



LABORatorio R. Revelli
Centre for Employment Studies

Another look at the Regression Discontinuity Design

Erich Battistin
Institute for Fiscal Studies, London

Enrico Rettore
Department of Statistics, University of Padova

Another look at the Regression Discontinuity Design

Erich Battistin
Institute for Fiscal Studies, London

Enrico RETTORE
Department of Statistics, University of Padova*

February 13, 2002

Abstract

The attractiveness of the Regression Discontinuity Design (RDD) rests on its close similarity to a formal experimental design. On the other hand, it is of limited applicability since it is not often the case that units are assigned to the treatment group on the basis of an observable (to the analyst) pre-program measure. Besides, it only allows to identify the mean impact on a very specific sub-population. In this paper we show that the RDD straightforwardly generalizes to the instances in which the units eligibility is established on an observable pre-program measure with eligible units allowed to freely self-select into the program. This set-up also turns out very convenient to build a specification test on conventional non-experimental estimators of the program mean impact. Data requirements are made explicit.

Keywords: program evaluation, second control group, specification tests

1 Introduction

The central issue in evaluating the impact of interventions is to separate their causal effect from the confounding effect of other factors influencing the outcomes of interest.

*Address for correspondence: enrico.rettore@stat.unipd.it

Random assignment of units to the intervention produces treatment and control groups that are equivalent in all respects, except for their exposition to the intervention. Thus, in a completely randomized experiment any post-intervention difference between the two groups by construction doesn't reflect pre-intervention differences. As a result, differences between exposed and control units is entirely due to the intervention itself.

However, in most instances assignment is not controlled by the analyst, hence random assignment is unfeasible. Besides, even in those instances in which in principle the analyst can randomize the assignment, units may not comply with the assigned status and either drop out of the intervention or seek an alternative program (see Heckman and Smith, 1995).

A well-known (and widely used) example of randomized assignment to the treatment group is the US JTPA program, which currently serves close to one million economically disadvantaged people every year (see Friedlander *et al.*, 1997). Random assignment occurs prior to the actual enrollment in the program, but a consistent fraction of those randomized into the treatment group don't participate. For certain components of the JTPA, such non-complying behavior seems to be non-negligible (see, for example, Heckman and Smith, 1998).

In this situation, the ideal experiment is not fully realized since participation turns out (at least partly) voluntary: training is provided only for those individuals who meet certain criteria of need and comply with the result of randomization. It follows that participation depends on observable and unobservable personal characteristics that might be correlated with the outcome of interest. In this situation, differences between treated and control groups with respect to the outcome of interest might be the result of units' self-selection into the intervention.

There are instances in which the so called Regression Discontinuity Design (RDD) arises (see Campbell, 1964, Rubin, 1977, Trochim, 1984). According to this design, assignment is solely based on whether a pre-intervention measure is above/below an established threshold. To fix ideas, consider the case in which a pool of units willing to participate are split into two groups according to whether the pre-intervention measure is above or below a specified threshold. Those who score above the threshold are exposed to the intervention while those who score below are denied it.

This design features both advantages and disadvantages relative to its competitors. On the one hand, in a neighborhood of the threshold for selection a RDD presents some features of a pure experiment. In this sense, it is certainly more attractive than a non-experimental design. By exploiting the fact that subjects assigned to the comparison and the intervention group solely differ with respect to the variable on which the assignment to the in-

tervention is established (and with respect to any other variable correlated to it), one can control for the confounding factors just by contrasting marginal participants to marginal non-participants.

In this context, the term *marginal* refers to those units *not too far* from the threshold for selection. Contrasting marginally treated and marginally control units identifies the mean impact of intervention locally with respect to the threshold for selection. Intuitively, for identifiability to hold it must not be that any spurious discontinuity in the relationship between the outcome and the variable on which selection is based happens to coincide with the cutoff point.

On the other hand, the design features two main limitation. Firstly, its feasibility is by definition confined to those instances in which selection takes place on an observable pre-intervention measure. As a matter of fact, this is not often the case.

