of outliers.

4.3 Simulation study

To assess the impact of mass points in RD designs with a discrete running variable, we developed a Monte Carlo simulation study. The goal is not only to compare different strategies for collapsing the data at the mass points—aggregating by the mean outcome, selecting a representative individual via the medoid, and aggregating by the median—but, above all, to provide guidance on how to choose the most appropriate method given the structure of the data. In particular, the analysis focuses on the sensitivity of these approaches to *sample size*, the *number of mass points*, and the *shape of the outcome distribution*. The strategies consist of simulating as many observations as there are mass points. In the first case we take the average of the outcome; in the second we use the medoid—that is, a representative observation of the dataset or of clusters (in our case, of each mass point); and in the third we use the median of the outcome instead of the mean. The latter two methods offer greater robustness to outliers. In each simulation, the running variable X was generated as a discrete vector on m grid of equally spaced values in the interval $[-2, 2]$:

$$X_i \in \{-2, ..., 2\}, \quad i = 1, ..., m$$

where m corresponds to the number of distinct discrete values assumed by the running variable. In general, four sample sizes were considered:

$$n \in \{5{,}000, 10{,}000, 15{,}000, 20{,}000\}.$$

The equations of our simulation study are defined as:

$$Y_{0i} = 0.3X_i + 0.2X_i^2 + \epsilon_i \tag{4.2}$$

$$Y_{1i} = Y_{0i} + \tau \cdot D_i \tag{4.3}$$

$$\epsilon_i \sim N(0, 1). \tag{4.4}$$

Where Y_{0i} denotes the outcome in the absence of treatment, Y_{1i} represents the outcome under treatment and the parameter τ corresponds to a treatment effect fixed a priori equal to 0.5. To assess the robustness of our approach, we considered two different distribution of the outcome within each mass point:

- Normal: symmetric, with skewness and kurtosis equal to 0
- Exponential: highly asymmetric, with skewness equal to 2 and kurtosis equal to 6, so as to emphasize the effect of outliers on mean-based estimates.

The density of the mass points was controlled by fixing three percentage levels relative to the total sample size:

$$Mass \quad points \in \{1\%, 1.5\%, 2\%\}$$

Table 4.3: Factors and levels of the simulation plan

Factors	**Levels**
n	5,000, 10,000, 15,000, 20,000
mass points	1%, 1.5%, 2%
Distribution of Y	Normal, Skewed

In this way, for each sample size we generated datasets in which the discrete values of the running variable appear with different concentrations. Combining the different factors of the study with the different levels, a total of 24 distinct scenarios were simulated. An overview of how many factors and how many levels are taken into account in the study is provided in Table 4.3, while an overview of the 24 scenarios is provided in Table 4.4. Furthermore, to ensure robustness of the results, we generated 500 samples for each scenario. Thus, the experimental design spans multiple dimensions, varying the sample size, the density of the mass points, and the distribution of the outcome, in order to explore scenarios with different characteristics in terms of symmetry and presence of outliers.

4.3.1 Simulation results

Figures 4.4 and 4.5 show the distribution of the estimated coefficients, summarized with density plots in which each curve is color-coded by the method used. Specifically, Figure 4.4 displays the results for a normal distribution of the outcome at each mass point, whereas Figure 4.5 refers to the case of a skewed distribution. In each

Table 4.4: Simulation study plan

Scenarios	**n**	**mass points**	**Distribution of Y**
1	5,000	50 (1%)	Normal
2	5,000	50 (1%)	Skewed
3	5,000	75 (1.5%)	Normal
4	5,000	75 (1.5%)	Skewed
5	5,000	100 (2%)	Normal
6	5,000	100 (2%)	Skewed
7	10,000	100 (1%)	Normal
8	10,000	100 (1%)	Skewed
9	10,000	150 (1.5%)	Normal
10	10,000	150 (1.5%)	Skewed
11	10,000	200 (2%)	Normal
12	10,000	200 (2%)	Skewed
13	15,000	150 (1%)	Normal
14	15,000	150 (1%)	Skewed
15	15,000	225 (1.5%)	Normal
16	15,000	225 (1.5%)	Skewed
17	15,000	300 (2%)	Normal
18	15,000	300 (2%)	Skewed
19	20,000	200 (1%)	Normal
20	20,000	200 (1%)	Skewed
21	20,000	300 (1.5%)	Normal
22	20,000	300 (1.5%)	Skewed
23	20,000	400 (2%)	Normal
24	20,000	400 (2%)	Skewed

figure, rows illustrate the effect of sample size, while columns show the different levels of the percentage of mass points considered. Note that the treatment effect was set to 0.5. The density plots in Figure 4.4 show that the mean, the medoid, and the median all yield essentially correct estimates: the distributions of the estimated coefficients are centered around 0.5, indicating no systematic bias. The key difference concerns dispersion. For small samples ($n = 5{,}000$), the density for the mean is more concentrated (higher peaks and shorter tails), whereas those for the medoid and the median are wider with more pronounced tails. In other words, for the same level of accuracy, the mean is more efficient (lower variance), while the medoid and the median incur a cost in terms of variability. This efficiency difference appears at all three levels of mass-point concentration ($1\%, 1.5\%, 2\%$). However, as the percentage of mass points increases, the variability differences between the approaches reduces. In fact, with $n = 5{,}000$ and 2% mass points, the three curves are more similar than

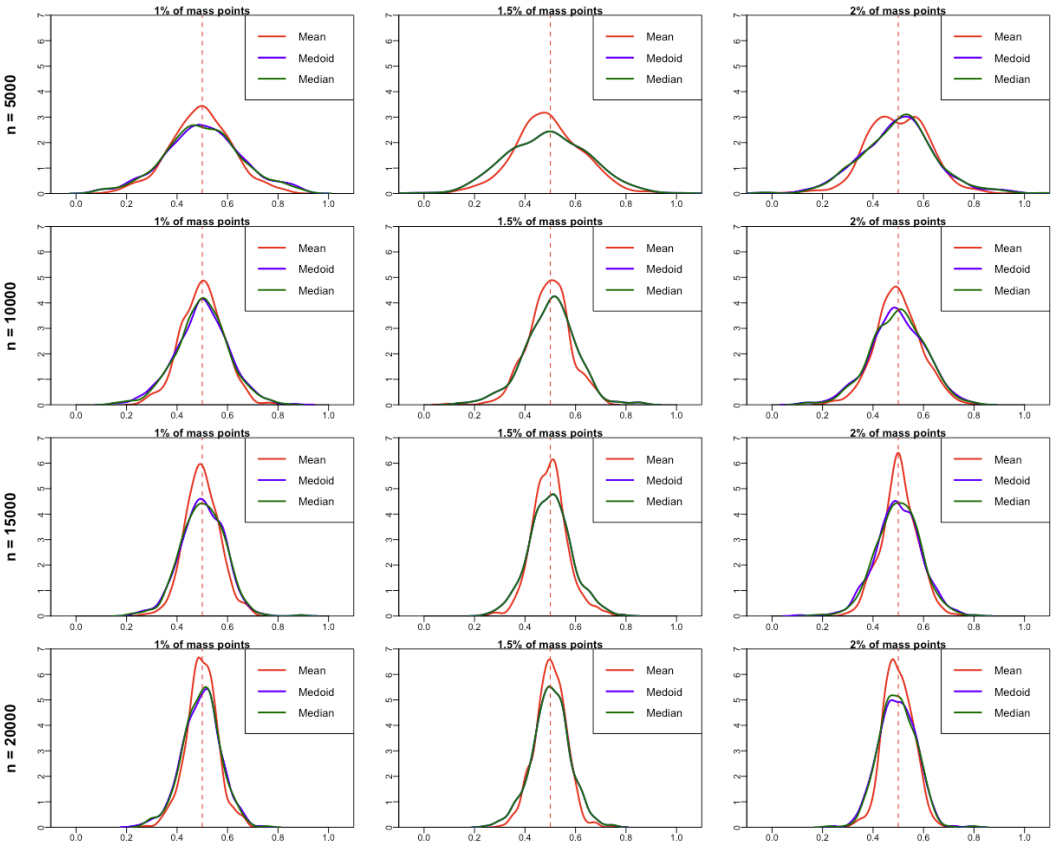


Figure 4.4: Estimated coefficient distribution: scenarios with normal outcome at each mass point

those with 1% or 1.5% of mass points. Moving to n = 10,000, we see a uniform contraction of all the densities: the peaks become sharper and the tails pull in, but the mean retains an advantage in precision (a narrower curve) compared with the medoid and the median, which continue to show slightly greater dispersion. The same pattern holds for n = 15,000 and n = 20,000: the densities progressively converge, with only marginal differences. Results provided in Figure 4.4 suggest that, under symmetry, increasing the sample size makes the approaches nearly equivalent in terms of the distribution of the estimated coefficients. Sensitivity to the percentage of mass points also remains limited across all panels: with 2% of mass points the densities are slightly "fatter" at small n, but the effect fades quickly as n increases. In summary, under normality: (i) greater efficiency of the mean for small or moderate samples (more concentrated density); (ii) differences that shrink as n grows, almost disappearing; (iii) a modest impact of the share of mass points, which does not change the ranking of methods but may slightly reduce—at small n—the higher variability of the medoid and the median as its percentage increases.

In the density plots of the estimated coefficients for the skewed distribution in Figure 4.5, marked differences emerge among the three approaches. With small samples (n = 5,000), the distribution of mean-based estimates is wider and more dispersed, with long tails and a less pronounced peak; this reflects the mean's greater sensitivity to outliers and to the long tail typical of the exponential distribution. By contrast, both the median and the medoid show curves more concentrated around 0.5, with sharper peaks and shorter tails, indicating that these robust approaches succeed in limiting the variability induced by asymmetry. This difference is already evident at a low percentage of mass points (1%) and becomes even more pronounced at 2%, when the mean further increases its dispersion while the median and the medoid maintain stable, compact distributions. As the sample size increases (n = 10,000 and n = 15,000), the three densities tend to tighten, but the relative advantage of the robust approaches persists: the mean remains more exposed to tails and variability, whereas the median and the medoid continue to produce more concentrated estimates, with higher peaks and estimates symmetrically distributed around 0.5 (τ). Only with very large samples (n = 20,000) a partial convergence is observed: the differences between the mean and the robust methods diminish, but the superiority of the median and the medoid in terms of stability and precision remains visible, especially when the share of mass points is high.

Overall, these results indicate that in asymmetric scenarios the mean suffers a systematic loss of efficiency, amplified in small samples and when the percentage of mass points is higher. By contrast, the median and the medoid offer a clear advan-

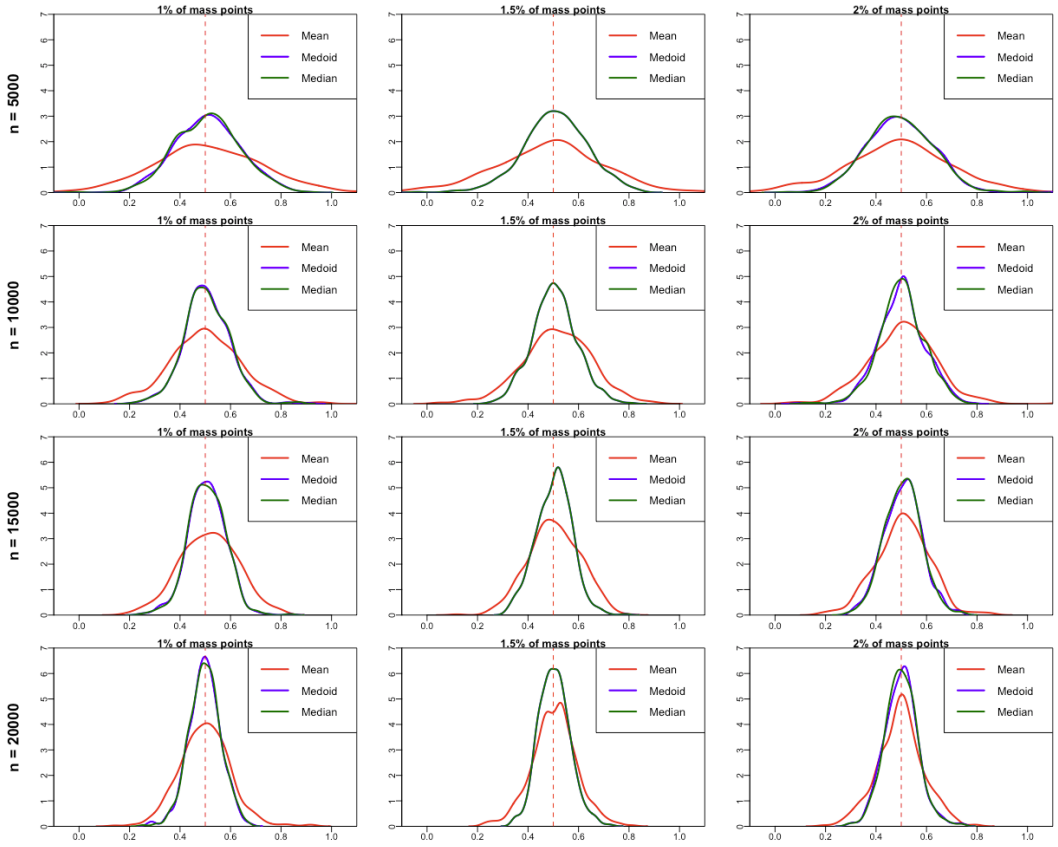


Figure 4.5: Estimated coefficient distribution: scenarios with a skewed outcome at each mass point

tage, producing distributions of the estimated coefficients that are more compact and concentrated around the true value, with robustness that makes them methodologically preferable in real applications where outcome asymmetry is the rule rather than the exception.

