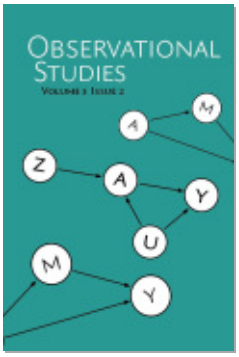
Regression Discontinuity Designs as Local Randomized Experiments

Alessandra Mattei, Fabrizia Mealli



Observational Studies, Volume 3, Issue 2, 2017, pp. 156-173 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2017.0004*

Observational Studies 2 (2016) 156-173 Submitted 10/16; Published 12/16

Regression Discontinuity Designs as Local Randomized Experiments

Alessandra Mattei **mattei@disia.unifi.it**
Department of Statistica, Informatica, Applicazioni
University of Florence
Viale Morgagni 59, 50134 Firenze, Italy

Fabrizia Mealli **mealli@disia.unifi.it**
Department of Statistica, Informatica, Applicazioni
University of Florence
Viale Morgagni 59, 50134 Firenze, Italy

Abstract

In the seminal paper from 1960, Thistlethwaite and Campbell (1960) introduce the key ideas underlying regression discontinuity (RD) designs, which, even if initially almost completely ignored, have then acted as a fuse of a blowing number of studies applying and extending RD designs starting from the late nineties. Building on the original idea by Thistlethwaite and Campbell (1960), RD designs have been often described as designs that lead to locally randomized experiments for units with a realized value of a so-called forcing variable falling around a pre-fixed threshold. We embrace this perspective, and in this discussion we offer our view on how the original proposal by Thistlethwaite and Campbell (1960) should be formalized. We introduce an explicit local overlap assumption for a subpopulation around the threshold, for which we re-formulate the Stable Unit Treatment Value Assumption (SUTVA), and provide a formal definition of the hypothetical experiment underlying RD designs, by invoking a local randomization assumption. A distinguishing feature of this approach is that it embeds RD designs in a framework that is fully consistent with the potential outcome approach to causal inference. We discuss how to select suitable subpopulation(s) around the threshold with adjustment for multiple comparisons, and how to draw inference for the causal estimands of interest in this framework. We illustrate our approach in a study concerning the effects of University grants on students' dropout.

Keywords: Causal Inference, Local Causal Effects, Local Randomization, Potential Outcomes, Regression Discontinuity Designs

1. Introduction

Thistlethwaite and Campbell (1960) are considered to be the fathers of the regression discontinuity (RD) design and they deserve great recognition for their outstanding insight. It is a pleasure and an honor for us to contribute to the discussion on the reprint of their original article on the RD design.

In RD designs, the assignment to the treatment is determined, at least partly, by the realized value of a variable, usually called the forcing or running variable, falling on either side of a prefixed threshold or cutoff point. Thistlewaite and Campbell's key intuition

was that in RD designs the comparisons of units with very close values of the forcing variable, namely around the point where the discontinuity is observed, but different levels of treatment, may lead to valid inference on causal effects of the treatment at the threshold. Nevertheless Thistlethwaite and Campbell (1960) provided no formal description of the design and no theoretical result. In practice, the approach they proposed was a regression analysis with a causal interpretation, and indeed they referred to it as a "RD analysis", rather than a "RD design." It was only later that Campbell (1969) called that type of analysis "a design," but again without giving any formal statistical presentation but only relying on intuitions and analogies to the Fisher's work on design.

Despite Thistlewaite and Campbell's brilliant intuition, RD designs did not attract much attention in the causal inference literature until recently, as the historical excursus in Cook (2008) describes. It is only starting from the late 1990s that RD designs have become increasingly popular in statistics, social science, economics and, more recently also in epidemiology and the medical sciences. In the last two decades, causal inference in RD designs has been a fertile area of research, and there has been a growing number of studies applying and extending RD methods. General surveys can be found in Imbens and Lemieux (2008) and Lee and Lemieux (2010). See also Athey and Imbens (2016) and the edited volume by Cattaneo and Escanciano (2016) for more recent reviews, discussions, and references.

In the modern causal inference literature, inference on causal effects in RD designs uses a formal approach to causal inference rather than the regression framework that was originally used by Thistlethwaite and Campbell (1960). Following one of the main strand of the literature, we will frame RD designs in the context of the potential outcome approach to causal inference (Rubin, 1974; Imbens and Rubin, 2015). See Constantinou and O'Keeffe (2016) for an alternative perspective embedded in the decision theoretic approach to causal inference (Dawid, 2000).

Traditionally, the forcing variable in RD settings is viewed as a pretreatment covariate and RD designs are usually described as quasi-experimental designs with a non-probabilistic assignment mechanism. Therefore inference in RD designs needs to rely on some kind of extrapolation: the traditional inference approach in RD designs invokes smoothness assumptions for the relationship between the outcome and the forcing variable, such as continuity of conditional regression functions (or conditional distribution functions) of the outcomes given the forcing variable. Under these smoothness assumptions, which imply randomization at the single threshold value (Battistin and Rettore, 2008), observations near the known cutoff are used to derive estimates of treatment effects *at the threshold*, using global polynomial series estimators or local-polynomial (non-)parametric regression methods and their asymptotic proprieties. In real applications, large-sample approximations might be unreliable, especially if the sample size around the threshold is small, and exact inference might be preferable. Some further discussion on this traditional approach and its implication for inference is offered in Section 5.

Building on the original idea by Thistlethwaite and Campbell (1960), RD designs have been often described as designs that lead to locally randomized experiments around the threshold (Lee, 2008; Lee and Lemieux, 2010; Dinardo and Lee, 2011). Expanding on this interpretation, a recent strand of the literature (e.g., Cattaneo et al., 2015; Li et al., 2015; Sales and Hansen, 2015) is moving towards a formal and well-structured definition of the con-

ditions under which RD designs can be formally described as local randomized experiments, also discussing the relationship between local randomization and smoothness/continuity RD assumptions (de la Cuesta and Imai, 2016; Skovron and Titiunik, 2015). We embrace this new perspective, to which we have also proudly contributed (Li et al., 2015).