Secondly, even when the design is feasible it only identifies the mean impact at the threshold for selection. Which in the presence of heterogeneous impacts tells us nothing about the impact on units away from the threshold for selection. In this sense, we only identify a *local* mean impact of the treatment. To identify the mean impact on the broader population one can only resort to a non-experimental estimator whose consistency for the intended mean impact intrinsically depends on behavioral assumptions.

In this paper we derive two results which to the best of our knowledge are new. Firstly, we show that the range of applicability of the RDD is wider than it has been thought before. It includes all the instances in which the relevant population is split into two subpopulation, eligible and non-eligible units respectively, provided that (i) the eligibility status is established with respect to a continuous variable and (ii) both non-eligible and eligible non-participant units are observable. Then, the mean impact on participant units in a neighborhood of the threshold for eligibility is identified under the standard RDD conditions no matter how eligible units self-select into the program.

Secondly, as a straightforward corollary of the previous result, the selection bias at the threshold for eligibility turns out identifiable. Then, one can formally test whether any of the long array of existing non-experimental estimators is able to compensate for such selection bias. On finding an estimator able to compensate for the selection bias even if only with reference to a particular subpopulation - namely, the units in a neighborhood of the threshold for eligibility - one might feel more confident to use it on the broader population.

Links to related literature are established. In particular, we show that our first result is closely linked to Bloom (1984) and to Angrist and Imbens

(1991). We also stress that our result is closely related to the idea stated by Rosenbaum (1987) of using two alternative comparison groups to better identify a program impact. Lastly, we point out the similarities between our specification test of a non-experimental estimator and the specification tests derived by Heckman and Hotz (1989) as well as the link to the characterisation of the selection bias provided by Heckman *et al.* (1998a).

The remaining of this paper is organized as follows. Section 2 discusses the similarities between a fully randomized experiment and a RDD. Section 3 generalizes the use of a RDD when participation into the treatment group is determined by self-selection. Section 4 shows how to validate the use of non-experimental estimators for the treatment effect using a RDD. Section 5 presents some concluding remarks.

2 Regression Discontinuity Design vs Randomized experiments

This section highlights similarities between a randomized experiment and a RDD.

Following the notation of the potential outcome approach to causal inference (see Rubin, 1974), let (Y_i^T, Y_i^{NT}) be the potential outcomes the i -th subject would experience by taking and not taking part into the program, respectively.

The causal effect of the treatment on a specific unit is then defined as the difference between these two potential outcomes, $\beta_i = Y_i^T - Y_i^{NT}$, which is not observable since being exposed to (denied) the program reveals Y_i^T (Y_i^{NT}) but conceals the other potential outcome.

Let E be the binary variable for the treatment status, with $E = 1$ signaling that the subject takes part into the program. If the assignment is determined by randomization, the treatment status doesn't depend on individual characteristics, hence the following condition holds true

$$(Y^T, Y^{NT}) \perp E. \tag{1}$$

Typically randomization is administered only to people who previously applied for a certain program, who in general are not representative of the overall population (as for the JTPA case in the US). In this situation condition (1) holds with respect to the group of units actually randomized, not with respect to the overall population.

The attractiveness of randomization is that the difference between the mean outcome for treated units and the mean outcome for control units

identifies the mean impact of the program

$$E(Y^T - Y^{NT}) = E(Y^T|E = 1) - E(Y^{NT}|E = 0), \quad (2)$$

since conditioning on E in the right-hand side of (2) is irrelevant by construction. In other words, randomization allows using information on non-participants to identify the mean counterfactual outcome for participants, that is what participants would have experienced had they not entered the program.