In addition to the distributions of the estimated coefficients obtained through the proposed method, we also assess the differences with the mean and median-collapsing method to evaluate performance by comparing two indices: the *Relative Bias (RBias)* and the *Root Mean Square Error (RMSE)* of the estimates based on the 500 replications. The RBias was calculated as:

$$RBias = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\tau}_s - \tau}{\tau}\right), \quad s = 1, 2, \ldots, 500 \tag{4.5}$$

Where S represents the number of replications in the simulation, $\hat{\tau}_s$ is the estimate of the treatment effect for the generic replication, and τ is the treatment effect fixed a priori (0.5). Instead, the RMSE was computed as:

$$RMSE = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\tau}_s - \tau)^2}, \quad s = 1, 2, \ldots, 500 \tag{4.6}$$

Tables 4.5 and 4.6 below report the results for RBias and RMSE. The upper part of the tables refers to a normal distribution of the outcome at each mass point, while the lower part refers to the skewed distribution. As with the previous figures, the tables show, for each row, the effect of sample size, while each column reports the different levels of the percentage of mass points considered. In all scenarios, RBias values are very close to zero, confirming the absence of systematic distortion for the mean, medoid, and median. No clear winner emerges: for n = 5,000 the mean is slightly closer to zero at 1% mass points (0.011 versus 0.008/0.004), whereas at 1.5% the medoid/median prevail (0.007/0.007 versus 0.015 for the mean). With n = 10,000 and n = 15,000, the differences remain minimal and mixed (e.g., n = 10,000, 2%: mean 0.003 vs medoid 0.002 vs median 0.003; n = 15,000, 1.5%: mean 0.004 vs 0.001/0.001). For n = 20,000, the values continue to fluctuate around zero, with a negligible level of distortion for all methods. In sum: under normality, RBias does not discriminate among the three approaches; it increases slightly in absolute value for smaller samples and, for a given n, varies little with the percentage of mass points. The evidence confirms that, under symmetric conditions, all estimators are unbiased, and the small differences observed reflect finite-sample fluctuations rather than structural differences. As for the skewed distribution, by contrast, marked differences emerge. With n = 5,000, the mean shows a larger RBias in absolute value: it

Table 4.5: RBias results for Normal outcome and Skewed outcome

RBias									
	1% of mass points			1.5% of mass points			2% of mass points		
	Mean	Medoid	Median	Mean	Medoid	Median	Mean	Medoid	Median
Normal distribution									
n=5000	-0.011	0.008	0.004	-0.015	-0.007	-0.007	-0.001	-0.002	0.003
n=10000	-0.003	-0.003	-0.002	-0.010	-0.001	-0.001	0.003	-0.002	-0.003
n=15000	0.001	0.003	0.004	-0.004	0.001	0.001	0.004	-0.001	0.001
n=20000	-0.004	-0.009	-0.008	0.001	0.003	0.003	0.001	-0.008	-0.008
Skewed distribution									
n=5000	0.025	0.018	0.018	-0.036	0.016	0.016	-0.013	-0.003	-0.005
n=10000	0.009	0.011	0.010	0.030	0.010	0.010	0.006	0.002	0.002
n=15000	0.027	0.012	0.013	0.008	0.005	0.005	0.011	0.013	0.012
n=20000	-0.003	-0.005	-0.004	0.008	0.008	0.008	-0.006	-0.001	-0.001

goes from +0.025 (1%) to 0.036 (1.5%) and 0.013 (2%), highlighting strong sensitivity to the share of mass points. By contrast, the medoid and the median remain close to zero and stable (e.g., 1.5%: 0.016/0.016), indicating clear robustness to asymmetry. With $n = 10{,}000$ the mean continues to show wider deviations (up to +0.030 at 1.5%), while medoid/median remain contained around 0.010 and even near zero or slightly negative at 2% (0.002/0.002). For $n = 15{,}000$ the pattern persists: the mean maintains higher positive RBias (0.027˘0.011), whereas the medoid/median settle at lower values (0.012˘0.005/0.013˘0.005). Only with $n = 20{,}000$ is partial convergence observed: the three procedures have very small RBias (between 0.006 and +0.008), but the robust methods remain slightly closer to zero, especially at 2% mass points (medoid/median 0.001 versus 0.006 for the mean). In short: under skewness, (i) the mean is more biased and more sensitive to the percentage of mass points (both sign and magnitude shifts); (ii) the medoid and the median show reduced, stable bias across all scenarios, with negligible differences between them; and (iii) increasing n reduces RBias for everyone, but convergence is faster for the robust approaches. Operationally, RBias therefore suggests: use the mean only in symmetric settings (where, in any case, the difference is minimal), whereas in the presence of skewness—especially with small n and many mass points—the medoid/median provide more reliable estimates and are less sensitive to the discrete structure of the data. Table 4.6 reports the RMSE for the three methods. RMSE decreases monotonically as n increases for all methods. Comparatively, the mean is systematically the most efficient: at $n = 5{,}000$ the mean's RMSE is 0.122˘0.132˘0.122 versus 0.154˘0.166˘0.151 for the medoid and 0.151˘0.166˘0.148 for the median—that is, an advantage on the order of 20–35%. The gap shrinks as the sample size grows (at $n = 20{,}000$: mean 0.058˘0.061; medoid/median 0.071˘0.076), but the hierarchy remains unchanged. The effect of the percentage of mass points is secondary: under normality, shifts between 1%, 1.5%, and 2% lead to modest changes (often on the order of 0.002˘0.01) and do not alter the ranking. Conclusion: in symmetric scenarios, the mean minimizes RMSE, while the median and medoid—though unbiased—are less efficient (higher variance), with differences that fade only for very large samples.

Turning to the skewed distribution, as seen with the other metrics, the picture reverses. For $n = 5{,}000$, the mean's RMSE is clearly higher ($\approx$ 0.214–0.218) than that of the medoid and median ($\approx$ 0.128-0.132), yielding a 35–45% reduction when adopting a robust approach. At $n = 10{,}000$ and $n = 15{,}000$ the robust methods still hold the edge (e.g., $n = 10{,}000$: mean 0.147–0.133 vs. medoid/median 0.088–0.092; $n = 15{,}000$: mean 0.116–0.106 vs. 0.071–0.075), and this persists at $n = 20{,}000$

Table 4.6: RMSE results for Normal outcome and Skewed outcome

RMSE									
	1% of mass points			1.5% of mass points			2% of mass points		
	Mean	Medoid	Median	Mean	Medoid	Median	Mean	Medoid	Median
Normal distribution									
n=5000	0.122	0.154	0.151	0.132	0.166	0.166	0.122	0.151	0.148
n=10000	0.083	0.106	0.103	0.085	0.104	0.104	0.090	0.110	0.107
n=15000	0.068	0.086	0.086	0.071	0.085	0.085	0.072	0.091	0.087
n=20000	0.061	0.076	0.074	0.061	0.076	0.076	0.058	0.072	0.071
Skewed distribution									
n=5000	0.214	0.131	0.129	0.218	0.128	0.128	0.215	0.132	0.130
n=10000	0.147	0.090	0.089	0.137	0.088	0.088	0.133	0.092	0.089
n=15000	0.116	0.075	0.074	0.103	0.071	0.071	0.106	0.075	0.074
n=20000	0.101	0.063	0.062	0.089	0.062	0.062	0.092	0.067	0.065

(mean 0.101–0.089 vs. 0.062–0.067), with margins still around 25˘35%. The percentage of mass points matters less than the other factors but, unlike in the normal case, the mean's RMSE shows greater sensitivity: in some panels non-monotonic changes appear (e.g., at n = 10,000 the mean's RMSE falls from 0.147 to 0.133 when moving from 1% to 2%), indicating that the combination of discreteness and skewness can amplify variability. The medoid and median, by contrast, are stable with respect to the percentage of mass points: small oscillations (typically ≤ 0.004) and near-overlapping curves, confirming their robustness.

In summary: (i) RMSE is driven by the shape of the distribution—under normality the mean wins; under skewness the medoid/median prevail. (ii) Increasing n reduces RMSE for everyone, but does not eliminate the relative advantage: even at n = 20,000 the robust methods maintain lower error in the presence of asymmetry. (iii) The percentage of mass points has a second-order effect: almost negligible under normality; more visible under skewness for the mean, while remaining mild for the robust methods.51

Operationally, this suggests clear guidance: use the mean in plausibly symmetric scenarios and when n is not too small; use the medoid/median when the outcome is skewed (or potentially so) and/or when the discrete structure induces many ties—especially with small or medium samples—since they substantially reduce the root mean squared error.

4.3.2 Simulation summary

Overall, the results from the density plots, RBias, and RMSE outline a coherent picture of the performance of the three collapsing approaches based on the mean, the medoids and the median. In scenarios with a normal distribution, all three estimators are unbiased, with medians well centered on the true effect ($\tau = 0.5$) and RBias values close to zero; what differentiates the approaches is efficiency: the mean yields more concentrated densities and a systematically lower RMSE than the medoid/median, with gaps that shrink but do not disappear as n increases, while the percentage of mass points plays a secondary role. In this context, using the medoid or the median brings no benefit and may entail a slight loss of precision.

By contrast, in scenarios with a skewed distribution, the differences between the two methods are reversed: the mean is affected by asymmetry and outliers, producing more dispersed estimates, RBias farther from zero, and higher RMSE—especially for small samples and when the density of mass points is high. In these situations, the medoid emerges as the more robust approach, able to keep estimates concentrated around the true value, limit relative bias, and reduce mean squared error.

Finally, increasing the sample size tends to soften the differences between the two methods, confirming that asymptotic properties lead to progressive convergence: in very large samples, the mean and the medoid are virtually equivalent.

In summary: (i) in symmetric scenarios the mean is preferable for efficiency; (ii) in skewed scenarios the robust methods (medoid/median) are clearly superior in stability and accuracy; (iii) increasing n improves all approaches but does not change the ranking; (iv) the share of mass points has little effect under normality and chiefly impacts the instability of the mean in the presence of asymmetry.

These findings therefore suggest a clear methodological trade-off: while the mean remains preferable under ideal conditions of symmetry, the medoid offers a significant advantage in realistic scenarios characterized by non-normal distributions and the presence of mass points, without substantially penalizing performance in other cases.

The results of our simulation study provide important insights for the literature on Regression Discontinuity with a discrete running variable, a context in which the presence of mass points represents a non-negligible methodological challenge. Traditional data collapsing strategies rely primarily on the mean of the outcomes within mass points, but our results show that while this approach is efficient under symmetry, it can be vulnerable in the presence of skewed or heavy-tailed distributions. In such scenarios, the mean is strongly affected by outliers and tends to generate more variable and biased estimates. The proposed alternative, namely collapsing based on the medoid, has proven capable of producing more robust estimates, containing bias and RMSE especially in small samples and when mass points are numerous. In this respect, it is important to emphasize that many studies with a discrete running variable are also characterized by low numbers of both mass points and observations (Cattaneo et al., 2024). Consider, for example, studies in the medical field where statistical units correspond to patients: in such contexts, the number of available observations is often limited, the running variable takes on discrete values (e.g., clinical scores or demographic characteristics), and the presence of mass points is common. Under these conditions, the use of the medoid proves particularly advantageous, as it reduces the influence of outliers and yields more stable and reliable estimates, even with small sample sizes.

From a practical perspective, our study suggests that the choice between mean, median and medoid should not be rigid, but rather depend on the empirical characteristics of the data: in real-world applications, where the symmetry of the outcome distribution is rarely guaranteed, adopting the medoid or median may represent a more cautious and methodologically sound strategy.

A final aspect worth discussing concerns the direct comparison between the me-

dian and the medoid. In our study, the two approaches produced almost identical results in terms of bias, variability, and RMSE, suggesting that—at least in the absence of further specifications—they can be regarded as equivalent from a statistical performance perspective. However, the key distinction between the two methods becomes evident when considering potential extensions of the design, particularly the inclusion of qualitative covariates. It is worth recalling that the inclusion of covariates is not required in simple RD settings, but it can be very useful for improving the robustness of the estimates and for conducting tests on them (see Section 2.2.3 for a detailed discussion on the use of covariates) (Cattaneo et al., 2023; Calonico et al., 2019; Frölich and Martin, 2019). In such cases, the medoid offers a notable advantage: it can also be used with mixed data (both quantitative and qualitative), identifying the most representative observation within the group without requiring arbitrary transformations of the variables. This property makes the medoid a more flexible and generalizable choice than the median, especially in complex applied settings where the available information is not exclusively numerical.

4.4 Application to the INVALSI data

In the following section, we apply the insights from the simulation study to the INVALSI case introduced in Chapter 3. In Section 3.2.2, we concluded that one of the main limitations of the analysis conducted up to that point was the neglect of the discrete nature of the running variable. Here, by contrast, we explicitly account for this feature and provide an estimate of the treatment effect that is not biased by this issue.

To do so, we first analyze the distribution of the running variable, namely the school ESCS (see Section 3.1.1 for a detailed discussion of the running variable under consideration). Recall that in Section 3.2.1 we examined the possibility of manipulation of the running variable, which could undermine the validity of the RD design. We applied both the McCrary density test (McCrary, 2008) and the Frandsen test (Frandsen, 2017), the latter specifically developed for discrete running variables. After conducting these tests and reviewing the results, we concluded that it is not possible for individuals to perfectly sort around the cutoff—choosing to be in the treatment group rather than the control group, or vice versa.

We should therefore carry out the analysis while accounting for mass points. However, before proceeding with the application, it is necessary to assess which of the collapsing procedures examined in the simulation study best fits the empirical structure of the INVALSI data. To this end, we evaluate the distribution of both

the running variable and the outcome across the mass points, in order to establish whether the nature of the data is closer to a symmetric context (where the mean proved to be more efficient) or to an asymmetric one (where the median and the medoid showed greater robustness). In this way, the empirical application will not only provide an estimate of the treatment effect but also serve as a concrete testbed for the guidelines derived from the simulation.