In this discussion we offer our view on how the original proposal by Thistlethwaite and Campbell (1960) should be formalized, that is, how their heuristic reasoning can be formally described. Our view is based on the approach we propose in Li et al. (2015). A distinguishing feature of this approach is that it embeds RD designs in a framework that is fully consistent with the potential outcome approach to causal inference, providing a formal definition of the hypothetical experiment underlying RD designs, based on a description of the assignment mechanism, formalized as a unit-exchangeable stochastic function of covariates and potential outcomes.

We provide a detailed description of this approach, discussing both theoretical and practical issues, and highlighting issues that we feel are valuable topics for further research. We focus on the sharp RD design, the original form of the design, where the treatment status is assumed to be a deterministic step function of the forcing variable: All units with a realized value of the forcing variable on one side of a prefixed threshold are assigned to one treatment regime and all units on the other side are assigned to the other regime. Nevertheless, our methodological framework applies also to fuzzy RD designs, where the realized value of the forcing variable does not alone determine the receipt of the treatment, although a value of the forcing variable falling above or below the threshold acts as an encouragement or incentive to participate in the treatment (see Li et al., 2015, for details on the probabilistic formulation of the assignment mechanism underlying fuzzy RD designs).

2. Our Interpretation of RD Designs as Local Randomized Experiments

Consider a sample or population of N units indexed by $i = 1 \ldots, N$. Let S_i denote the forcing variable, on the basis of which a binary treatment Z_i is assigned according to a RD rule: If a unit has a value of S falling below (or above, depending on the specific application) a predetermined threshold, s_0, that unit is assigned to the active treatment, and s/he is assigned to the control treatment otherwise. Therefore the treatment status Z_i for each unit i is a deterministic function of S_i: $Z_i = \mathbf{1}\{S_i \leq s_0\}$ where $\mathbf{1}\{\cdot\}$ is the indicator function.

Thistlethwaite and Campbell describes the approach they propose arguing that

> The argument [justifying a RD analysis] – and the limitations on generality of the result – can be made more specific by considering a "true" experiment for which the regression-discontinuity analysis may be regarded as a substitute. ... a group of commended students who narrowly missed receiving the higher award might be given opportunity of receiving extra recognition. Thus students in Interval 10 in Figure 1 [in a neighborhood of the threshold, s_0] might be randomly assigned to the different treatment of C of M award and no C of M award (Thistlethwaite and Campbell, 1960, page 310).

We propose to formalize their argument, formally reconstructing the hypothetical "true" experiment underlying a RD design using a framework that is fully consistent with the potential outcome approach. Throughout our discussion we also highlight the key differences

between our approach and both the standard approach to RD designs, where smoothness assumptions are invoked to estimate causal effects at the threshold, as well as alternative, more recent, attempts aiming at formally describing RD designs as local randomized experiments.

Our reconstruction starts from re-defining RD designs step-by-step using the potential outcome approach, which has two essential parts: (a) the definition of the primitive concepts – units, treatments and potential outcomes; and (b) the definition of an assignment mechanism determining which potential outcomes are realized, and possibly observed. Formally, the assignment mechanism is a probabilistic model for the assignment variable as a function of potential outcomes and covariates. The careful implementation of these steps is absolutely essential for drawing objective inferences on causal effects in any study, and thus also in RD designs.

In RD designs, the treatment status, which a unit may be exposed to, depends on the forcing variable, which is the assignment variable. Potential outcomes need to be defined accounting for the alternative levels of the forcing variable and the assignment mechanism needs to be specified as probabilistic model for the conditional probability of the forcing variable given potential outcomes and covariates.

In the literature, the forcing variable is traditionally viewed as a pretreatment covariate and RD designs are typically described as designs with an irregular assignment mechanism breaching the overlap assumption: $Pr(Z_i = 1) = Pr(\mathbf{1}\{S_i \leq s_0\})$ and $Pr(\mathbf{1}\{S_i \leq s_0\}) = \mathbf{1}\{S_i \leq s_0\}$, if S is a fixed pretreatment covariate, and thus the probability of assignment to treatment versus control is equal to zero or one for all units.

We revisit this perspective viewing the forcing variable, S, as a random variable with a probability distribution, and propose to break the longtime interpretation of RD designs as an *extreme* violation of the overlap assumption. Specifically, we formulate the following assumption:

Assumption 1 *(Local overlap). Let $\mathcal{U}$ be the random sample (or population) of units in the study. There exists a subset of units, $\mathcal{U}_{s_0}$, such that for each $i \in \mathcal{U}_{s_0}$, $Pr(S_i \leq s_0) > \epsilon$ and $Pr(S_i > s_0) > \epsilon$ for some sufficiently large $\epsilon > 0$.*

Assumption 1 is essentially a local overlap assumption implying that there exists a subpopulation of units, each of whom has a probability of having a value of the forcing variable falling on both sides of the threshold sufficiently faraway from both zero and one. Assumption 1 implies that each unit belonging to a subpopulation $\mathcal{U}_{s_0}$ has a non-zero marginal probability of being assigned to either treatment levels: $0 < Pr(Z_i = 1) < 1$ for all $i \in \mathcal{U}_{s_0}$. Therefore for units belonging to the subpopulation $\mathcal{U}_{s_0}$, an overlap assumption holds, and this represents a main and key distinction with the traditional description of RD designs. Assumption 1 is a *local* overlap assumption in the sense that a unit with a realized value of the forcing variable falling very faraway from the threshold does not probably belong to the subpopulation $\mathcal{U}_{s_0}$ and may have a zero probability of having a value of the forcing value falling on the other side of the threshold.