Although the RDD lacks random assignment of units to the treatment group, it shares some interesting features with an experimental design. Let S be the random variable according to which units are selected into the treatment and let \bar{s} be the threshold for selection. Units are assigned to the treatment if and only if they score at or above \bar{s} , that is

$$E = \mathbb{1}(S \geq \bar{s}). \quad (3)$$

The probability of selection into the treatment conditional on S is then discontinuous at \bar{s} , stepping from zero to one as S crosses the threshold \bar{s} . Following Trochim (1984), we will refer to this situation as *sharp* RDD, since the status with respect to the program is a deterministic function of an observable pre-program characteristic.

In this context, conditioning on S allows to identify the average impact of the program on subjects scoring \bar{s} , thus a local version of the parameter in (2). In fact, in a neighborhood of \bar{s} this design presents the same features of a ‘pure’ randomized experiment (see Rubin, 1977), since for any positive ε the following condition holds approximately

$$(Y^T, Y^{NT})|S = \bar{s} - \varepsilon \approx (Y^T, Y^{NT})|S = \bar{s} + \varepsilon.$$

Exploiting the relationship between S and E in (3), it follows that the following condition holds true

$$(Y^T, Y^{NT}) \perp E | S = \bar{s}. \quad (4)$$

Because of this property, the RDD is often referred to as a quasi-experimental design (Cook and Campbell, 1979).

In a finite sample for the condition to hold ε needs to go to zero at a proper rate as the sample size grows to infinity, implying a non-standard asymptotic theory for the resulting estimator of the mean impact (see Hahn, Todd and Van der Klaauw, 2001).

Note that to meaningfully define marginal units (with respect to \bar{s}) thus allowing the use of a RDD, S needs to be continuous.

In some cases, units do not comply with the mandated status, dropping out of the program or seeking alternative treatments. Any of these violations of the original assignment might lead to biased conclusions about program effects, since conditions (1) or (4) are no longer valid. The presence of non-complying units in a RDD leads to the so called *fuzzy* RDD, which is not directly relevant in what follows (see Hahn, Todd and Van der Klaauw, 2001 and Battistin and Rettore, 2002).

Two major drawbacks hamper the applicability of RDD. Firstly, in an observational study it is more often the case that units self-select into the treatment rather than being exogenously selected on a pre-program measure. Secondly, even in those instances in which the RDD applies, such a design is not informative about the impact on units away from \bar{s} . These are the two issues we look at in the next sections.

3 A generalization of the Regression Discontinuity Design

Let the *eligibility* to a specific program be determined on the basis of the value taken on by the *continuous* variable S according to the rule (3). If all eligible units participated into the program, the standard RDD would arise and the mean impact on units in a neighborhood of \bar{s} would be identifiable.

In fact, it is a widespread evidence that not all eligible units participate into the program they are eligible for. Across units heterogeneity in the information available on the program, in the individual preferences and in the opportunity costs are likely factors influencing participation in several instances.

As a result of both the eligibility rule and the process leading to participation, the population turns out split into three subgroups: *non-eligibles*, *eligible non-participants* and *participants*. To label these subgroups we introduce a further binary variable to distinguish, amongst units eligible for the treatment, those who actually received it. Let $D = 1$ ($D = 0$) indexes the group of eligible units who participate (do not participate) into the program.

To stress that the eligibility status E affects the actual treatment status we will write D^E . Non-participants are therefore a mixture of those who don't meet eligibility criteria, ($E = 0$), and those who choose not to enter the program, ($E = 1, D^1 = 0$). D^E is a potential variable itself and it is logically defined also for units belonging to the $E = 0$ group.

Carefully note that in the set-up we are considering participation into the program amongst eligible units does not take place by design; it is due to

self-selection.

Let

$$E(Y^T|E = 1, D^1 = 1, S = s) \tag{5}$$

be the mean outcome for eligible units scoring $S = s$ and actually receiving the treatment, with $s \geq \bar{s}$. This quantity is identified exploiting information on the outcome of participants for any given value of S . Let

$$E(Y^{NT}|E = 1, D^1 = 1, S = s) \tag{6}$$

be the counterfactual mean outcome for the same group of units, that is what their response would have been had they not participated. The mean impact of the program on treated units scoring $S = s$ is then defined as the difference between factual and counterfactual results in (5) and (6)

$$\tau(s) = E(Y^T|E = 1, D^1 = 1, S = s) - E(Y^{NT}|E = 1, D^1 = 1, S = s).$$

Accordingly, the mean impact on participants τ is obtained as a weighted average of these quantities, with weights given by the proportion of eligible units scoring $S = s$.