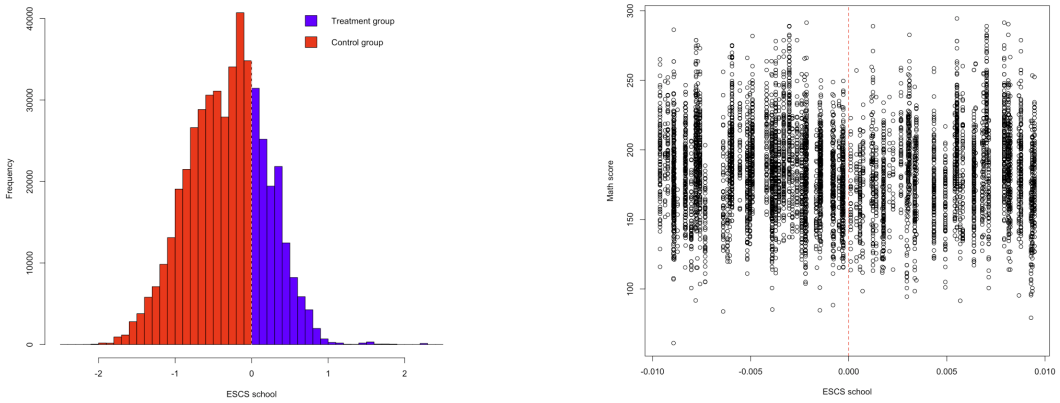
ESCS school value	**N**	**Group**
-0.000800915815791725	80	Control
-0.000769094348092414	75	Control
-0.000493575347122854	34	Control
-0.000341353380736931	94	Control
-0.000278840437243799	179	Control
0	0	\
0.000101513764336481	42	Treatment
0.000393309398053154	24	Treatment
0.000577354214352677	84	Treatment
0.000706378963977960	12	Treatment
0.000734973760698421	20	Treatment

Table 4.7: Analysis on the mass points: the five points closest to zero from the left and right of the running variable

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted.

In total, the running variable consists of 6,432 discrete values, defined as mass points. Table 4.7 provides an example of their distribution, showing the first five points to the left and right of the cutoff. As the table illustrates, some mass points are associated with relatively large numbers of observations, while others have fewer. Figure 4.6 presents an analysis of the distribution of the school ESCS and the outcome, the Math score. The left panel shows a histogram of the running variable, with bars colored according to treatment status (blue for treated observations and red for control observations). The right panel displays a scatterplot of the running variable against the outcome, restricting the sample to observations within the interval -0.10 to 0.10 of the running variable to reduce point overlap. Both graphs once again highlight the discrete nature of the running variable.

To better determine whether we are dealing with a symmetric scenario, where it would be appropriate to rely on the mean, or an asymmetric one, where the use of



(a) Histogram-ESCS school

(b) Scatter plot-ESCS school vs Math score

Figure 4.6: Bivariate analysis on the running variable and the outcome

Note. The cutoff was centered at 0 (0 corresponds to -0.31243). The running variable was inverted. To create the scatter plot, we filtered only those observations with the ESCS school value between -0.010 and 0.010.

the median or medoid may be preferable, we calculated the skewness of the outcome within each mass point.

The boxplot in Figure 4.7, which depicts the distribution of skewness in Math score within each mass point, clearly highlights the presence of a substantial number of outliers, distributed across both the lower and upper tails of the distribution. These are not isolated cases, but rather a consistent concentration indicating that, within many mass points, the outcome takes on extreme values with non-negligible frequency. This phenomenon substantially alters the symmetry of the distribution: while the central core of the data appears relatively compact, the tails are particularly extended, giving rise to strong local skewness. In practical terms, this implies that the mean, although theoretically appropriate in symmetric scenarios, tends here to be pulled toward the tails, losing representativeness and producing more variable and less reliable estimates. In other words, the mean is systematically penalized by the presence of outliers, which exert a disproportionate influence on the final result—especially in smaller mass points, where even a few extreme values can significantly distort the average.

These findings justify the use of alternative, more robust collapsing strategies such as the median and the medoid. Both methods have the advantage of drastically reducing the influence of anomalous values, focusing instead on the true central tendency of the data and thereby preserving the stability of the estimates. The median, in particular, represents the value that splits the distribution of outcomes within the mass point exactly in half, making it, by construction, insensitive to

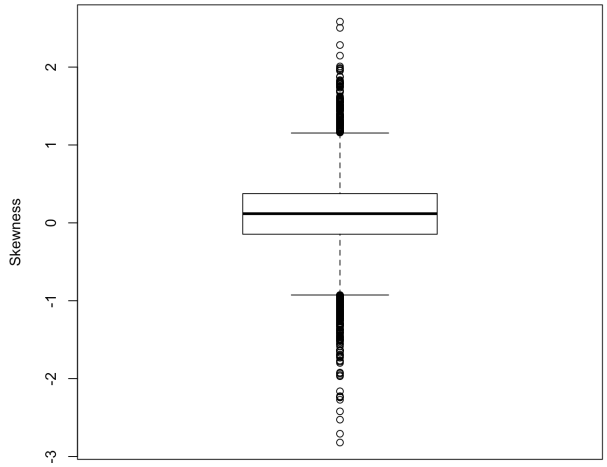


Figure 4.7: Distribution of the Skewness of the outcome in the mass points

long tails and extreme values. The medoid, on the other hand, selects the most representative observation within the group, maintaining adherence to the empirical data and allowing for extension to contexts that include qualitative variables.

Given our specific needs, the final choice falls on the median. As discussed in the previous section, the performances of the median and the medoid are nearly equivalent in terms of bias, variability, and RMSE. The distinction between the two becomes relevant primarily when qualitative covariates are included in the design, in which case the medoid retains an important advantage. In our case, however, there is no need to incorporate qualitative covariates into the analysis, and thus the medoid does not provide any additional benefit over the median. The latter is therefore preferable: it is simpler to implement, fully suited to managing the large number of outliers observed in the boxplots, and capable of producing robust and stable estimates, minimizing the distortions generated by the asymmetric nature of the mass points.

Following the recommendation of Cattaneo et al. (2024), we re-estimate the technique on the new dataset collapsed on the median and, in a second step, compare the results obtained on the collapsed data with those obtained on the raw data—most cases should lead to the same conclusions.

Table 4.8 reports the main position indices for the running variable and the outcome after collapsing on the median. As shown in the second column, the number of observations is $6,432$, corresponding to the number of mass points.

By contrast, Table 4.9 reports the estimated treatment effect obtained after applying the collapsing strategy. As in the previous analyses, the optimal bandwidth was selected using the `mserd` function from the `rdrobust` package, which chooses the bandwidth that minimizes the MSE while accounting for the bias–variance trade-off

Table 4.8: RD effect on the outcome Math score–Data collapsed on the median

Variable	N	NA	Min	Max	Median	Mean	CV
ESCS school	6,432	0	−2.12	2.42	−0.26	−0.29	1.97
Math score	6,432	0	93.21	276.38	185.35	187.00	0.15

(see Section 2.2.1). As weighting function, we used a triangular kernel, since when combined with the MSE-optimal bandwidth it yields a point estimator with optimal properties (Cattaneo et al., 2020). To estimate the causal effect, we employed a local polynomial of order 1, the most commonly used specification in the literature (see Section 2.1 for a discussion of local polynomial order).

With a bandwidth width of 0.278 on either side of the cutoff, the treatment effect remains negative and equal to −0.425, indicating that attending a low-ESCS school has adverse consequences for Math score. However, considering that the outcome ranges from 93.21 to 276.38, this effect is very small and can be regarded as practically negligible. Compared to the effect obtained in Chapter 3 equal to −1.071(Table 3.5, Section 3.2.2), and in line with the literature (Cattaneo et al., 2024), the effect has remained negative and has changed very little. This time, however, it is no longer statistically significant at any conventional significance level (p-value equal to 0.81).

Table 4.9: RD effect on the outcome Math score–Data collapsed on the median

Variable	**MSE-Optimal Bandwidth**	**RD Estimator**	**Kernel**	**Robust Inference**		**N of Obs**	
				p-value	*95% CI*	*left*	*right*
Math score	0.278	−0.425	Triangular	0.81	[−3.997 , 3.146]	1,192	1,012

4.5 Concluding remarks

In conclusion, this work has proposed a methodological extension to the literature on Regression Discontinuity design with discrete running variable, introducing the use of alternative and more robust collapsing strategies beyond the simple mean—specifically, the median and the medoid—and examining the conditions under which these approaches are most appropriate. The simulation study showed that while, in symmetric distributions, the mean remains the most efficient estimator, in the presence of skewness and outliers the robust methods offer significant advantages

in terms of stability and reliability of the estimates.

We then applied this framework to our case study using INVALSI data, with the aim of estimating the peer effect of attending a school classified as low-ESCS according to the ministerial threshold, by comparing Mathematics outcomes among students. Preliminary data analysis revealed pronounced skewness of scores within mass points, driven by the presence of numerous outliers. For this reason, we adopted the median as our collapsing method—rather than the medoid, given the absence of qualitative covariates. The results partly confirm the findings of the literature (Cattaneo et al., 2024), according to which estimates should remain stable after collapsing, but with some important differences: whereas the estimate based on the raw data yielded an effect of -1.071 (albeit modest in size), the new estimates based on the median return a coefficient of -0.425, very close in magnitude but no longer statistically significant.

This outcome suggests that once the discrete structure of the running variable and the asymmetric distribution of the outcome are properly accounted for, the evidence of a peer effect on students' Math scores is further attenuated to the point of being statistically insignificant. In other words, while the initial analysis pointed to a small peer effect, the robust analysis indicates that such an effect is, in practice, negligible.

Chapter 5

Handling with the hierarchical structure of the data

5.1 The Multilevel approach

The goal of Multilevel regression analysis is to estimate predictive models in much the same way as classical regression models. However, in this case one must account for the fact that the dataset is structured across a series of higher-level groups or clusters, deriving from nested data structures—for example, students nested within classes or schools, or employees nested within firms.

Depending on the field of application, multilevel models are referred to by different names in the literature. In Statistics, they are known as Mixed Models (Harville, 1977) or Hierarchical Linear Models; in Econometrics, as Random Coefficient Models (Swamy, 1970) or Random Effects Models for panel data; in Biostatistics, as Mixed Models for repeated measures (Laird and Ware, 1982) or Random Effect Models; and in Educational Statistics, they are commonly called Multilevel Models (Aitkin and Longford, 1986).

The literature on multilevel modelling is extensive and well established. Among the major references, Snijders and Bosker (1999) and Snijders and Bosker (2012) provide a particularly clear and comprehensive introduction to the fundamentals of multilevel models. Hox et al. (2010) offers a more concise overview but addresses a broader range of topics related to hierarchical data analysis. A systematic exposition accompanied by numerous detailed applications is offered by Raudenbush and Bryk (2002), while Goldstein (2011) is considered a classic in the field, with a more technical and advanced treatment, especially in the context of education. The volume edited by Leeuw and Meijer (2008) is a useful reference manual, particularly suitable for an overall understanding of multilevel methods. A broader treatment, which includes not only multilevel models but also Rasch models, IRT, factor models, and structural equation models, is provided by Skrondal and Rabe-Hesketh (2004). Several authors have also summarised theoretical and empirical developments in this area (Grilli and Rampichini, 2009; Grilli and Rampichini, 2015).

The issue with nested data structures is that they violate the independence assumption required by traditional statistical techniques such as ANOVA and Ordinary Least Squares (OLS) regression. Therefore, when working with nested data structures, it is important to recognize that individuals belonging to the same group tend to behave similarly because they share the same environment, and are likely to behave differently from individuals in different groups. These violations of independence often make multilevel modelling necessary, since traditional analytical models can produce inflated Type I errors and biased parameter estimates. One of the key assumptions underlying classical regression models is that residuals are uncorrelated. This implies that the data for each statistical unit are independent of those for other units. However, when data are nested—for instance, students within different schools, regions, or countries—this assumption of independence does not hold.

An example of nested data is illustrated in Figure 5.1, where students are located within schools, which are in turn located within regions. In this example, the highest-level grouping variable is defined by the region of residence. It is up to the researcher to determine which higher-level variable defines the hierarchical structure of the data. However, the presence of nested data does not automatically imply the need to use multilevel models (Peugh, 2010). Before proceeding with multilevel analysis, it is good practice to verify whether the hierarchical structure of the data has a meaningful quantitative impact on the way variance is partitioned. This is assessed using the *Intraclass Correlation Coefficient (ICC)*. The ICC quantifies the proportion of the total variance of the outcome that is attributable to differences between groups (higher level), rather than to differences between individuals within those groups (lower level). In other words, it measures the extent to which

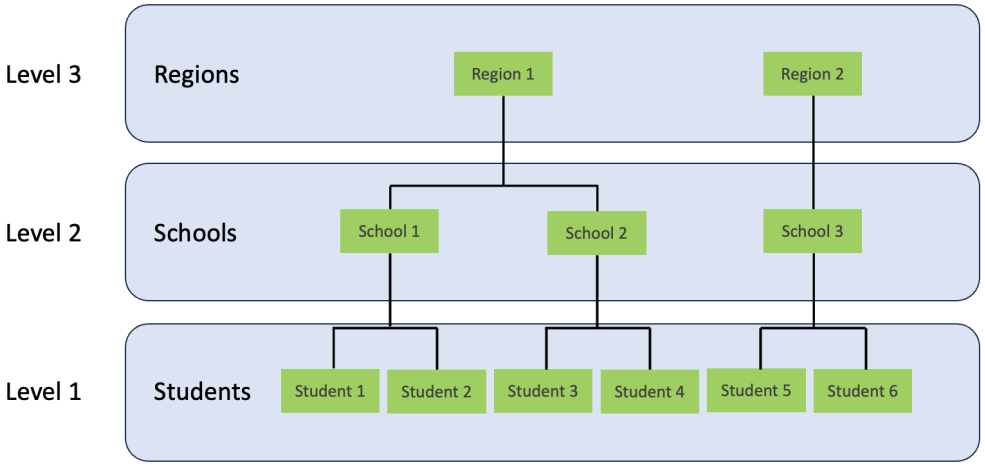


Figure 5.1: Nested data example: students in schools in regions

individuals within the same group resemble one another.

There are two main statistical models that form the basis of multilevel modelling: the Random Intercept model and the Random Intercept and Slope model. In a Random intercept model, as the name suggests, the intercept term is allowed to vary across the clusters. To do so, we introduced a random variable u_j to account for the variance caused by clusters, that is the random variable responsible for unique intercepts for each group. While simple regression uses a single line to represent the best fit for all data, a random intercept model features multiple distinct regression lines–one for each group–in addition to an overall or common regression line. The Random intercept model is expressed as:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij} \\ &= (\beta_0 + u_{0j}) + \beta_1 x_{ij} \end{aligned} \tag{5.1}$$

where:

- y_{ij} is the outcome for individual i in group j;
- β_0 represents the overall mean intercept common to all groups;
- x_{ij} is the independent variable for individual i in group j;
- β_1 is the slope associated with x_{ij};
- u_{0j} is the level-2 random effect, representing the deviation of group j from the overall intercept β_0, distributed as $u_{0j} \overset{\text{iid}}{\sim} N(0, \sigma^2_{u0})$;
- e_{ij} is the level-1 individual error term, distributed as $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_e)$, with $e_{ij} \perp u_{0j}$.