It is worth noting that Assumption 1 does not require that the subpopulation $\mathcal{U}_{s_0}$ is unique; it only requires that there exists at least one subpopulation $\mathcal{U}_{s_0}$. Also the value ϵ in Assumption 1 has not a substantive meaning, but it is only a methodological tool for formally describing the subpopulation $\mathcal{U}_{s_0}$.

Assumption 1 plays a key role in the definition of the causal estimands: Under Assumption 1, we can focus on causal effects for a subpopulation, $\mathcal{U}_{s_0}$, rather than on causal effects at the threshold, which are the causal estimands typically considered in RD designs. The correct definition of causal effects depends on the specification of potential outcomes. Each unit in the subpopulation $\mathcal{U}_{s_0}$ can be exposed to alternative values of the forcing variable, therefore, in principle, potential outcomes need to be defined as function of the forcing variable. Let $N_{\mathcal{U}_{s_0}}$ be the number of units belonging to a subpopulation $\mathcal{U}_{s_0}$ and let $\mathbf{s}$ be an $N_{\mathcal{U}_{s_0}}-$dimensional vector of values of the forcing variables with ith element s_i. For each unit $i \in \mathcal{U}_{s_0}$, let $Y_i(\mathbf{s})$ denote the potential outcomes for an outcome variable Y: $Y_i(\mathbf{s})$ is the value of Y for unit i given the vector of values of the forcing variable, $\mathbf{s}$.

Working with the potential outcomes $Y_i(\mathbf{s})$ raises serious challenges to causal inference because the forcing variable is a continuous variable, and so generates a continuum of potential outcomes, and potential outcomes for a unit may be affected by the value of the forcing variable of other units. To face these challenges, within the subpopulation $\mathcal{U}_{s_0}$, we formulate a modified Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1980), specific to RD settings:

Assumption 2 *(Local RD-SUTVA). For each $i \in \mathcal{U}_{s_0}$, consider two treatment statuses $z_i^{'} = \mathbf{1}(s_i^{'} \leq s_0)$ and $z_i^{''} = \mathbf{1}(s_i^{''} \leq s_0)$, with possibly $s_i^{'} \neq s_i^{''}$. If $z_i^{'} = z_i^{''}$, that is, if either $s_i^{'} \leq s_0$ and $s_i^{''} \leq s_0$, or $s_i^{'} > s_0$ and $s_i^{''} > s_0$, then $Y_i(\mathbf{s}^{'}) = Y_i(\mathbf{s}^{''})$.*

Assumption 2 introduces two important simplifications. First, it rules out interference between units, implying that potential outcomes for a unit cannot be affected by the value of the forcing variable (and by the treatment status) of other units. Second, Local RD-SUTVA implies that for units in the subpopulation $\mathcal{U}_{s_0}$, potential outcomes depend on the forcing variable solely through the treatment indicator, z, but not directly, so that, values of the forcing variable leading to the same treatment status define the same potential outcome. The key implication of Assumption 2 is that it allows us to write $Y_i(\mathbf{s})$ as $Y_i(z_i)$ for each unit $i \in \mathcal{U}_{s_0}$, avoiding to define potential outcomes as functions of the forcing variable. Therefore under local RD-SUTVA for each unit within $\mathcal{U}_{s_0}$ there exist only two potential outcomes, $Y_i(0)$ and $Y_i(1)$: they are the values Y if the unit had a value of the forcing variable falling above and below the threshold, respectively.

Local RD-SUTVA is an important limitation and its plausibility depends on the substantive meaning of the forcing variable and on the support of S for each unit. It may be plausible for the subpopulations $\mathcal{U}_{s_0}$, comprising units who have a relatively large probability that the realized values of S fall in a neighborhood around s_0, but it is arguably plausible for the whole study population, and this may be a major obstacle to the generalization of results from RD designs.

Under local RD-SUTVA, causal effects are defined as comparisons of the potential outcomes $Y_i(0)$ and $Y_i(1)$ for a common set of units in $\mathcal{U}_{s_0}$. They are local causal effects in that they are causal effects for units belonging to a subpopulation $\mathcal{U}_{s_0}$. Typical causal estimands of interest in RD designs are average treatment effects. If focus is on the finite population $\mathcal{U}_{s_0}$, then the average treatment effect is the Sample Average Treatment Effect defined as

$$\tau_{s_0}^S = \frac{1}{N_{\mathcal{U}_{s_0}}} \sum_{i \in \mathcal{U}_{s_0}} [Y_i(1) - Y_i(0)]$$

If the $N_{\mathcal{U}_{s_0}}$ units are considered as a random sample from a large superpopulation (where Assumptions 1 and 2 hold), the causal estimand of interest is the Population Average Treatment Effect:

$$\tau_{s_0} = \mathbb{E}\left[Y_i(1) - Y_i(0) | i \in \mathcal{U}_{s_0}\right]$$

Statistical inference for causal effects requires the specification of an assignment mechanism, i.e., the process that determines which units has a value of the forcing variable falling above or below the threshold, and so which potential outcomes are realized and which are missing. In our approach to RD designs the assignment mechanism is a probabilistic model for the conditional probability of the forcing variable given potential outcomes and covariates. The specification of this assignment mechanism is the distinguishing feature of the approach we propose. Specifically, we formalize the concept of a RD design as local randomized experiment invoking the following assumption:

Assumption 3 *(Local randomization) For each* $i \in \mathcal{U}_{s_0}$*,*

$$Pr\left(S_i | \mathbf{X}_i, Y_i(0), Y_i(1)\right) = Pr(S_i)$$

where $\mathbf{X}_i$ *is a vector of pretreatment variables.*

Note that Assumption 3 can be relaxed assuming that local randomization holds conditional on pretreatment variables, and the analysis of RD designs under ignorable assignment mechanisms given covariates is a valuable topic for future research. This is an assumption similar to those considered in Mealli and Rampichini (2012); Angrist and Rokkanen (2015) and Keele et al. (2015).