Neither $\tau(s)$ nor τ are directly identifiable, since the counterfactual mean outcome in (6) is not observed by construction. Nor we can replace it by the factual mean outcome observed on eligible non-participants. In fact, due to the self-selection process determining the group of participants ($E = 1$ and $D^1 = 1$) and the group of non-participants ($E = 1$ and $D^1 = 0$), eligible non-participants are not a random sample from the pool of eligible units, implying that in general

$$E(Y^{NT}|E = 1, D^1 = 0, S = s) \tag{7}$$

is different from (6). Note that this result holds true for any given value of S , in particular when $S = \bar{s}$.

Now let information on the outcomes experienced by non-eligible units, ($E = 0$), be available. Since this group of units is by construction characterized by values of S below the threshold for selection \bar{s} , it cannot be used to approximate the counterfactual mean outcomes of participants. Nor we can use non-eligible units in a neighborhood of \bar{s} to approximate the counterfactual mean outcome of participant units in a neighborhood of \bar{s} . The quantity

$$E(Y^{NT}|E = 0, S = \bar{s}) \tag{8}$$

is in fact different from the counterfactual result (6) evaluated at \bar{s} because of the non-random selection of units into the treatment group discussed above.

Non-eligibles alone do not allow to solve the problem. It is the joint use of information on non-eligibles and eligible non-participants to allow solving the problem at least for a particular subpopulation of participants. The key relationship to obtain this result is the following

$$E(Y^{NT}|E = 1, S = \bar{s}) = E(Y^{NT}|E = 0, S = \bar{s}). \quad (9)$$

In a neighborhood of the cutoff point \bar{s} eligible and non-eligible units are nearly alike with respect to S , so that in the counterfactual scenario the two marginal groups would experience the same mean outcome. This result rests on the standard RDD as reviewed in the previous section.

The left-hand side of equality (9) can be written as the weighted mean of the mean outcomes experienced by eligible participants and eligible non-participants, respectively, in a neighborhood of \bar{s}

$$E(Y^{NT}|E = 1, D^1 = 1, S = \bar{s})\phi(\bar{s}) + E(Y^{NT}|E = 1, D^1 = 0, S = \bar{s})[1 - \phi(\bar{s})],$$

where $\phi(\bar{s}) = Pr(D^1 = 1|E = 1, S = \bar{s})$ is the probability of self-selection into the program for units marginally eligible. Substituting the last expression in (9) and solving for $E(Y^{NT}|E = 1, D^1 = 1, S = \bar{s})$ we obtain

$$\begin{aligned} E(Y^{NT}|E = 1, D^1 = 1, S = \bar{s}) &= \frac{E(Y^{NT}|E = 0, S = \bar{s})}{\phi(\bar{s})} \\ &- E(Y^{NT}|E = 1, D^1 = 0, S = \bar{s}) \frac{1 - \phi(\bar{s})}{\phi(\bar{s})} \end{aligned} \quad (10)$$

Namely, the counterfactual mean outcome for participants presenting $S = \bar{s}$ is a linear combination of the factual mean outcome for non-eligible units at $S = \bar{s}$ and of the factual mean outcome for eligible non-participants at \bar{s} . The coefficients of the linear combination add up to one and are function of the probability $\phi(\bar{s})$ which in turn is identifiable. Hence, equation (10) implies that $\tau(\bar{s})$, the mean impact on participants at \bar{s} , is identifiable.

Since subtracting (10) from (5) the mean impact can be expressed as

$$\frac{E(Y|E = 1, S = \bar{s}) - E(Y|E = 0, S = \bar{s})}{\phi(\bar{s})},$$

it can be interpreted as the ratio of the intention to treat effect, the mean impact we would observe if all eligible units actually took part in the program, to the mean impact of E on D at \bar{s} .