The total error is represented by the sum between u_{0j} and e_{ij}. Thus, the model has:

- Homoschedasticity:

$$Var(y_{ij}|x_{ij}) = \sigma^2_{u0} + \sigma^2_e$$

  where:

  - σ^2_{u0} is the between-cluster variance
  - σ^2_e is the within-cluster variance

- Homogeneous correlation among the responses of the units of the same cluster:

$$Cov(y_{ij}, y_{i'j}|x_{ij}, x_{i'j}) = \sigma^2_{u0}$$

- No correlation among the responses of units of different clusters:

$$Cov(y_{ij'}, y_{i'j'}|x_{ij'}, x_{i'j'}) = 0$$

The second model, the Random Intercept and Slope Model, is a more complex model that allows the slope to vary across groups. It is represented by the following equation:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij} \\ &= (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{ij} + e_{ij} \end{aligned} \tag{5.2}$$

where:

- y_{ij} is the outcome for individual i in group j;
- β_0 and β_1 are, respectively, the mean intercept and the mean slope in the population;
- x_{ij} is the independent variable of the individual i in group j;
- u_{0j} is the level-2 random term, representing the deviation of group j from the average intercept β_0;
- u_{1j} is the random deviation of the slope for group j from the mean slope β_1;
- e_{ij} is the level-1 individual error.

Graphically, the Random Intercept and Slope Model is characterized by the fact that each group has its own intercept and slope. This implies that groups can differ not only in the mean level of the outcome, but also in the relationship between the outcome and the predictor. This time, the element that varies is the slope, β_{1j}, which is composed of the sum of the mean slope, β_1, and the random slope deviation, u_{1j}: each cluster has its own slope. The random terms (u_{0j}, u_{1j}) follow a bivariate normal distribution:

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \overset{\text{iid}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix} \right), \qquad e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_e), \qquad e_{ij} \perp (u_{0j},\, u_{1j}).$$

The total error is represented by the sum $u_{0j} + u_{1j}x_{ij} + e_{ij}$. Thus, the model has:

- Heteroscedasticity:

$$Var(y_{ij}|x_{ij}) = [\sigma^2_{u0} + 2\sigma_{u01}x_{ij} + \sigma^2_{u1}x^2_{ij}] + \sigma^2_e$$

  where:

  - σ^2_{u0} is the between-cluster variance
  - σ^2_e is the within-cluster variance

- Non-homogeneous covariance among the responses of the units of the same cluster:

$$Cov(y_{ij}, y_{i'j}|x_{ij}, x_{i'j}) = \sigma^2_{u0} + \sigma_{u01}(x_{ij} + x_{i'j}) + \sigma^2_{u1}x_{ij}x_{i'j}$$

- No covariance among the responses of units of different clusters:

$$Cov(y_{ij'}, y_{i'j'}|x_{ij'}, x_{i'j'}) = 0$$

5.2 The Multilevel Regression Discontinuity

Thanks to the Regression Discontinuity Design (RDD), as demonstrated in the previous chapters, it is possible to estimate the treatment effect of a running variable on an outcome. Multilevel models, on the other hand, make it possible to account for any hierarchical structure in the data, estimating how and how much an effect changes within and between the groups considered. A combination of the two approaches would provide a measure of the treatment effect and how this varies across groups.

Let us: i) a running variable X_{ij} which is the variable on which the assignment mechanism to the treatment or control group is based assuming that there are n units indexed with $i = 1, 2, \ldots, n$ belonging to the group j ii) a cutoff c equal to 0 for semplicity; iii) a binary variable D_{ij} equal to 1 if the unit i in the group j is treated, 0 otherwise; iv) y_{ij} a potential outcome of the unit i belonging to the group j. The two-level regression model according to the equation 5.2 can be expressed as:

$$y_{ij} = \beta_{0j} + \beta_1(X_{ij} - c) + \beta_{2j}D_{ij} + e_{ij} \tag{5.3}$$

where:

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$\beta_{2j} = \beta_2 + u_{2j}$$

β_0 represents the mean outcome score, across groups, of units with a running variable equal to the cutoff; β_{2j} is the parameter of interest and represents the treatment effect across groups; u_{0j} and u_{2j} are, respectively, the deviation of each group's coefficient from the overall intercept and slope parameters; β_1 is the general effect of the running variable on the outcome.

The proposed approach seeks to go beyond standard practices for implementing multilevel RD in educational settings, as described in previous works (Luyten, 2006; Steinmann and Olsen, 2006), by incorporating both conventional and more advanced RD procedures—and, more generally, Local Polynomial techniques—such as optimal bandwidth selection and the construction of observation-specific weighting functions (see Section 2.2 for a detailed discussion). In the studies reviewed, the issue of the bandwidth—namely, the interval of observations around the cutoff within which the assumption of quasi-randomness underlying Regression Discontinuity is plausible (see Section 2.1)—is not taken into account. In this work, the question was raised whether selecting a single, overall bandwidth without considering the hierarchical structure of the data may conceal or attenuate the treatment effect within individual groups (such as schools or regions), thereby compromising the ability to detect potential territorial or institutional heterogeneity. For this reason, the study compares a global approach—which estimates a single bandwidth for the entire sample—with alternative approaches that incorporate the grouped structure of the data from the outset. These alternative approaches aggregate information from bandwidths estimated separately for each group. In doing so, it becomes possible to assess whether, and to what extent, accounting for the hierarchical nature of the data improves the identification of the treatment effect in local contexts.

To this end, we introduce three different methods: the *General Bandwidth Method*, the *Average Bandwidth Method*, and the *Weighted Average Bandwidth Method*. These three methods are conceptually similar; their differences lie in how the optimal bandwidth is selected and how it is weighted.

The General Bandwidth Method consists of three steps. **Step 1:** Estimate the RD to select the optimal bandwidth h_{opt} using the `mserd` function from the `rdrobust` package, which chooses the bandwidth that minimizes the MSE while accounting for the bias–variance trade-off (see Section 2.2.1). **Step 2:** Filter the dataset by retaining only observations within the bandwidth h_{opt} and apply triangular kernel

weights, thereby assigning weights that decline symmetrically and linearly as the value of the score moves away from the cutoff (Cattaneo et al., 2020). **Step 3:** Estimate a multilevel model using the weighting system created in Step 2 and the optimal bandwidth h_{opt} selected in Step 1.

Unlike the previous method, the Average Bandwidth Method is designed to account explicitly for the fact that different groups may be characterized by different optimal bandwidths. To capture this heterogeneity without losing information, bandwidth selection is performed at the group level. **Step 1:** Estimate separate RD models within each group j to select the group-specific optimal bandwidths $h_{opt,j}$ using the `mserd` function. **Step 2:** Aggregate these group-specific bandwidths by computing their arithmetic mean:

$$h_{opt,\text{Average}} = \frac{1}{J} \sum_{j=1}^{J} h_{opt,j}$$

Step 3: Filter the data, retaining only observations within $h_{opt,\text{Average}}$, and apply triangular kernel weights. **Step 4:** Estimate the multilevel model using the weighting system from Step 3 and the optimal bandwidth $h_{opt,\text{Average}}$ obtained in Step 2.

The Weighted Average Bandwidth Method, like the previous approach, aims to capture heterogeneity across groups during the bandwidth selection phase. However, in contrast to the simple average method, it also accounts for differences in group sizes: larger groups are assigned greater weight when aggregating bandwidths. **Step 1:** Estimate separate RD models for each group j to obtain group-specific optimal bandwidths $h_{opt,j}$ using `mserd`. **Step 2:** Aggregate the group-specific bandwidths into a weighted mean, using weights w_j proportional to group size:

$$w_j = \frac{n_j}{\sum_{k=1}^{J} n_k}$$

where n_j is the number of observations in group j. The resulting bandwidth is:

$$h_{opt,\text{Weighted}} = \sum_{j=1}^{J} w_j \cdot h_{opt,j}$$

Step 3: Filter the dataset to include only observations within $h_{opt,\text{Weighted}}$ and apply triangular kernel weights. **Step 4:** Estimate the multilevel model using the weighting system created in Step 3 and the optimal bandwidth $h_{opt,\text{Weighted}}$ selected in Step 2.

The integration of these concepts is systematically synthesized and presented in Table 5.1.

Table 5.1: MRD: A summary of the three different approaches

Method	Functioning
General h	Step 1: estimate a RD to select the optimal bandwidth h_{opt} on the whole set of data; Step 2: create a triangular kernel function for all the observations in the subset defined by h_{opt}; Step 3: estimate a multilevel regression according to h_{opt} selected in step 1 and the kernel function created in step 2.
Average h	Step 1: estimate different RD for each group j and select the optimal bandwidth $h_{opt,j}$; Step 2: aggregating the different bandwidth: $h_{opt,\text{Average}} = \frac{1}{J}\sum_{j=1}^{J} h_{opt,j}$ Step 3: create a triangular kernel function on the subset defined by $h_{opt,\text{Average}}$; Step 4: estimate a multilevel regression according to $h_{opt,\text{Average}}$ from step 2 and the kernel from step 3.
Weighted average h	Step 1: estimate different RD for each group j and select $h_{opt,j}$; Step 2: aggregating the different bandwidth according to the weights w_j $w_j = \frac{n_j}{\sum_{k=1}^{J} n_k}, \qquad h_{opt,\text{Weighted}} = \sum_{j=1}^{J} w_j \cdot h_{opt,j}$ Step 3: create a triangular kernel on the subset defined by $h_{opt,\text{Weighted}}$; Step 4: estimate a multilevel regression according to $h_{opt,\text{Weighted}}$ from step 2 and the kernel from step 3.

5.2.1 Simulation study

In the following section, a simulation study will be conducted to evaluate the three proposed methods in order to highlight their limitations and advantages. Different simulation scenarios are used to compare the performance of the methods under

various conditions.

Within the potential outcomes framework, the data generation for an observation i belonging to the group j is based on a set of group-specific parameters (a_{0j}, a_{1j}, b_{0j}, b_{1j}, r_j). In particular, the equations of our simulation study are defined as follows:

$$y_{0_{ij}} = a_{0j} + b_{0j}X_{ij} + e_{ij} \tag{5.4}$$

$$y_{1_{ij}} = y_{0_{ij}} + a_{1j} + b_{1j}X_{ij} \tag{5.5}$$

$$X_{ij} = \mu_X + r_j + \pi_{ij} \tag{5.6}$$

with:

$$e_{ij} \sim N(0, \sigma_e^2)$$
$$\pi_{ij} \sim N(0, 1 - \text{ICC}_X)$$

where $y_{0_{ij}}$ denotes the outcome in the absence of treatment, while $y_{1_{ij}}$ represents the observed outcome under treatment. The error term e_{ij} follows a normal distribution with variance σ_e^2, which remains constant across all observations and groups. The term r_j reflects a group-level shift effect on the running variable, which also incorporates the random component π_{ij}, constructed to ensure that the overall distribution of the score has mean zero and unit variance. The parameter ICC_X instead represents the *Intraclass Correlation* of the running variable X. Each group includes a number n_j of observations. The generation of group-level coefficients and parameters follows the distributions:

$$a_{0j} \sim N(\mu_{a0}, \sigma_{a0}^2)$$
$$b_{0j} \sim N(\mu_{b0}, \sigma_{b0}^2)$$
$$a_{1j} \sim N(\mu_{a1}, \sigma_{a1}^2)$$
$$b_{1j} \sim N(\mu_{b1}, \sigma_{b1}^2)$$
$$r_j \sim N(0, \text{ICC}_X)$$

where μ_{a0}, μ_{b0}, μ_{a1} and μ_{b1} correspond to the means of the coefficients and σ_{a0}^2, σ_{b0}^2, σ_{a1}^2 and σ_{b1}^2 to their respective variances. All coefficients are assumed to be independent of one another.

Within each simulation scenario, we vary the parameter σ_{a1}^2 (heterogeneity of the Local Average Treatment Effect (LATE) across groups), the ICC (0.00, 020, 0.70), and the group size, simulating scenarios with perfectly balanced observations across groups as well as unbalanced scenarios. In the balanced case, each group has the

same size, set to N/J, resulting in groups that are perfectly equal in terms of observations. To simulate imbalance, the size of each group was generated by drawing from a multinomial distribution with parameters N and p, where N represents the total number of observations and $p = (p_1, \ldots, p_J)$ is a probability vector of length J. The probabilities p_j were randomly drawn from a uniform distribution and then normalized so that $\sum_{j=1}^{J} p_j = 1$. In this way, each simulation assigns to every group a size proportional to its corresponding p_j, introducing heterogeneity in group sizes while keeping the total number of observations fixed. This setup allows us to model more realistic scenarios where groups do not all have the same sample size, avoiding results that would otherwise depend solely on an artificially balanced design. In each scenario, we decided to keep both the total sample size and the number of groups j fixed at $20,000$ and 20, respectively. Furthermore, to ensure robustness of the results, we performed resampling with 500 replications and applied each of the three proposed methods to each of the 500 samples. Below are the details of the factors and levels considered for the simulation:

- σ^2_{a1}: 0.00; 0.20; 0.60
- Group balancing: balanced; unbalanced
- ICC: 0.00; 0.20; 0.70

The fixed parameter are: $\mu_{a0} = 0.70$, $\mu_{b0} = 0.05$, $\mu_{a1} = 0.07$, $\mu_{b1} = 0.025$, $\sigma^2_{a0} = 0.09$, $\sigma^2_{b0} = 0.00$, $\sigma^2_{b1} = 0.00$, $\sigma^2_e = 0.40$, $\mu_X = 0.00$. The total number of scenarios obtained from the combination of the above-described levels of the design-factors is equal to 18.

Table 5.2 presents the detailed simulation design. The simulation study explores the impact of treatment effect heterogeneity across groups, differences in the balancing among the groups and the intraclass correlation of the running variable. Part of this specification, as well as the simulation strategy adopted, follows the approach proposed by Litschwartz and Miratrix (2021).