Assumption 3 implies that for each unit $i \in \mathcal{U}_{s_0}$, $Pr(S_i \leq s_0 | \mathbf{X}_i, Y_i(0), Y_i(1)) = Pr(S_i \leq s_0) = Pr(Z_i = 1)$, which amounts to state that within the subpopulation $\mathcal{U}_{s_0}$ a Bernoulli trial has been conducted, with individual assignment probabilities depending only on the distribution of the forcing variable, not on either the potential outcomes or pretreatment variables. In other words, Assumption 3 implies that the treatment is randomly assigned in some small neighborhood, $\mathcal{U}_{s_0}$, around s_0, formalizing the key idea by Thistlethwaite and Campbell (1960) that a "true" experiment has been conducted in a neighborhood of the threshold (Thistlethwaite and Campbell, 1960, page 310).

3. Inference on Local Causal Effects for a Subpopulation $\mathcal{U}_{s_0}$

3.1 Selection of subpopulations $\mathcal{U}_{s_0}$

Assumptions 1-3 amount to assuming that within subpopulations $\mathcal{U}_{s_0}$ a classical randomized experiment has been conducted, therefore if at least a true subpopulation $\mathcal{U}_{s_0}$ were known, we could draw inference on causal effects for the subpopulation $\mathcal{U}_{s_0}$ using standard methods for analyzing randomized experiments (e.g., Imbens and Rubin, 2015). Unfortunately, in practice, the true subpopulations $\mathcal{U}_{s_0}$ are usually unknown. Therefore an important issue, in practice, is the selection of a subpopulation $\mathcal{U}_{s_0}$.

In principle, a subpopulation may come in any shape or form. Following Li et al. (2015), we limit our choice to symmetric intervals around s_0 for convenience, assuming that for units belonging to a supposedly existing subpopulation $\mathcal{U}_{s_0}$, the realized value of the forcing variable falls in a symmetric interval around the threshold. Formally, we assume:

Assumption 4 *There exists* $h > 0$ *such that for each* $\epsilon > 0$, $Pr(s_0 - h \leq S_i \leq s_0 + h) > 1 - \epsilon$, *for each* $i \in \mathcal{U}_{s_0}$.

Recall that Assumptions 1-3 (and Assumption 4) do not imply that $\mathcal{U}_{s_0}$ has to be unique, therefore we are not interested in finding the largest h, but we only aim at determining plausible values for h.

It is worth noting that the bandwidth choice problem also arises in more conventional RD approaches but for a very different objective. In standard RD approaches, where focus is on estimating causal effects at the threshold, neighborhood selection approaches are usually based on criteria related to local or global polynomial regression methods used to approximate the unknown conditional expectations of the potential outcomes and to obtain an "optimal" extrapolation towards the threshold (see Cattaneo and Vazquez-Bare, 2016, for a review of these methods). In our framework, the objective is to find a subpopulation where Assumptions 1 through 3 are plausible. Consistently the approach for selecting bandwidths h we proposed in Li et al. (2015) exploits Assumption 3. Assumption 3 is a "local" randomization assumption, in the sense that it holds for a subset of units, but may not hold in general for other units. Specifically, Assumption 3 implies that within a subpopulation $\mathcal{U}_{s_0}$ all observed and unobserved pretreatment variables are well balanced in the two subsamples defined by assignment, Z. Therefore, under the assumption that all relevant variables known (or believed) to be related to both treatment assignment and the potential outcomes are observed, within a subpopulation $\mathcal{U}_{s_0}$ any test of the null hypothesis of no effect of assignment on covariates should fail to reject the null. Rejection of the null hypothesis can be interpreted as evidence against the local randomization assumption, at least for the specific subpopulation at the hand. Cattaneo et al. (2015) also exploits balance tests of covariates to select a suitable subpopulation around the threshold, but their approach aims at selecting the largest subpopulation.

Assessing balance in the observed covariates raises problems of multiple comparisons, which may lead to a much higher than planned type I error if they are ignored (e.g., Benjamini and Hochberg, 1995). Cattaneo et al. (2015) prefer to take a conservative approach, by conducting tests for the null hypothesis of balance for each covariate separately, and ignoring the problem of multiplicities. We believe that it may be extremely valuable to account for multiplicities in RD settings, also to avoid to end up with overly small subpopulations.

In the literature, there exist several approaches to tackle the problem of multiple comparisons. From a causal inference perspective, we can use a randomization-based mode of inference, and implement randomization tests adjusted for multiplicities (Lee et al., 2016). As an alternative we can opt for a Bayesian model-based approach, using a Bayesian multiple testing method (e.g., Berry and Berry, 2004; Scott and Berger, 2006). The Bayesian procedure provides a measure of the risk (posterior probability) that a chosen interval around the threshold defines a subpopulation of units that does not exactly matches any true subpopulation, including subjects for which Assumptions 1 through 3 do not hold (see Li et al., 2015).

3.2 Inference

Once subpopulations where Assumptions 1 through 3 are plausible have been selected, we can move to the analysis phase, using any procedure for estimating causal effects from classical randomized experiments, including randomization-based or Bayesian model-based modes of inference.

Randomization inference and Bayesian methods, not relying on asymptotic approximations, are particularly attractive in RD settings where the analysis may rely on a small sample size. Randomization inference provides exact inferences for the finite selected population $\mathcal{U}_{s_0}$, focusing on finite sample causal estimands. From a Bayesian perspective, all inferences are based on the posterior distributions of causal estimands, which are functions of potential outcomes. Therefore inference about sample-average and population-average estimands can be drawn using the same inferential procedures.