Results by Hahn *et al.* (2001) on non-parametric estimation in a RDD apply straightforwardly.

Note that

- condition (9) is the cornerstone on which we build the result. Otherwise stated, it is crucial that in the absence of self-selection amongst eligibles, conditions for the RDD to hold are met.
- to derive the result we don't need to specify how eligible units self-select into the treatment. Thus, identifiability of $\tau(\bar{s})$ doesn't require any behavioral assumption on the process itself.
- to identify $\tau(\bar{s})$ information on three different groups of units, participants, eligible non-participants and non-eligibles, are required.

3.1 Related results

In a fully randomized experiment, Bloom (1984) deals with the case where some units assigned to the program do not actually participate (no-shows). Exploiting information on participants, eligible non-participants and non-eligibles the author proves that the mean impact on participants is identifiable. The result in the previous section can be seen as a special case of Bloom (1984) since according to the condition (4) it is as if randomization took place at the threshold for eligibility \bar{s} . In our case eligible non-participants at \bar{s} play the role of Bloom's (1984) no-shows.

Our result (as well as Bloom's one) can also be derived as a special case of Angrist and Imbens (1991). The authors prove that, even if participation takes place as a result of self-selection, the mean impact on participants is identifiable provided that (i) there exists a random variable Z affecting the participation into the program and orthogonal to the potential outcomes (Y_i^T, Y_i^{NT}) and (ii) the probability of participation conditional on Z is zero for at least one value of Z . Condition (i) qualifies Z as an Instrumental Variable for the problem.

In the Bloom (1984) context, self-selection arises as a consequence of the non-complying behavior of some units randomly assigned to the program. The natural choice for Z in that case is the mandated status as it results from randomization. Condition (i) is satisfied since $Pr(D = 1|Z = 1) > Pr(D = 1|Z = 0)$ and Z is orthogonal to the potential outcomes while condition (ii) is satisfied since $Pr(D = 1|Z = 0) = 0$. In our case, since E is orthogonal to the potential outcomes in a neighborhood of \bar{s} and $Pr(D = 1|E = 0) = 0$, E meets the conditions stated by Angrist and Imbens (1991) in a neighborhood of \bar{s} . Hence the identification of the mean impact on participants at \bar{s} follows.

4 Validating non-experimental estimators of the mean impact on participants

In the previous section we have shown that the existence of an eligibility rule allows to identify the mean impact of an intervention on marginally eligible participants even if participants are self-selected from the eligible pool. If the gain of being treated is heterogeneous with respect to S , such mean impact is not informative on the impact of the intervention on units away from the threshold for eligibility. Nor non-eligible units and eligible non-participants can be used as valid comparison groups, since they differ systematically from participants (the former with respect to S and the latter with respect to the variables driving the self-selection process).

In order to identify the mean impact on the overall population of participants, one has to resort to one of the long array of non-experimental estimators available in the literature which adjust for the selection bias under different assumptions (see Heckman *et al.*, 1999, and Blundell and Costa Dias, 2000, for a review). The problem with these non-experimental estimators is that the assumptions on which their consistency relies most times are not testable.

Over the years the literature took two main routes to deal with this problem. The first route amounts to seek whether any over-identifying restriction on the data generating process arises from a behavioral theory of the phenomenon under investigation, possibly exploiting it to test the assumptions on which the non-experimental estimator rests (see Rosenbaum, 1984 and Heckman and Hotz, 1989).

The second route is feasible only when an experimental design has been run, so that an experimental estimate of the impact comes available. Then, besides estimating the mean impact, one can exploit the experimental set up to study the selection bias and to assess whether the non-experimental estimators are able to reproduce the experimental estimate (see LaLonde, 1986 and Heckman *et al.*, 1998a). When information from a randomized experiment is available, one can meaningfully check how closely non-experimental comparison groups methods approximate experimental impact estimates. At the same time, this allows us to assess the performance of alternative non-experimental estimators for the treatment effect, thus suggesting the best strategy to follow when experimental data are not available.