5.2.2 Simulation results

The results of the simulation study were analyzed using a graphical summary to facilitate interpretation. In particular, for each scenario considered, the empirical distributions of the main metrics of interest is presented — LATE Bias, LATE heterogeneity Bias, the bandwidth, and the number of observations contained within the bandwidth — using side-by-side boxplots as a summary tool. The LATE bias

Table 5.2: MRD: Simulation study plan

Scenarios	σ^2_{a1}	**Groups balancing**	ICC
1	0.00	Balanced	0.00
2	0.00	Balanced	0.20
3	0.00	Balanced	0.70
4	0.00	Unbalanced	0.00
5	0.00	Unbalanced	0.20
6	0.00	Unbalanced	0.70
7	0.20	Balanced	0.00
8	0.20	Balanced	0.20
9	0.20	Balanced	0.70
10	0.20	Unbalanced	0.00
11	0.20	Unbalanced	0.20
12	0.20	Unbalanced	0.70
13	0.60	Balanced	0.00
14	0.60	Balanced	0.20
15	0.60	Balanced	0.70
16	0.60	Unbalanced	0.00
17	0.60	Unbalanced	0.20
18	0.60	Unbalanced	0.70

and LATE heterogeneity bias were calculated as:

$$LATE \text{ Bias} = \hat{\mu}_{a1} - \mu_{a1} \tag{5.7}$$

$$LATE \text{ heterogeneity Bias} = \hat{\sigma}^2_{a1} - \sigma^2_{a1} \tag{5.8}$$

In addition, we assess the differences and evaluate performance by comparing two indices: the LATE Relative Bias (RBias) and the LATE Root Mean Square Error (RMSE) of the estimates based on the 500 replications. The LATE RBias was calculated as:

$$LATE \text{ RBias} = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\mu}_{a1} - \mu_{a1}}{\mu_{a1}}\right), \qquad s = 1, 2, \ldots, 500 \tag{5.9}$$

Where S represents the number of replications in the simulation, $\hat{\mu}_{a1}$ is the estimate of the Local Average Treatment Effect for the generic replication, and μ_{a1} is the treatment effect fixed a priori. Instead, the LATE RMSE was computed as:

$$LATE \text{ RMSE} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\mu}_{a1} - \mu_{a1}\right)^2}, \qquad s = 1, 2, \ldots, 500 \tag{5.10}$$

The graphical representation allowed us to immediately highlight the variabil-

ity of the estimates obtained, as well as to compare the performance of the three proposed methods under different simulation conditions.

Figure 5.2 shows the distribution of LATE Bias for each of the 18 simulated scenarios, comparing the three estimation approaches: General bandwidth, Average bandwidth, and Weighted average bandwidth. Each scenario combines a different level of treatment effect heterogeneity (σ^2_{a1}= 0.00, 0.20, 0.60), group balance (Balanced vs. Unbalanced), and intra-group correlation level (ICC= 0.00, 0.20, 0.70). Looking at the results, it emerges that in the first six scenarios (in which σ^2_{a1} is equal to 0.00), the estimated bias is close to zero for all three approaches, regardless of balance and ICC. This is expected, since in the absence of treatment heterogeneity, the estimation of the average effect is not affected by between-group variation. As the variance of the treatment effect increases (σ^2_{a1}= 0.20 in scenarios from 7 to 12 and σ^2_{a1}= 0.60 in scenarios from 13 to 18), a progressive increase in the dispersion of the bias is observed, with a tendency for the mean bias to rise, especially in scenarios with unbalanced data. In particular, scenario 18 (σ^2_{a1}= 0.60, unbalanced, ICC = 0.70) shows the most distorted estimates, with extreme values and high variability.

Comparing the approaches, it can be seen that the General bandwidth approachtends to produce slightly lower bias than the other two in the presence of high heterogeneity and imbalance, suggesting greater structural robustness. However, overall, the three approaches appear to be nearly equivalent.

Taken together, the simulation confirms that the simultaneous presence of high treatment heterogeneity, imbalance, and high ICC leads to greater difficulty in accurately estimating the LATE, with a marked impact on bias.

Figure 5.3 depicts the distribution of LATE heterogeneity Bias. The analysis shows that in the first six scenarios (in which σ^2_{a1} is equal to 0.00), whereby design no heterogeneity exists across groups, all three approaches return a variance estimate close to zero, as expected. In these cases, the bias is essentially null with negligible dispersion, confirming the efficacy of the model. In the next six scenarios (from 7 to 12), where treatment effect variance is moderate (σ^2_{a1}), a slight decrease in negative bias emerges, suggesting a tendency to overestimate the true variance. However, the median remains low and quite similar across the three approaches. The variability of the bias increases slightly but remains manageable. Finally, in the six scenarios characterized by high heterogeneity (in which σ^2_{a1} is equal to 0.60), the LATE heterogeneity Bias increases substantially, with much wider dispersion. In particular, extreme positive bias values are observed in the presence of high intra-group correlation (ICC = 0.70) and imbalanced groups (scenario 18). In these contexts, all three approaches tend to underestimate the true variance, but relative

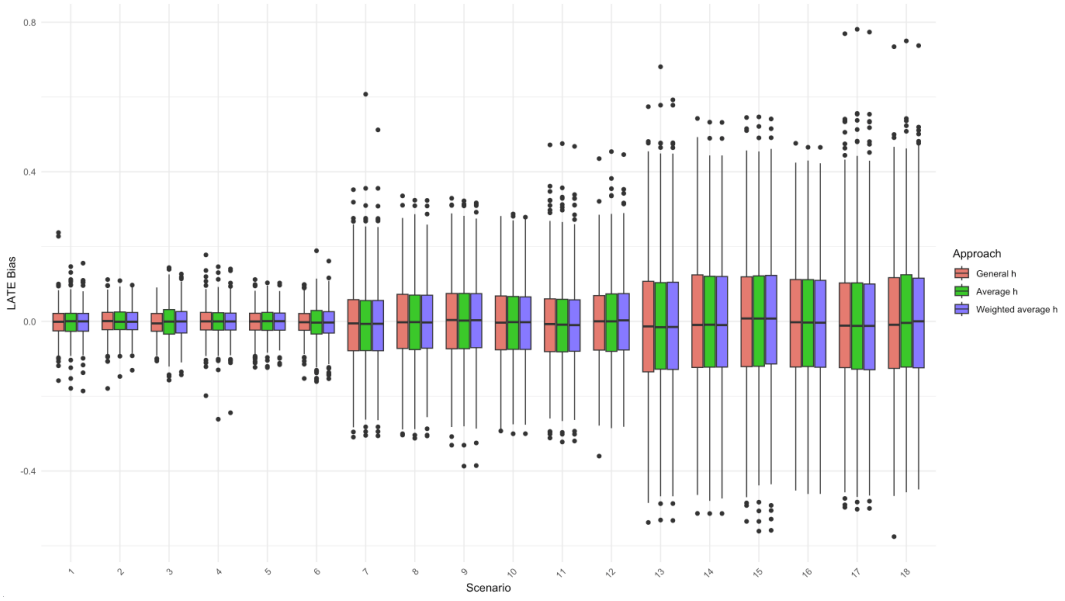


Figure 5.2: Distribution of the LATE Bias value for each scenario

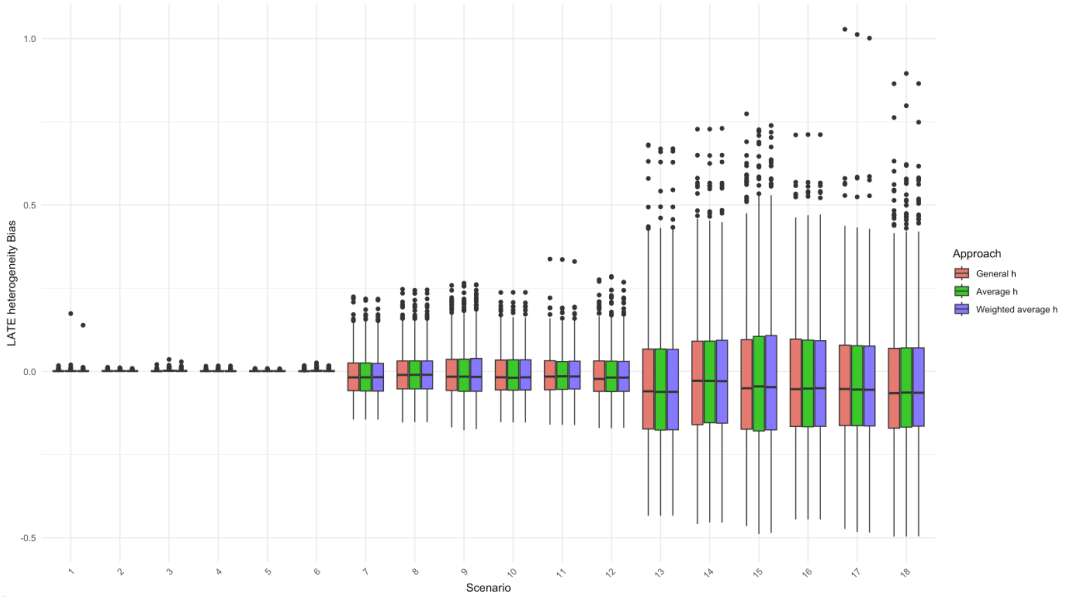


Figure 5.3: Distribution of the LATE heterogeneity Bias value for each scenario

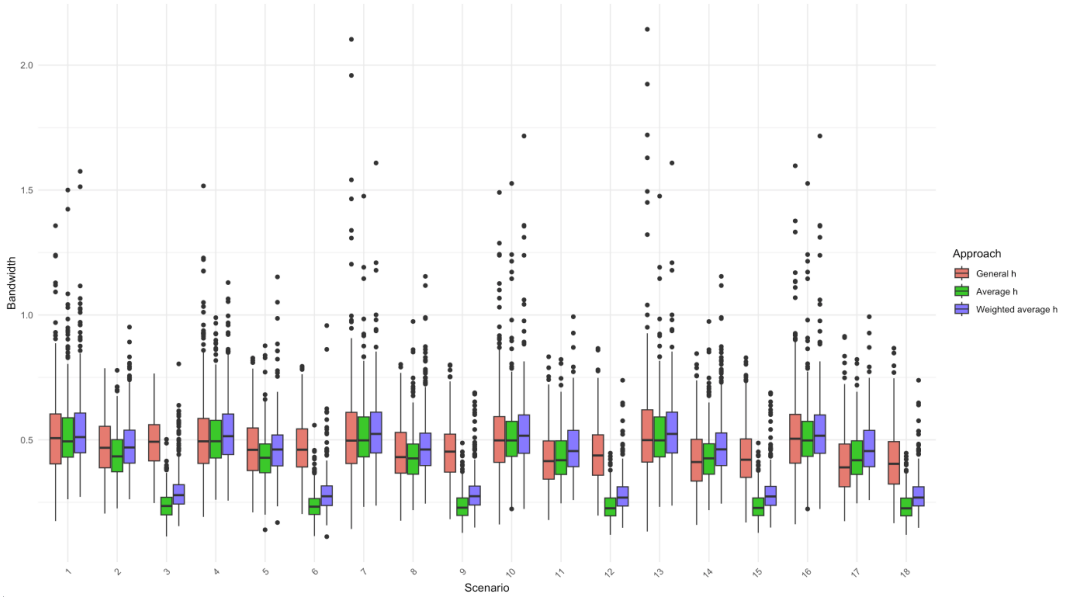


Figure 5.4: Distribution of the bandwidth selected for each scenario

consistency among methods is preserved, with the General bandwidth approach being slightly more centered on average.

Thus, the precision in estimating the variance of the treatment effect decreases as heterogeneity and scenario complexity increase (high ICC and unbalanced groups). However, all approaches behave stably in simpler contexts, showing only marginal differences in the more complex ones.

From Figure 5.4, which shows the distribution of the bandwidth selected for each of the three approaches, a clear variability in bandwidth is observed in relation to the complexity of the simulated scenario. In general, in the first six scenarios (with zero variance of the treatment effect and thus no heterogeneity), the bandwidths tend to be smaller and relatively stable, especially for the Average bandwidth approach, which exhibits a more contained distribution compared to the other two. As treatment effect heterogeneity increases (σ^2_{a1}= 0.20 in scenarios from 7 to 12, and σ^2_{a1} =0.60 in scenarios from 13 to 18), the selected bandwidth tends to widen. In these more complex scenarios, the General bandwidth approach shows a tendency to select, on average, wider bandwidths compared to the other two, perhaps as a compromise strategy to smooth across group heterogeneities. Indeed, Weighted average bandwidth and Average bandwidth methods tend to produce a more contained distribution, slightly shifted toward smaller values, especially in the presence of high ICC (scenarios 3, 6, 9, 12, 15, and 18). One possible explanation lies in the fact that, when estimating the RD separately at the group level, the strong within-group similarity given by an higher ICC produces a high density of observations around

the cutoff for regions close to the threshold. This, in turn, leads the selection algorithms to narrow the optimal bandwidth. In the General bandwidth approach, by contrast, the estimation is carried out on the entire aggregated sample: in this setting, the density of observations at the cutoff is more diluted, which results in a broader estimated optimal bandwidth.

However, such narrow bandwidths in these scenarios lead to estimation issues. Specifically, for both the Average bandwidth and Weighted average bandwidth methods, bandwidths that are too restrictive prevent the algorithm from estimating the coefficients of some groups in: 0.40% of cases in scenario 3, 1% in scenario 6, 1% in scenario 9, 0.60% in scenario 12, 1% in scenario 15, and 0.60% in scenario 18. This occurs because, with narrow bandwidths and separate regressions for each group, it can happen that some treatment and/or control groups contain no observations—or very few, and the algorithm is unable to estimate any effect. By contrast, with the General bandwidth approach, we did not encounter any estimation problems.