A model-based approach requires to specify a model for the potential outcomes. It is worth noting, however, that modeling assumptions play a distinctive role in our setting. They are not necessary and are mainly introduced to adjust for covariates and improve inference: In our setting, model assumptions essentially play the same role as in classical randomized experiments. Conversely, model assumptions are generally crucial in conventional approaches to RD design, where focus is on specifying 'optimal' functional forms relating the outcome to the forcing variable to draw inference on causal effects at the threshold.

Adjusting for both pretreatment variables and the realized values of the forcing variable may be valuable in our approach to RD designs. If the true subpopulations $\mathcal{U}_{s_0}$ were known, in theory, we would not need to adjust for S, because local randomization guarantees that for units in $\mathcal{U}_{s_0}$ values of the forcing variable falling above or below the threshold are independent of the potential outcomes. Nevertheless, in practice, the true subpopulations $\mathcal{U}_{s_0}$ are usually unknown and the risk that a chosen interval around the threshold defines a subpopulation that includes units not belonging to the any true subpopulation, $\mathcal{U}_{s_0}$, is not zero. Systematic differences in the forcing variable S that, by definition, occur between treatment groups may affect inference in the presence of units who do not belong to any subpopulation $\mathcal{U}_{s_0}$. Therefore in order to account for the presence of these units, it might be sensible to conduct inference conditioning on both covariates and the realized values of the forcing variable.

Covariates and forcing variable can be easily incorporated in a Bayesian approach, and they may also help reduce posterior variability of the estimates. Adjusting for S, and possibly for covariates, may be more difficult in randomization-based inference, even if there exist some results in the literature that may be fruitfully exploited in our RD setting (Rosenbaum, 2002; Conti et al., 2014).

4. An Illustrative Example: The Effect of University Grants on Dropout

We illustrate our framework in an example concerning the impact of University student-aid policies on academic careers, using data from the cohort of first-year students enrolled between 2004 to 2006 at University of Pisa and University of Florence (Italy). In Italy, state universities offer grants every year to a limited number of eligible freshmen. In order to get a grant, a student must both meet some eligibility criteria, which are based on an economic indicator of the student's family income and assets falling below or above a

prefixed threshold, as well as apply for the grant. Therefore the grant assignment rule appeals to a RD design, with the economic indicator acting as the forcing variable. Let S be the student's family economic indicator.

In this study, for simplicity, we focus on the effect of eligibility, thus neglecting both the application status and the actual receipt of the grant. The effect of eligibility must be interpreted as an intention-to-treat effect (ITT). The eligibility rule appeals to a *sharp* RD design: Students are eligible if their family economic indicator is below the threshold of 15 000 Euros, and are ineligible otherwise. Therefore for each student i the eligibility indicator is equal to $Z_i = \mathbf{1}\{S_i \leq 15\,000\}$. The outcome variable of primary interest is dropout at the end of the first year. Let $Y_i(z)$ be an indicator for dropout given eligibility status z, and let $Y_i = Y_i(Z_i)$ be the actual dropout indicator observed. In addition, a vector of pretreatment variables, $\mathbf{X}_i$, is observed for each student.

Table 1 presents means for the sample of 16 361 students grouped by eligibility status, Z_i. Eligible freshmen, including students from very low-income families, show different characteristics from ineligible students: on average they have lower high-school grades, and are less likely to come from a science high school and to choose a technical major in University.

We first apply the regression-based approach proposed by Thistlethwaite and Campbell (1960). We divide the forcing variable into evenly-spaced bins and calculate the proportion of students dropping out in each bin. Then, we fit linear regression functions to the observations on either side of the cutoff point, under the assumption that there exists a linear relationship between the outcome (dropout) and the forcing variable.

Figure 1 presents the results. As we can see in Figure 1, there exists a discontinuity at the threshold, which can be interpreted as average treatment effect of eligibility at the threshold according to the original heuristic reasoning of Thistlethwaite and Campbell (1960). The estimate of the ITT effect at the threshold based on the linear regression approach is approximately equal to -0.037%, suggesting that the eligibility reduces dropout for students from families with a value of the economic indicator near the threshold.

Since the publication of Thistlethwaite and Campbell's paper in the early sixties the literature has evolved, and regression or modeling assumptions have been replaced by smoothness/continuity assumptions on the relationship between the outcome and the forcing variable. Table 2 shows estimates of, and 95% confident intervals for, the (population) ITT effects at the threshold derived under the assumption that the conditional distribution functions of the potential outcomes given the forcing variable are continuous in the forcing variable at the threshold. We apply local polynomial estimators, using both a rectangular and a triangular kernel, where the smoothing parameter, the bandwidth, is selected using modern fully data-driven methods, namely, the Coverage Error Rate (CER)-optimal bandwidth proposed by Calonico et al. (2016), used to derive confidence intervals for the average causal effect at the threshold, and two Mean Square Error (MSE)-optimal bandwidths, the Imbens-Kalyanaraman (IK) optimal bandwidth proposed by Imbens and Kalyanaraman (2012) and an upgraded version of it proposed by Calonico et al. (2014). For illustrative purposes, in Table 2 we focus on estimates based on standard local polynomial estimators. Nevertheless, estimates from bias-corrected/robust local polynomial estimators can be also easily applied (see, e.g., Calonico et al., 2014, for details).

Table 1: Italian University Grant Study: Summary Statistics

Variable	All ($n = 16\,361$)	$Z = 0$ ($n = 4\,281$)	$Z = 1$ ($n = 12\,080$)
Assignment variables			
Forcing variable (S)	11148.16	17373.12	8942.12
Grant receipt status (Z)	0.74	0.00	1.00
Outcome variable			
Dropout (Y)	0.38	0.36	0.39
Pre-treatment variables ($\mathbf{X}$)			
Gender	0.60	0.58	0.60
High School Type			
Humanity	0.27	0.26	0.27
Science	0.30	0.36	0.28
Tech	0.39	0.36	0.40
Other	0.05	0.02	0.05
High School grade	81.13	81.94	80.84
Year			
2004	0.40	0.40	0.39
2005	0.34	0.36	0.34
2006	0.26	0.23	0.27
University (Pisa)	0.42	0.39	0.43
Major in University			
Humanity	0.23	0.22	0.23
Social Science	0.26	0.23	0.26
Science	0.13	0.13	0.13
Bio-Med	0.14	0.14	0.14
Tech	0.19	0.22	0.18
Other	0.06	0.06	0.06

Figure 1: Regression of dropout on the forcing variable

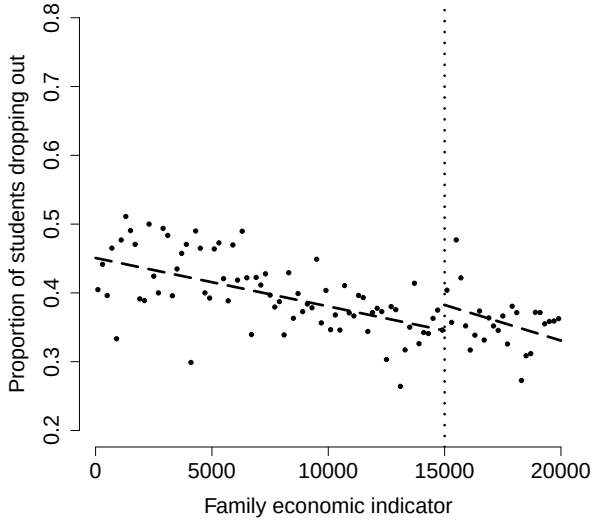


Table 2: Italian University Grant Study: Local Polynomial Estimates of the ITT Effect

Local polynomial	Rectangular Kernel			Triangular Kernel		
regression of order p	τ_{s_0}	*s.e.*	95% CI	τ_{s_0}	*s.e.*	95% CI
CER-optimal bandwidth $= 1\,316.695\,(n = 2\,796)$						
$p = 0$	-0.034	0.018	$[-0.069;\ 0.002]$	-0.045	0.021	$[-0.087; -0.003]$
$p = 1$	-0.066	0.037	$[-0.138;\ 0.006]$	-0.056	0.040	$[-0.134;\ 0.023]$
$p = 2$	-0.039	0.054	$[-0.145;\ 0.067]$	-0.039	0.058	$[-0.152;\ 0.074]$
MSE-optimal bandwidth $= 2\,138.827\,(n = 4\,451)$						
$p = 0$	-0.027	0.014	$[-0.056;\ 0.001]$	-0.032	0.017	$[-0.065;\ 0.001]$
$p = 1$	-0.041	0.029	$[-0.098;\ 0.016]$	-0.057	0.032	$[-0.119;\ 0.005]$
$p = 2$	-0.082	0.043	$[-0.166;\ 0.003]$	-0.068	0.046	$[-0.157;\ 0.022]$
IK optimal bandwidth $= 3\,619.086\,(n = 7\,346)$						
$p = 0$	-0.005	0.011	$[-0.027;\ 0.017]$	-0.022	0.013	$[-0.047;\ 0.004]$
$p = 1$	-0.054	0.022	$[-0.098; -0.011]$	-0.051	0.025	$[-0.099; -0.003]$
$p = 2$	-0.045	0.033	$[-0.111;\ 0.021]$	-0.056	0.036	$[-0.126;\ 0.014]$

As we can see in Table 2, the results are quite sensitive to the choice of both the bandwidth and the polynomial order. The size of the effects changes substantially across different bandwidths, although most of the 95% confidence intervals includes zero. Esti-

mates are also rather unstable across different polynomial orders, especially when the MSE- and IK-optimal bandwidths are used. Nonzero-order polynomials lead to estimate somewhat larger effects than the zero-order polynomial. In some scenario even the choice of the kernel makes a difference. For instance, when the IK-optimal bandwidth and zero-order polynomial are used, the size of the estimate based on the rectangular kernel is about 1/5 of that based on the triangular kernel (-0.005 versus -0.022).

The high sensibility of the inferential results to the critical choices underlying standard RD analyses casts serious doubts on the credibility of the estimates. We argue that these results might strongly rely on extrapolation and model assumptions, especially if the local randomization assumption does not hold for subpopulation of students with a value of the forcing variable falling within a neighborhood defined by some optimal bandwidths, such as the MSE- or IK-optimal bandwidth.

We finally apply the approach we propose, starting by selecting suitable subpopulations $\mathcal{U}_{s_0}$ (see Section 3). We apply randomization-based tests with adjustment for multiplicities to find subpopulations of units where our RD assumptions are plausible. All the covariates listed in Table 1 are considered and we believe that they include all relevant potential confounders.

Table 3 shows randomization-based adjusted p-values for the null hypotheses that the covariates have the same distribution between treated and untreated students for subpopulations defined by various bandwidths, included the optimal bandwidths used in the standard RD analysis. Table 3 also shows $p-$values for the whole sample with S between 0 to 20 000 Euros (column named "ALL") for comparison.

All variables are well balanced for subpopulations defined by bandwidths strictly lower than 1 500. For larger subpopulations some covariates, such as the "indicator of university" are clearly unbalanced. Therefore reasonable subpopulations include students with realized values of the forcing variable within at most 1 500 Euro around the threshold. It is worth noting that only the CER-optimal bandwidth is lower than 1 500 Euro; the MSE- and IK-optimal bandwidths are larger, and define subpopulations where there is clear evidence that covariates are significantly different between eligible and ineligible students. This imbalance justifies, at least partially, the high sensibility of standard RD results to the choice of the bandwidth and model assumptions.