Table 5.3 reports the values of the LATE Rbias, as defined in equation 5.9, for each scenario. It provides an overview of how the bias varies as σ^2_{a1} and the ICC increase, in both balanced and unbalanced group settings. All three approaches yield essentially unbiased estimates, with LATE Rbias values close to zero under every condition. However, as σ^2_{a1} increases, the Rbias tends to become more negative, indicating a systematic tendency to underestimate the average treatment effect. This phenomenon is particularly evident in scenarios with low or moderate ICC (0.00 and 0.20), where the higher between-group variability amplifies distortion in aggregated estimates, and it is more pronounced in unbalanced settings.

As the intra-class correlation increases ($ICC = 0.70$), this trend partially reverses: the Rbias decreases in absolute value and in some cases becomes positive, indicating a slight overestimation. This reflects the fact that, when there is strong within-group homogeneity, fewer effective observations are sufficient to estimate the local treatment effect at the cutoff. In these scenarios, the *Average bandwidth* and *Weighted average bandwidth* methods tend to select smaller bandwidths and produce less biased estimates in unbalanced contexts, whereas the *General bandwidth method* continues to perform better in both balanced and unbalanced settings when the ICC is equal to 0.20 or 0.70. Overall, the latter approach appears more robust when groups are of similar size and the intra-group correlation is moderate or strong ($ICC = 0.20$ or $ICC = 0.70$), whereas the bandwidth-by-group procedures have a slight advantage when groups are homogeneous and ICC is low.

Table 5.4 presents the LATE RMSE values as defined in equation 5.10 for each scenario. As with the LATE Rbias, differences between the three methods are

Table 5.3: LATE RBias results

LATE RBias									
	ICC= 0.00			*ICC*= 0.20			*ICC*= 0.70		
	General h	Average h	Weighted average h	General h	Average h	Weighted average h	General h	Average h	Weighted average h
Balanced									
σ^2_{a1} =0.00	-0.018	-0.018	-0.021	-0.005	-0.006	-0.008	-0.041	-0.025	-0.034
σ^2_{a1} =0.20	-0.116	-0.111	-0.115	-0.050	-0.055	-0.058	0.015	0.019	0.035
σ^2_{a1} =0.60	-0.177	-0.166	-0.169	-0.080	-0.085	-0.088	0.030	0.038	0.032
Unbalanced									
σ^2_{a1} =0.00	0.003	0.001	-0.004	-0.014	-0.005	-0.009	-0.053	-0.036	-0.046
σ^2_{a1} =0.20	-0.056	-0.046	-0.053	-0.091	-0.100	-0.110	-0.028	0.007	0.006
σ^2_{a1} =0.60	-0.076	-0.061	-0.069	-0.120	-0.141	-0.153	-0.043	0.016	0.006

minimal, as all exhibit stable performance with LATE RMSE values close to zero. RMSE increases as σ^2_{a1} grows: moving from 0.00 to 0.20 and 0.60, the root mean squared error increases across all settings. When groups are balanced, the three approaches are nearly equivalent for low or moderate ICC.

5.2.3 Simulation summary

The joint analysis of LATE Bias, LATE heterogeneity Bias, the bandwidth distribution and the number of effective observations in the bandwidth selected, the LATE Rbias and the LATE RMSE across the 18 simulated scenarios highlight important differences among the three approaches. Overall, the General bandwidth approach proves to be the most stable and reliable, showing robust performance in terms of Bias, RBias and RMSE even under more complex conditions, such as scenarios with high treatment effect heterogeneity ($\sigma^2_{a1} = 0.60$) or high intra-group correlation ($ICC = 0.70$). However, this greater stability appears to stem from a tendency to select wider bandwidths, thereby increasing the number of observations but potentially trading off local estimation precision.

The Average bandwidth method, which applies a simple average of local bandwidths, performs well in simpler contexts (low heterogeneity and balanced groups), but becomes more unstable and sometimes more biased as scenario complexity grows, especially in the presence of unbalanced groups. Its strength lies in its simplicity, but it tends to underestimate variability across groups.

The Weighted average bandwidth method, which accounts for group size, behaves very similarly to Average h in balanced scenarios (as expected), but shows slight improvement in unbalanced contexts, where giving more weight to larger groups helps yield more centered estimates. However, overall, its variability remains high in the presence of strong heterogeneity or intra-group correlation.

In conclusion, the General bandwidth method appears to be the most versatile, better able to adapt to different scenarios, offering a good compromise between bias, variance, and estimation stability, while also consistently providing group-level coefficient estimates. The Average bandwidth method is more suitable for simple and homogeneous scenarios, while Weighted average bandwidth may serve as an intermediate choice in contexts with group size imbalances. Nonetheless, both of the latter methods occasionally suffer from selecting bandwidths that are too narrow, which leads to estimation problems.

Table 5.4: LATE RMSE results

LATE RMSE									
	ICC= 0.00			*ICC*= 0.20			*ICC*= 0.70		
	General h	Average h	Weighted average h	General h	Average h	Weighted average h	General h	Average h	Weighted average h
Balanced									
σ^2_{a1} =0.00	0.001	0.001	0.001	0.000	0.000	0.001	0.003	0.002	0.002
σ^2_{a1} =0.20	0.008	0.008	0.008	0.004	0.004	0.004	-0.005	-0.005	-0.005
σ^2_{a1} =0.60	0.012	0.012	0.012	0.006	0.006	0.006	0.002	0.003	0.003
Unbalanced									
σ^2_{a1} =0.00	0.000	0.000	0.000	0.001	0.001	0.001	0.003	0.004	0.004
σ^2_{a1} =0.20	0.004	0.003	0.004	0.006	0.007	0.008	0.002	0.001	0.001
σ^2_{a1} =0.60	0.004	0.005	0.005	0.008	0.010	0.011	0.001	0.003	0.003

5.3 Application to the INVALSI data

In the following section, we apply the insights from the simulation study to the INVALSI case introduced in Chapter 3. The goal is to examine whether—and to what extent—the LATE varies across the 20 Italian regions. Table 5.5 reports the number of observations, mean, median, and coefficient of variation for both the outcome (*Math score*) and the running variable (*ESCS school*) for each of the 20 regions. As already highlighted in Chapter 3, there is a marked disparity in both the mean and the median between the southern regions and the rest of the country. The relatively high coefficients of variation—especially for the outcome—further indicate substantial variability both between and within regions.

Table 5.5: Analysis of Math score and ESCS school across regions

Region	**N**	**Math score Mean**	**Math score Median**	**Math score CV**	**ESCS school Mean**	**ESCS school Median**	**ESCS school CV**
Valle D'Aosta	851	199.68	196.18	6.19	0.06	0.02	0.14
Piemonte	30,002	200.66	197.93	5.68	0.06	-0.01	0.13
Liguria	10,128	195.92	192.77	5.55	0.11	0.18	0.34
Lombardia	70,343	205.74	203.34	5.90	0.16	0.06	0.21
Trentino Alto Adige	7,448	208.58	206.60	5.83	0.21	0.16	0.60
Veneto	36,513	206.98	204.75	5.78	0.03	-0.04	0.07
Friuli Venezia Giulia	8,366	206.81	204.75	6.08	0.12	0.05	0.14
Emilia Romagna	32,931	200.48	200.35	5.41	0.11	0.06	0.20
Toscana	27,463	197.17	195.03	5.33	0.09	0.01	0.23
Umbria	6,898	195.91	193.47	5.40	0.13	0.01	0.28
Marche	12,765	196.04	193.94	5.48	0.10	0.02	0.22
Lazio	46,658	187.66	185.23	5.03	0.20	0.18	0.41
Abruzzo	10,048	191.09	188.59	5.35	0.10	0.02	0.23
Molise	2,369	190.74	188.17	5.13	0.05	-0.05	0.12
Campania	67,068	176.13	172.75	4.64	-0.27	-0.24	0.31
Puglia	32,068	188.98	186.76	4.90	-0.16	-0.16	0.31
Basilicata	4,624	190.80	188.89	5.00	-0.07	-0.15	0.18
Calabria	15,509	180.00	177.61	5.26	-0.04	-0.14	0.09
Sicilia	40,335	181.09	179.27	5.29	-0.14	-0.17	0.27
Sardegna	10,584	178.18	175.90	5.27	-0.09	-0.04	0.22

In light of the simulation study results, we chose to apply the *General bandwidth method* to the real data, as it displayed overall more stable performance in terms of bias, variance, and RMSE—particularly in scenarios that most closely resemble the empirical configuration of the dataset, namely those characterized by a medium-to-low ICC. Indeed, to support this choice, the ICC was calculated and found to be 0.179. Furthermore, compared to approaches based on the simple or weighted average of region-specific bandwidths, the General bandwidth method is also computationally less demanding, as it requires estimating only a single regression discontinuity model on the full sample, rather than 20 separate models—one for each region.

This procedure makes it possible to combine the advantages of local RD estimation with those of hierarchical modelling: on the one hand, isolating the causal effect at the discontinuity margin, and on the other, accounting for the nested structure of the data (students within regions).

The analyses were carried out using the `lme4` package (Bates et al., 2015). Following the analytical strategy of the General bandwidth method, we first implemented the Regression Discontinuity design using, as in previous chapters, the MSE-optimal bandwidth selection algorithm, which chooses the bandwidth that minimizes the MSE while accounting for the bias–variance trade-off (see Section 2.2.1). After selecting a bandwidth of 0.131 to the left and 0.131 to the right of the cutoff (equal to 0 due to the centering of the running variable), we constructed a triangular weighting function for the observations within this interval. Finally, as the last step of our analytical strategy (see Table 5.1), we implemented multilevel regression. Specifically, we estimated three different models:

- Random intercept model: the intercept term is allowed to vary across the regions
- Random slope model: the slope term is allowed to vary across the regions
- Random intercept and slope model: both the intercept and the slope terms are allowed to vary across the regions.

We compared the fully random model specified in Equation 5.3 with the two simpler versions. The best model selection was performed based on information criteria like the Akaike Information Criterion (AIC) (Akaike, 1973; Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978), as well as the log-likelihood value and the Likelihood Ratio Test (LRT), as suggested in the literature for comparing multilevel models (Snijders and Bosker, 1999; Snijders and Bosker, 2012). Results reported in Table 5.6 suggest that the Random intercept and slope model is preferred for our data, indicating that both intercept and slope (the treatment) significantly vary across regions. Specifically, in this model, the fixed effects are equal to $\beta_0 = 184.56$ for the intercept, $\beta_1 = -30.37$ for the centered running variable $(X_{ij} - c)$ and $\beta_2 = -1.07$ for the treatment effect. Hence, overall, students enrolled in a school with a low ESCS experience a negative treatment, resulting in a loss of 1.07 points in the Math score. Concerning random effects, the variances of level 2 (regions) residuals are equal to $\sigma^2_{u_0} = 88.65$ and $\sigma^2_{u_2} = 53.05$.

Figures 5.5 and 5.6 show the caterpillar plots, which allow us to visualise the estimated random effects (intercepts and slopes) for each higher-level unit (region) in the model.

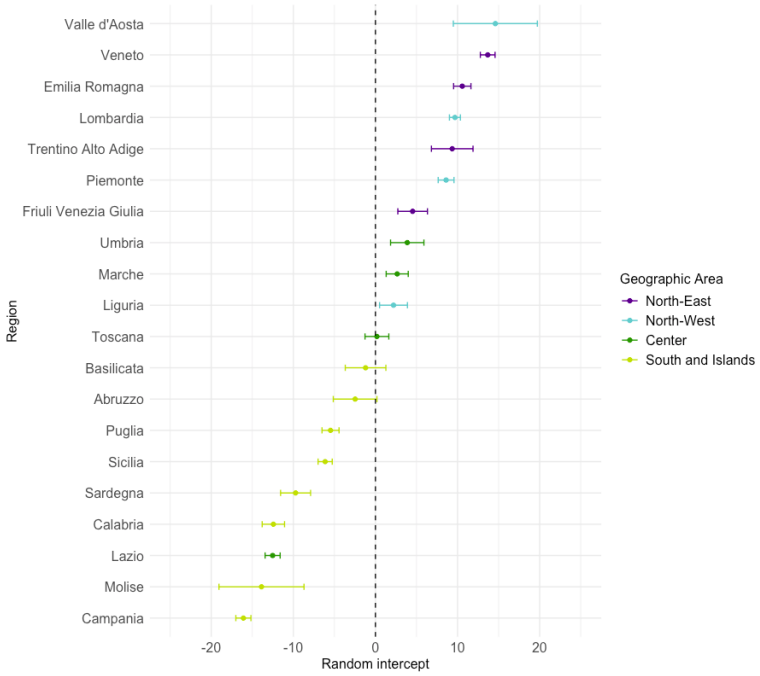


Figure 5.5: Random intercept: deviation of each region from the overall intercept

Note. 0 corresponds to the overall intercept β_0 equal to 184.56.

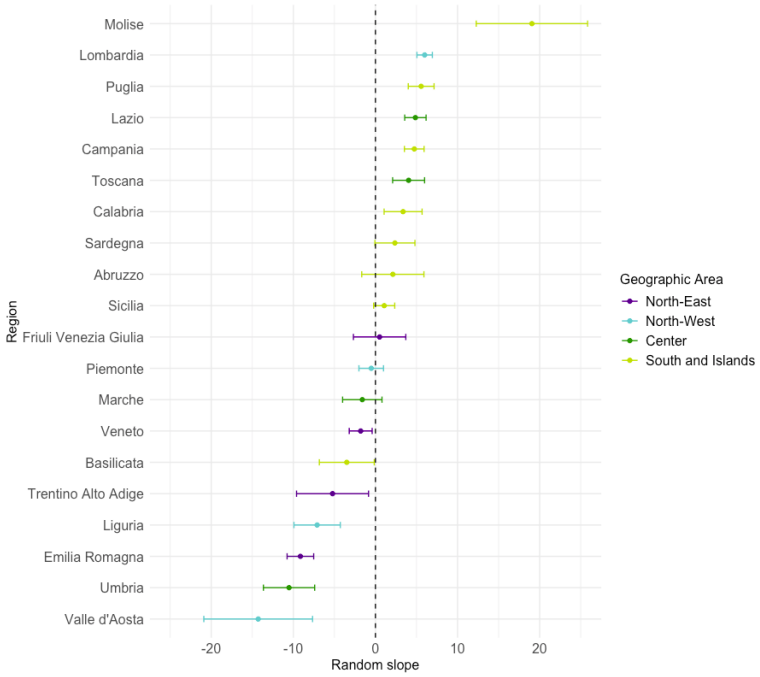


Figure 5.6: Random slope: deviation of each region from the overall slope

Note. 0 corresponds to the overall slope β_2 equal to -1.07.