Given the selected subpopulations $\mathcal{U}_{s_0}$, we use a Neyman approach for inference. Table 4 shows estimates of, and 95% confidence intervals (based on the Normal approximation) for the ITT effect for bandwidths ranging from 500 to 1 500 Euros. The estimated ITT effects are similar across different bandwidths: All the estimates are negative, suggesting that eligibility reduces dropout, but most of them are not significant at the 5% level. Only for the subpopulation of students within 1 000 Euros around the threshold, the 95% confidence interval do not cover zero. For this subpopulation the estimated ITT effect of eligibility is a reduction in dropout rate of about 4.7%. The precision of the estimates could be improved adjusting for covariates using a model-based Bayesian approach, which involves model assumptions. Recall that, however, under our framework, identification does not rely on model assumptions; they are only used to improve inference.

Table 3: Italian University Grant Study: Adjusted p-values for the null hypothesis that covariates have the same distribution between eligible and ineligible students for various subpopulations

	Local Randomization Bandwidths						*Local Polynomial Bandwidths*		
Variable	500	1000	1500	2000	5000	All	1316.695	2138.827	3619.086
(Sample size)	(1 042)	(2 108)	(3 166)	(4 197)	(9 846)	(16 361)	(2 796)	(4 451)	(7 346)
Gender	1.000	1.000	1.000	1.000	0.307	0.058	1.000	1.000	0.953
High School Type									
Humanity	1.000	0.999	1.000	1.000	0.973	0.996	1.000	1.000	1.000
Science	1.000	1.000	1.000	0.909	0.001	0.001	1.000	0.686	0.227
Tech	1.000	1.000	1.000	1.000	0.084	0.001	0.998	1.000	1.000
Other	0.432	0.720	0.402	0.281	0.004	0.001	0.541	0.250	0.081
High School Grade	0.991	1.000	1.000	1.000	1.000	0.001	1.000	1.000	1.000
Year									
2004	1.000	1.000	1.000	1.000	1.000	0.987	1.000	1.000	1.000
2005	1.000	0.943	1.000	1.000	0.847	0.066	0.999	1.000	0.877
2006	1.000	1.000	1.000	1.000	0.788	0.001	1.000	1.000	0.939
University (Pisa)	0.998	1.000	0.117	0.006	0.001	0.001	0.602	0.018	0.004
Major in University									
Humanity	0.965	0.295	0.405	0.910	0.969	0.970	0.562	0.955	1.000
Science	1.000	1.000	0.999	0.998	1.000	1.000	1.000	0.995	0.991
Social Science	1.000	1.000	1.000	1.000	0.998	0.001	1.000	1.000	1.000
Bio-Med	0.995	0.698	0.999	0.999	1.000	1.000	0.992	0.990	1.000
Tech	0.965	0.984	1.000	1.000	0.123	0.001	1.000	1.000	0.717
Other	0.989	1.000	1.000	1.000	0.858	0.993	1.000	1.000	1.000

Table 4: Italian University Grant Study: Estimates of, and 95% Confidence Intervals for, the ITT Effect for various subpopulations $\mathcal{U}_{s_0}$ based on Neyman's approach

Bandwidth	τ^S_{s0}	*s.e.*	95% CI	
500	−0.026	0.030	[−0.085;	0.033]
1000	−0.047	0.021	[−0.088;	−0.005]
1500	−0.020	0.017	[−0.054;	0.014]
CER-optimal bandwidth				
1316.695	−0.034	0.018	[−0.069;	0.002]

5. Discussion

There exist alternative approaches to formalize and analyze RD designs as local randomized experiments. Simultaneously with, but separately from Li et al. (2015), Cattaneo et al. (2015) and Sales and Hansen (2015) propose different sets of assumptions within a neighborhood of the threshold that allow one to exploit a local randomization assumption as an identification and estimation strategy in RD designs. Our approach presents subtle but important differences with the methodological framework proposed by Cattaneo et al. (2015) and Sales and Hansen (2015). In particular, we develop a framework for RD analysis that is fully consistent with the potential outcome approach, by clearly defining the treatments and potential outcomes and separating and defining the critical assumptions.

Sales and Hansen (2015) propose to use regression methods to remove the dependence of the outcome from the forcing variable, and then assume that the transformed version of the outcome is independent of treatment assignment, that is, Z in our notation.

The key assumption in Cattaneo et al. (2015) – named 'local Randomization' – does not actually define an assignment mechanism as the conditional probability of the assignment variable given covariates and potential outcomes, which is the general definition of assignment mechanism in the potential outcome approach to causal inference (Imbens and Rubin, 2015). The local randomization assumption proposed by Cattaneo et al. (2015) has two components. The first component amounts to assuming that the marginal distributions of the forcing variable are the same for all units inside a specific subpopulation. This assumption does not formally define an assignment mechanism but simply implies that the values of the forcing variable can be considered "as good as randomly assigned." The second component requires that potential outcomes depend on the values of the forcing variable only through treatment indicators. We view this assumption as part of SUTVA, that is, as part of the definition of potential outcomes, rather than as an assumption on the assignment mechanism.

The birth of these alternative interpretations and formalizations of a RD designs has raised some discussion on the relationship between local randomization and continuity RD assumptions (e.g., de la Cuesta and Imai, 2016; Skovron and Titiunik, 2015).

It is worth noting that in approaches to RD designs where the forcing variable is viewed as a pre-treatment covariate, the conditional independence assumption trivially holds, but it cannot be exploited directly due to the violation of the overlap assumption. In these settings some kind of extrapolation is required, and in order to avoid that estimates heavily rely on extrapolation, previous analyses focus on causal effects of the treatment for units at the threshold under smoothness assumptions, such as continuity assumptions.

Some authors (de la Cuesta and Imai, 2016; Cattaneo et al., 2015; Skovron and Titiunik, 2015; Sekhon and Titiunik, 2016) argue that the local randomization assumption is not required for the RD design. According to us, this sentence may be misleading and deserves some discussion.

Continuity assumptions and our local randomization assumption are different assumptions: they lead to identify and estimate different causal estimands. Local randomization is not required to identify and estimate causal effects *at the threshold*, the causal effects typically considered in the RD design literature, but it is required to identify and estimate causal effects *around the threshold*.

Although causal effects at the threshold are identified under continuity assumptions, which imply that randomization took place precisely at the threshold, we argue that inference under local randomization may be more robust. Specifically, even if focus is on causal effects at the threshold, and continuity assumptions are invoked for inference, in practice, in any analysis of data we are always forced to actually use information on units that are far away from the threshold, relying on some form of extrapolation. In the literature, the choice of a sub-sample of units in a neighborhood of the threshold is usually based on local or global polynomial regression approximations of the unknown conditional expectations of the potential outcomes given the forcing variable. Recently fully data-driven methods, based on selecting an optimal bandwidth under squared error loss (for the local-linear regression estimator, the local polynomial estimator and generalizations) have become increasingly popular (Imbens and Kalyanaraman, 2012; Calonico et al., 2014). These methods do not guarantee, however, that units with a value of the forcing variable falling above and below the threshold have similar distributions of the covariates. If pre-treatment variables are not well-balanced between units above and below the threshold, inference may be highly sensitive to functional assumptions, such as the choice of a local estimator, that is, the choice of the weights from the kernel defining the local estimator. Conversely, if the local randomization assumption holds, and the neighborhood around the threshold is selected aiming at choosing a sub-sample of units where pre-treatment variables are well-balanced between units above and below the threshold, we expect that inference is robust with respect to model assumptions, including the choice of kernels of local estimators. This is analogous to the result about consistency of regression-based estimates of average causal effects from randomized experiments, where consistency does not rely on the linearity of the relationship between outcome, treatment and covariates (Imbens and Rubin, 2015, Chapter 7).

Under local randomization causal estimands of interest are causal effects for units belonging to a sub-population $\mathcal{U}_{s_0}$, which generally includes units with values of the forcing variable falling in a neighborhood "away" from the threshold. Therefore, under local randomization we can identify and estimate causal effects away from the threshold. Alternative ways to generalize RD results aways from the cutoff point require additional ignorability-type assumptions (e.g., Battistin and Rettore, 2008; Mealli and Rampichini, 2012; Angrist and Rokkanen, 2015). Mealli and Rampichini (2012) combine unconfoundedness and differences-in-differences approaches to extend estimates of causal effects from RD analyses away from the cutoff point. Ways to further exploiting randomization-type assumptions to generalize results from RD analyses away from the threshold are still under investigation.

Acknowledgments

The authors acknowledge financial support from the Italian Ministry of Research and Higher Education through grant Futuro in Ricerca 2012 RBFR12SHVV_003.

References

Angrist, D. J. and Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

Athey, S. and Imbens, G. (2016). The state of applied econometrics - causality and policy evaluation. *ArXiv working paper*, No 1607.00699.

Battistin, E. and Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, 142:715–730.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society - Series B*, 57:289–300.

Berry, S. M. and Berry, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60:418–426.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016). On the effect of bias estimation on coverage accuracy in nonparametric inference. *arXiv Working paper: 1508.02973*, 82.

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

Campbell, D. (1969). Reforms as experiments. *American Psychologist*, 24:409–442.

Cattaneo, M. and Escanciano, J. C. (2016). Regression discontinuity designs: Theory and applications. *Advances in Econometrics*, 38. Emerald Group Publishing. To appear.

Cattaneo, M., Frandsen, B. R., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. and Vazquez-Bare, G. (2016). The choice of neighborhood in regression discontinuity designs. *Observational Studies*, 2:134–146.

Constantinou, P. and O'Keeffe, A. G. (2016). Regression discontinuity designs: A decision theoretic approach. *ArXiv working paper*, No 1601.00439.

Conti, E., Duranti, S., Mattei, A., Mealli, F., and Sciclone, N. (2014). The effects of a dropout prevention program on secondary students' outcomes. *RIV Rassegna Italiana di Valutazione*, 58:15–49.

Cook, T. D. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142:636–654.

Dawid, A. P. (2000). Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Society*, 95:407–448.

de la Cuesta, B. and Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19:375–396.

Dinardo, J. and Lee, D. S. (2011). Program evaluation and research designs. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 4A, pages 463–536. Elsevier Science B.V.

Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3).

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

Imbens, W. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction.* Cambridge University Press, New York, NY, USA.

Keele, L., Titiunik, R., and Zubizarreta, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical. Society, Series A*, 178(1).

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

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 485:281–355.

Lee, J. J., Miratrix, L., Pillai, N. S., and Forastiere, L. (2016). More powerful multiple testing in randomized experiments with non-compliance. *Statistica Sinica*, To appear.

Li, F., Mattei, A., and Mealli, F. (2015). Bayesian inference for regression discontinuity designs with application to the evaluation of italian university grants. *The Annals of Applied Statistics*, 9(4):1906–1931.

Mealli, F. and Rampichini, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *Journal of the Royal Statistical. Society, Series A*, 175(3):775–798.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–304.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (1980). Discussion of "randomization analysis of experimental data in the fisher randomization test" by basu. *Journal of the American Statistical Association*, 75:591–593.

Sales, A. and Hansen, B. B. (2015). Limitless regression discontinuity: Causal inference for a population surrounding a threshold. *ArXiv working paper*, No 1403.5478.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136:2144–2162.

Sekhon, J. S. and Titiunik, R. (2016). On interpreting the regression discontinuity design as a local experiment. *Advances in Econometrics*, 38. Emerald Group Publishing. To appear.

Skovron, C. and Titiunik, R. (2015). A practical guide to regression discontinuity designs in political science. *Working paper. Department of Political Science University of Michigan.*

Thistlethwaite, D. and Campbell, D. (1960). Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.