Table 5.6: Selection of the best multilevel model

Model	**AIC**	**BIC**	**LogLik**	χ^2	**P-value**
Random intercept model	855,750	855,778	$-427,872$		
Random slope model	855,280	855,327	$-427,635$	473.92	<0.001
Random intercept and slope model	854,827	854,892	$-427,406$	457.48	<0.001

Figure 5.5 displays the deviations of each region from the overall intercept β_0, which is equal to 184.56. Regions to the right of the plot show positive deviations from β_0, while those on the left show negative deviations. Most northern regions—particularly Valle d'Aosta, Veneto, Emilia-Romagna, Lombardy, Trentino-Alto Adige and Piedmont—exhibit significantly higher values than the national average. This implies that, holding school ESCS constant, students in these regions achieve higher-than-average Math scores. Conversely, the southern regions—such as Campania, Molise, Calabria, Sardinia, Sicily and Apulia—show negative deviations from the intercept, indicating structurally lower academic performance. This North–South divide confirms, in line with the literature and with what was discussed in previous chapters, the persistence of substantial territorial disparities in learning outcomes, likely linked to differences in school resources, quality of educational provision, family and social capital, and the effectiveness of regional education policies.

The second caterpillar plot, shown in Figure 5.6, displays the deviations of each region from the average treatment effect β_2, which is equal to -1.07. Since the overall treatment effect is negative, the random effects u_{2j} indicate whether the negative impact of attending a low-ESCS school is more or less intense than the national average. The results are highly revealing: in almost all northern and central regions, this effect is even more negative than in the South, suggesting that a disadvantaged school context has a stronger impact on student performance in the Centre–North. Some interesting exceptions emerge for Lombardia, Toscana and Lazio, where deviations are positive: in these regions, the negative effect of low ESCS is attenuated, indicating a greater capacity of the school system to mitigate internal inequalities. By contrast, in many southern regions the marginal effect of low ESCS, deriving from positive residual values, is less pronounced, implying that students in the South do not experience an additional sharp drop in performance when attending schools with low socioeconomic composition.

These findings reveal an apparent paradox: average performance is lower in the South, yet the negative peer effect is stronger in the Centre–North. This can be interpreted in at least two ways. A first, structural interpretation suggests that

southern schools already operate in generally more disadvantaged environments; therefore, the additional effect of low ESCS is embedded within a broader context of deprivation and is relatively less perceptible. A second, adaptive or resilience-based interpretation suggests that students in the South may have developed coping strategies in response to widespread disadvantage, thereby reducing the incremental impact of school-level deprivation. See Figure 3.1 in Chapter 3 for evidence of the structural ESCS disadvantage affecting southern regions. In contrast, in northern regions—characterised by stronger-performing education systems and higher average standards—attending a low-ESCS school is a more exceptional and penalising condition, generating clearer and more substantial negative effects on performance.

In conclusion, by combining Regression Discontinuity analysis with Multilevel Modelling, we show that the peer effect of attending schools with low ESCS—according to the threshold set by the Ministry of Education and Merit—is characterised by substantial heterogeneity across Italian regions.

Conclusions

In conclusion, the objective of this thesis was to address several methodological challenges related to Regression Discontinuity (RD) designs by proposing innovative solutions and applying them to real-world data. The work pursued a twofold aim: on the one hand, to contribute to the methodological debate on RD in the presence of discrete running variables and heterogeneous contexts; and on the other, to provide empirical evidence on the role of the school socio-economic environment in shaping Italian students' mathematics performance, with a particular focus on territorial differences between the North, Centre, and South of the country.

From an empirical perspective, the case study focused on data from the INVALSI tests administered to grade 13 (final-year high school) students between March and May of the 2023/2024 academic year. The goal was to assess the influence of school socio-economic status on students' Mathematics outcomes. To this end, a RD framework was employed, where the running variable was the school-level *Economic, Social, and Cultural Status (ESCS)*, the outcome was the individual Math score, and the treatment threshold was set by the Ministry of Education and Merit in Ministerial Decree No. 90 of 19 May 2023 at -0.31243 for upper secondary schools. Below this threshold, a school is classified as low-ESCS. The objective, therefore, was to estimate the peer effect of attending a school classified as low-ESCS under this policy rule.

The empirical analysis yielded a negative treatment effect of -1.079, indicating that attending a low-ESCS school has a detrimental impact on Math scores. However, given that the outcome ranges between 57.99 and 306.64, this effect is economically very small and can be considered negligible. Socio-economic composition thus emerges as a real but non-dominant factor in explaining mathematics performance and, more generally, school achievement. The peer effect arising from attendance at low-ESCS schools is statistically detectable, yet modest in magnitude and not the primary driver of performance differentials. This implies that effective strategies to reduce learning gaps cannot be limited to altering student composition. Instead, they must combine careful management of social composition with targeted instructional interventions and systemic educational policies. In other words, peer effects exist but are limited: the decisive policy lever lies in improving classroom practices and providing stronger support to low-ESCS schools, for example by enhancing teaching quality or implementing targeted support programmes.

From a methodological perspective, the thesis first addressed the issue of RD

designs with discrete running variables and the presence of so-called *mass points*, i.e. multiple individuals sharing the same value of the running variable. The presence of mass points reduces the effective sample size. Local polynomial methods—the most widely used approach in RD—require sufficient variation in the running variable to construct functions on either side of the threshold; what truly matters, therefore, is not the number of individual observations but the number of unique values of the running variable. In RD, local polynomial estimators behave as though the effective sample size is determined by the number of mass points rather than the number of observations. This phenomenon is referred to as *sample size inflation*. The literature suggests avoiding this issue by aggregating the data at the level of the mass points using the mean outcome—thereby producing a new dataset with a number of observations equal to the number of mass points.

This is where the first methodological contribution of the thesis is introduced: in many cases, summarising a group of individuals sharing the same running variable using the mean provides an incomplete picture, potentially masking outliers, skewness, or heavy tails. Therefore, when the distribution of the outcome within mass points is asymmetric and affected by outliers, alternative collapsing strategies—such as the median and the medoid—are preferable. The median is robust to extreme values, while the medoid is particularly suitable when qualitative covariates must also be summarised within mass points, as it originates from cluster analysis with mixed data and is likewise minimally influenced by outliers.

A simulation study was conducted to reinforce this proposal, allowing sample size, proportion of mass points, and outcome distribution to vary, and comparing normal and skewed distributions. The real aim was to provide guidance on which collapsing method to adopt depending on data characteristics. The analysis of the estimated coefficients, Rbias, and RMSE showed that: in normally distributed scenarios, all three methods (mean, median, medoid) are unbiased; in skewed scenarios, robust methods (median/medoid) clearly outperform the mean in terms of stability and accuracy.

In the empirical application, the running variable, the ESCS of the school, was discrete. This issue, initially neglected, was later addressed by collapsing the data using the median, as preliminary analysis revealed high asymmetry and the presence of numerous outliers within mass points. Accounting for the discrete running variable reinforced earlier conclusions and avoided sample size inflation: the raw-data estimate of -1.071 was reduced to -0.425 after collapsing on the median—similar in sign, but no longer statistically significant. This suggests that, once data discreteness and outcome asymmetry are appropriately addressed, the peer effect on

Math scores becomes even weaker—further supporting the interpretation that school socio-economic composition is a real, though not predominant, factor.

The thesis concludes with a second methodological contribution, aimed at studying RD in contexts where treatment effects may be influenced by heterogeneity and hierarchical data structures. For this reason, the Multilevel regression framework was introduced. The rationale is straightforward: RD allows estimation of the Local Average Treatment Effect (LATE), while multilevel models allow exploration of how such effects vary across hierarchical levels (e.g. students nested within schools or regions). Combining the two approaches enables estimation of both the average treatment effect and its variation across higher-level grouping variables.

The proposed framework advances existing literature on multilevel RD by incorporating core RD procedures such as optimal bandwidth selection and kernel weighting. Three integration strategies were introduced—*General bandwidth*, *Average bandwidth*, and *Weighted average bandwidth*—which differ in how the optimal bandwidth is selected and weighted. A simulation study varying the heterogeneity of the LATE, the intraclass correlation (ICC), and group composition (balanced vs. unbalanced groups) was conducted. The methods were evaluated in terms of LATE Bias, LATE Heterogeneity Bias, bandwidth distribution, LATE Relative Bias (Rbias), and LATE Root Mean Square Error (RMSE).

Overall, the General bandwidth method proved to be the most versatile, achieving a good compromise between bias, variance, and stability, while consistently providing group-level coefficient estimates.

Finally, the selected method was applied to the INVALSI data to address potential LATE heterogeneity across regions—the higher-level grouping variable. The results confirmed significantly lower average Math scores in southern regions, consistent with the well-documented territorial divide in Italy, but also revealed greater sensitivity to negative peer effects in central and northern regions. This stronger treatment effect in the Centre–North was interpreted through two lenses: structurally, southern schools already operate in systematically more disadvantaged contexts, making the additional penalty of low ESCS less visible; adaptively, students in southern regions may have developed resilience strategies in response to widespread disadvantage. Conversely, in the Centre–North—where school systems are generally more effective and expectations higher—attending a low-ESCS school represents a more atypical and penalising condition, producing sharper negative effects.

Future developments will concern both the applied and methodological dimensions of this research. On the applied side, since the results reveal the presence of a modest peer effect, future investigations will focus more deeply on what hap-

pens within classrooms. In particular, it will be of interest to explore how student interaction dynamics, the quality of teaching practices, and teachers' professional preparation contribute to explaining differences in educational outcomes. As discussed, the most effective policy lever for reducing educational disparities does not lie solely in altering the socioeconomic composition of schools, but rather in strengthening internal educational processes—through the enhancement of teaching quality and the implementation of targeted support programs for low-ESCS schools. In this perspective, a particularly significant role is also played by the surveys conducted by INVALSI at the teacher level, such as the Mathematics Teacher Questionnaire[1], which collects information on teaching practices, instructional strategies, and the use of technology in the classroom. Integrating these data with the results of standardized tests would make it possible to deepen the understanding of the mechanisms linking teaching quality to student outcomes, thereby providing additional tools to guide educational policies aimed at improving instructional effectiveness and reducing territorial inequalities.

From a methodological perspective, future developments will aim at refining the proposed approaches. Concerning the case of Regression Discontinuity with a discrete running variable, a next step will involve incorporating qualitative covariates into the simulation study to better assess the use of medoids as a data-collapsing strategy. With regard to the extension of the Multilevel approach to the RD framework, future research could focus on constructing a unified model—a *single-step approach*—capable of coherently and simultaneously integrating the various stages envisaged in the three proposed methods, with the aim of improving the precision and stability of the estimates.

[1] https://serviziostatistico.invalsi.it/invalsi_ss_data/microdati-campione-g13-2023-24-questionario-insegnante/

Bibliography

Abdullah, A., Doucouliagos, H., and Manning, E. (2015). "Does Education Reduce Income Inequality? A Meta-Regression Analysis". In: *Journal of Economic Surveys* 29.2, pp. 301–316.

Agasisti, T. (2014). "The efficiency of public spending on education: An empirical comparison of EU countries". In: *European Journal of Education* 49.4, pp. 543–557.

Agasisti, T. and Vittadini, G. (2012). "Regional Economic Disparities as Determinants of Students' Achievement in Italy". In: *Research in Applied Economics* 4, pp. 33–54.

Aitkin, M. and Longford, N. (1986). "Statistical modelling issues in school effectiveness studies". In: *Journal of the Royal Statistical Society: Series A (General)* 149.1, pp. 1–26.

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle". In: *Proceedings of the 2nd International Symposium on Information Theory*. Ed. by Boris N. Petrov and Frigyes Csáki. Budapest: Akadémiai Kiadó, pp. 267–281.

– (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Anderson, M. L. (2014). "Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion". In: *American Economic Review* 104.9, pp. 2763–2796.

Angrist, J. D. and Krueger, A. B. (2001). "Instrumental variables and the search for identification: From supply and demand to natural experiments". In: *Journal of Economic perspectives* 15.4, pp. 69–85.

Angrist, J. D. and Lavy, V. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". In: *Quarterly Journal of Economics* 114.2, pp. 533–575.

Angrist, J. D. and Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, pp. 251–252.

Auffhammer, M. and Kellogg, R. (2011). "Clearing the Air? The Effects of Gasoline Content Regulation on Air Quality". In: *American Economic Review* 101.6, pp. 2687–2722.

Bates, Douglas, Mächler, Martin, Bolker, Ben, and Walker, Steve (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Becker, S. O., Egger, P. H., and Ehrlich, M. von (2013). "Absorptive Capacity and the Growth and Investment Effects of Regional Transfers: A Regression Discontinuity Design with Heterogeneous Treatment Effects". In: *American Economic Journal: Economic Policy* 5.4, pp. 29–77.

Black, D. A., Galdo, J., and Smith, J. (2007). "Evaluating the Bias of the Regression Discontinuity Design Using Experimental Data". Mimeo, University of Chicago.

Bond, T. G. and Fox, R. (2007). *Applying The Rasch Model, Foundamental Measurement in the human Sciences*. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Borman, G. D. and Dowling, M. (2010). "Schools and Inequality: A Multilevel Analysis of Coleman's *Equality of Educational Opportunity* Data". In: *Teachers College Record* 112.5, pp. 1201–1246.

Buddelmeyer, H. and Skoufias, E. (2004). *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA*. Tech. rep. 827. World Bank Publications.

Burger, N. E., Kaffine, D. T., and Yu, B. (2014). "Did California's Hand-Held Cell Phone Ban Reduce Accidents?" In: *Transportation Research Part A: Policy and Practice* 66, pp. 162–172.

Calonico, S., Cattaneo, M. D., Farrel, M. H., and Titiunik, R. (2019). "Regression discontinuity designs using covariates". In: *Review of Economics and Statistics* 101.3, pp. 442–451.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). "On the effect of bias estimation on coverage accuracy in nonparametric inference". In: *Journal of the American Statistical Association* 113.552, pp. 767–779.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). "Robust nonparametric confidence intervals for regression-discontinuity designs". In: *Econometrica* 82.6, pp. 2295–2326.

Campodifiori, E., Figura, E., Papini, M., and Ricci, R. (2010). *Un indicatore di status socio-economico-culturale degli allievi della quinta primaria in Italia*. Invalsi.

Card, D. and Krueger, A. B. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". In: *American Economic Review* 84.4, pp. 772–793.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.

– (2024). *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge University Press.

Cattaneo, M. D., Keele, L., and Titiunik, R. (2023). "Covariate adjustment in regression discontinuity designs". In: *Handbook of matching and weighting adjustments for causal inference*, pp. 153–168.

Cattaneo, M. D. and Titiunik, R. (2022). "Regression Discontinuity Designs". In: *Annual Review of Economics* 14, pp. 821–851.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2016). "Inference in regression discontinuity designs under local randomization". In: *The Stata Journal* 16.2, pp. 331–367.

– (2018). *rdlocrand: Local randomization methods for rd designs*. R package version 0.3.

Chen, X., John, G., Hays, J. M., Hill, A. V., and Geurs, S. E. (2009). "Learning from a Service Guarantee Quasi Experiment". In: *Journal of Marketing Research* 46.5, pp. 584–596.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). "Inference on Counterfactual Distributions". In: *Econometrica* 81.6, pp. 2205–2268.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: Government Printing Office.

Cook, T. D. and Wong, V. C. (2008). "Empirical Tests of the Validity of the Regression Discontinuity Design". In: *Annales d'Economie et de Statistique*, pp. 127–150.

De Paola, M. and Scoppa, V. (2010). "Peer group effects on the academic performance of Italian students". In: *Applied Economics* 42.17, pp. 2203–2215.

Dong, Y. (2015). "Regression Discontinuity Applications with Rounding Errors in the Running Variable". In: *Journal of Applied Econometrics* 30.3, pp. 422–446.

Elster, J. (1983). *Explaining technical change: A case study in the philosophy of science*. Cambridge, UK: Cambridge University Press.

Epple, D. and Romano, R. E. (2011). "Peer effects in education: A survey of the theory and evidence". In: *Handbook of social economics* 1.11, pp. 1053–1163.

Forastiere, L., Airoldi, E. M., and Mealli, F. (2006). "Identification and estimation of treatment and interference effects in observational studies on networks". In: *Journal of the American Statistical Association* 116.534, pp. 901–918.

Frandsen, B. R. (2017). "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design when the Running Variable is Discrete". In: *Regression Discontinuity Designs*. Vol. 38. Emerald Group Publishing Limited, pp. 281–315.

Frandsen, B. R., Frölich, M., and Melly, B. (2012). "Quantile Treatment Effects in the Regression Discontinuity Design". In: *Journal of Econometrics* 168.2, pp. 382–395.

Frölich, M. and Martin, H. (2019). "Including covariates in the regression discontinuity design". In: *Journal of Business & Economic Statistics* 37.4, pp. 736–748.

Furnée, C. A., Groot, W., and Den Brink, H. M. van (2008). "The Health Effects of Education: A Meta-Analysis". In: *European Journal of Public Health* 18.4, pp. 417–421.

Goldstein, H. (2011). *Multilevel Statistical Models*. 4th ed. Wiley.

Grainger, C. A. and Costello, C. J. (2014). "Capitalizing Property Rights Insecurity in Natural Resource Assets". In: *Journal of Environmental Economics and Management* 67.2, pp. 224–240.

Granger-Serrano, A. and Villarraga-Orjuela, A. (2021). "Peer effects on first-year university students' results: The role of classmates' academic performance and socioeconomic status". In: *Mathematics* 9.23, p. 3115.

Grilli, L. and Rampichini, C. (2009). "Multilevel models for the evaluation of educational institutions: a review". In: *Statistical methods for the evaluation of educational services and quality of products*, pp. 61–80.

– (2015). "Specification of random effects in multilevel models: a review". In: *Quality Quantity* 49.3, pp. 967–976.

Gu, X. (2023). "Classmates and friends matter! Peer effects on cognitive ability formation". In: *China Economic Review* 79, p. 101910.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design". In: *Econometrica* 69.1, pp. 201–209.

Hanushek, E. A. (2016). "What Matters for Student Achievement: Updating Coleman on the Influence of Families and Schools". In: *Education Next* 16.2. URL: https://www.educationnext.org/what-matters-for-student-achievement/.

Hanushek, E. A. and Woessmann, L. (2012). "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation". In: *Journal of economic growth* 17.4, pp. 267–321.

Harville, D. A. (1977). "Maximum likelihood approaches to variance component estimation and to related problems". In: *Journal of the American statistical association* 72.358, pp. 320–338.

Hausman, C. and Rapson, D. S. (2018). "Regression Discontinuity in Time: Considerations for Empirical Applications". In: *Annual Review of Resource Economics* 10.1, pp. 533–552.

Holland, P. W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

Hox, J., Moerbeek, M., and Schoot, R. van de (2010). *Multilevel Analysis: Techniques and Applications*. 2nd ed. New York: Routledge.

Huber, G. A. and Arceneaux, K. (2007). "Identifying the Persuasive Effects of Presidential Advertising". In: *American Journal of Political Science* 51.4, pp. 957–977.

Imbens, G. W. (2022). "Causality in Econometrics: Choice vs Chance". In: *Econometrica* 90.6, pp. 2541–2566.

Imbens, G. W. and Lemieux, T. (2008). "Regression Discontinuity Designs: A Guide to Practice". In: *Journal of Econometrics* 142.2, pp. 615–635.

Jacob, B. A. and Lefgren, L. (2004). "Remedial Education and Student Achievement: A Regression–Discontinuity Analysis". In: *The Review of Economics and Statistics* 86, pp. 226–244.

Jacob, R., Zhu, P., Somers, M. A., and Bloom, H. (2012). *A Practical Guide to Regression Discontinuity*. Tech. rep. MDRC.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience. DOI: 10.1002/9780470316801.

Keele, L. J. and Titiunik, R. (2015). "Geographic Boundaries as Regression Discontinuities". In: *Political Analysis* 23.1, pp. 127–155.

Krasno, J. S. and Green, D. P. (2008). "Do Televised Presidential Ads Increase Voter Turnout? Evidence from a Natural Experiment". In: *The Journal of Politics* 70.1, pp. 245–261.

Laird, N. M. and Ware, J. H. (1982). "Random-effects models for longitudinal data". In: *Biometrics*, pp. 963–974.

Lamdin, D. J. (1996). "Evidence of Student Attendance as an Independent Variable in Education Production Functions". In: *Journal of Educational Research* 89.3, pp. 155–162.

Lang, C. and Siler, M. (2013). "Engineering Estimates versus Impact Evaluation of Energy Efficiency Projects: Regression Discontinuity Evidence from a Case Study". In: *Energy Policy* 61, pp. 360–370.

Lee, D. S. and Lemieux, T. (2010). "Regression Discontinuity Designs in Economics". In: *Journal of Economic Literature* 48.2, pp. 281–355.

Leeuw, J. de and Meijer, E. (2008). *Handbook of Multilevel Analysis*. New York: Springer.

Litschwartz, S. and Miratrix, L. (2021). *Characterizing Cross-Site Variation in Local Average Treatment Effects in Multisite Regression Discontinuity Design Contexts*

with an Application to Massachusetts High School Exit Exam. EdWorkingPaper No. 21-422. Annenberg Institute for School Reform at Brown University.

Londoño-Vélez, J., Rodríguez, C., and Sánchez, F. (2020). "Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia". In: *American Economic Journal: Economic Policy* 12.2, pp. 193–227.

Ludwig, J. and Miller, D. L. (2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design". In: *The Quarterly Journal of Economics* 122.1, pp. 159–208.

Luyten, H. (2006). "An empirical assessment of the absolute effect of schooling: regression-discontinuity applied to TIMSS-95". In: *Oxford Review of Education* 32.3, pp. 397–429.

Manski, C. F. (1993). "Identification of endogenous social effects: The reflection problem". In: *The review of economic studies* 60.3, pp. 531–542.

McCrary, J. (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test". In: *Journal of Econometrics* 142.2, pp. 698–714.

Meyersson, E. (2014). "Islamic Rule and the Empowerment of the Poor and Pious". In: *Econometrica* 82.1, pp. 229–269.

Mulvey, J. M. and Crowder, H. P. (1979). "Cluster Analysis: An Application of Lagrangian Relaxation". In: *Management Science* 25.4, pp. 329–340.

OECD (1999). *Classifying Educational Programmes: Manual for ISCED-97 Implementation in OECD Countries*. Paris: OECD Publishing.

– (2007). *PISA 2006: Science Competencies for Tomorrow's World. Volume 1: Analysis*. PISA. Paris: OECD Publishing.

– (2008). *PISA 2006: Volume 2: Data*. PISA. Paris: OECD Publishing.

OECD (2016). *PISA 2015 Results: Excellence and Equity in Education (Vol. I)*. OECD Publishing.

– (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do*. PISA. Paris: OECD Publishing.

Papadogeorgou, G., Mealli, F., and Zigler, C. M. (2019). "Causal inference with interfering units for cluster and population level treatment allocation programs". In: *Biometrics* 75.3, pp. 778–787.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, US: Cambridge University Press.

– (2010). "An Introduction to Causal Inference". In: *The International Journal of Biostatistics* 6.2.

Peugh, J. L. (2010). "A practical guide to multilevel modeling". In: *Journal of School Psychology* 48.1, pp. 85–112.

Rao, M. R. (1971). "Cluster Analysis and Mathematical Programming". In: *Journal of the American Statistical Association* 66.335, pp. 622–626.

Rasch, G. (1966). "An individualistic approach to item analysis". In: *Readings in mathematical social science*, pp. 89–108.

– (1980). *Probabilistic models for some Intelligence and attainment tests*. Chicago: University of Chicago Press.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models*. 2nd ed. Thousand Oaks, CA: Sage Publications.

Ripple, C. H. and Luthar, S. S. (2000). "Academic Risk Among Inner-City Adolescents: The Role of Personal Attributes". In: *Journal of School Psychology* 38.3, pp. 277–298.

Sacerdote, B. (2001). "Peer effects with random assignment: Results for Dartmouth roommates". In: *The Quarterly journal of economics* 116.2, pp. 681–704.

Sacerdote, B. (2011). "Peer effects in education: How might they work, how big are they and how much do we know thus far?" In: *Handbook of the Economics of Education* 3, pp. 249–277.

Sannino, P., Davino, C., De Benedictis, L., Romano, R., and Vistocco, D. (2025a). "Investigating the Effect of Economic, Social and Cultural Conditions on Student's Performance". In: *Methodological and Applied Statistics and Demography II*. Ed. by Alessio Pollice and Paolo Mariani. Cham: Springer Nature Switzerland, pp. 586–591. ISBN: 978-3-031-64350-7.

Sannino, P., Davino, C., and Romano, R. (2025b). "Assessing policies for schools with low socioeconomic opportunities: insights from INVALSI tests". Conference abstract. 15th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) (Napoli, Italy, Sept. 8–10, 2025). Abstract presented at the 15th Scientific Meeting of the Classification and Data Analysis Group (CLADAG).

– (2025c). "Evaluating Student Performance: Unveiling the Impact of School's Economic, Social, and Cultural Status with Regression Discontinuity". In: *IES 2025-Innovation Society: Statistics and Data Science for Evaluation and Quality*, pp. 258–265.

Schneeweis, N. and Winter-Ebmer, R. (2007). "Peer effects in Austrian schools". In: *Empirical economics* 32.2, pp. 387–409.

Schwarz, G. (1978). "Estimating the dimension of a model". In: *The annals of statistics* 6, pp. 461–464.

Seyfried, S. F. (1998). "Academic Achievement of African American Preadolescents: The Influence of Teacher Perceptions". In: *American Journal of Community Psychology* 26.3, pp. 381–402.

Sirin, S. R. (2005). "Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research". In: *Review of Educational Research* 75.3, pp. 417–453.

Skovron, C. and Titiunik, R. (2015). "A Practical Guide to Regression Discontinuity Designs in Political Science". In: *American Journal of Political Science*, pp. 1–36.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC Press.

Snijders, T.A. and Bosker, R.J. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.

– (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press.

Steinmann, I. and Olsen, R. V. (2006). "Equal opportunities for all? Analyzing within-country variation in school effectiveness". In: *Large-Scale Assessments in Education* 10.1, p. 2.

Stolberg, H. O., Norman, G., and Trop, I. (2004). "Randomized Controlled Trials". In: *American Journal of Roentgenology* 183.6, pp. 1539–1544.

Sutton, A. and Soderstrom, I. (1999). "Predicting Elementary and Secondary School Achievement with School-Related and Demographic Factors". In: *Journal of Educational Research* 92.6, pp. 330–338.

Swamy, P. A. (1970). "Efficient inference in a random coefficient regression model". In: *Econometrica: Journal of the Econometric Society*, pp. 311–323.

Tan, C. Y., Hong, X., Gao, L., and Song, Q. (2025). "Meta-Analytical Insights on School SES Effects". In: *Educational Review* 77.1, pp. 274–302.

Thistlethwaite, D. L. and Campbell, D. T. (1960). "Regression–Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment". In: *Journal of Educational Psychology* 51.6, pp. 309–317.

Van der Klaauw, W. (2002). "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–Discontinuity Approach". In: *International Economic Review* 43.4, pp. 1249–1287.

Vardardottir, A. (2013). “Peer effects and academic achievement: a regression discontinuity approach”. In: *Economics of Education review* 36, pp. 108–121.

Vinod, H. D. (1969). “Integer Programming and the Theory of Grouping”. In: *Journal of the American Statistical Association* 64.326, pp. 506–519.

White, K. R. (1982). “The Relation Between Socioeconomic Status and Academic Achievement”. In: *Psychological Bulletin* 91.3, p. 461.

Zimmer, R. W. and Toma, E. F. (2000). “Peer effects in private and public schools across countries”. In: *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management* 19.1, pp. 75–92.

Zimmerman, D. J. (2003). “Peer effects in academic outcomes: Evidence from a natural experiment”. In: *Review of Economics and statistics* 85.1, pp. 9–23.