In this section we show that if the three groups of units are available resulting from the set-up of Section 3, then one can test the validity of any non-experimental estimators on a specific subpopulation. To fix the ideas, we will focus on the well known matching estimator, but the same line of

reasoning applies to other non-experimental estimators.

The key assumption on which the matching estimator rests is that all the variables driving the self-selection process *and* correlated to the outcome are observable to the analyst.

Formally, the assignment to the treatment is told *strongly ignorable* given a set of characteristics x if conditional on x the treatment can be thought as randomly assigned to units provided that at each value x there is a positive probability of being treated

$$(Y^T, Y^{NT}) \perp D | x, \quad 0 < Pr(D = 1 | x) < 1. \quad (11)$$

If this condition holds, then it is as if units were randomly assigned to the treatment with a probability depending on x ; the counterfactual outcome for participants presenting characteristics x can be approximated by the actual outcome of non-participants presenting the same characteristics. Since units presenting x have a common probability to enter the program, then an operational rule to obtain an *ex post* experimental-like data set is to match participants to non-participants on such probability (the so called *propensity score*), whose dimension is invariant with respect to the dimension of x (see Rosenbaum and Rubin, 1983).

The critical assumption of this procedure is that the available x is enough rich to guarantee the orthogonality condition in (11). In principle, this imposes strong requirements on data collection. Moreover, the violation of the second condition in (11) would rise the so called common support problem (see for example Heckman *et al.*, 1998a, and Lechner, 2001).

Let

$$sb(s) = E(Y^{NT} | E = 1, D^1 = 1, S = s) - E(Y^{NT} | E = 1, D^1 = 0, S = s) \quad (12)$$

be the *selection bias* affecting the raw comparison of eligible participants to eligible non-participants. The first term on the right-hand side is a counterfactual mean outcome while the second is a factual one. This quantity captures pre-intervention differences between eligible units self-selected in and out the intervention, respectively, at each level of S , with $S \geq \bar{s}$.

Using the results of the previous section, the mean counterfactual outcome for participants is identifiable in a neighborhood of \bar{s} by means of (10). This also implies that the selection bias for units marginally eligible, $sb(\bar{s})$, is identifiable as the difference between (10) and (7) evaluated at \bar{s} .

Note that the identification of the counterfactual term on the right-hand side of (12) at \bar{s} exploits information on the subgroup of non-eligible units closest to the group of eligible units, thus in a neighborhood of the threshold

for eligibility. Apparently, identification is precluded as S moves away from \bar{s} .

Then, let

$$sb(s, x) = E(Y^{NT}|E = 1, D^1 = 1, x, S = s) - E(Y^{NT}|E = 1, D^1 = 0, x, S = s)$$

be the selection bias on the specific subpopulation indexed by x , where x are the variables advocated to properly account for the selection bias in a matching estimation of the intervention impact. If the orthogonality condition in (11) holds, then

$$sb(s, x) = 0$$

uniformly with respect to x and s . In particular, a necessary condition for the matching estimator to work is that $sb(\bar{s}, x) = 0$, which is directly testable.

Operationally, in a neighborhood of \bar{s} any test of the equality of the mean outcomes of the non-eligible units and of the eligible non-participants, respectively, conditional on x is a test of the strong ignorability of the assignment to intervention, thus a test of the validity of the matching estimator. Clearly, the rejection of the null hypothesis is sufficient to conclude that condition (11) does not hold.

On the other hand, on accepting the null hypothesis one might feel more confident in using the matching estimator but by no means it can be said that the validity of the estimator has been proved. In fact, the specification test tells nothing on whether the strong ignorability condition holds away from \bar{s} .

4.1 Related results

Since the RDD can be seen as a formal experiment at $S = \bar{s}$, the specification test developed above displays a similarity to what Heckman *et al.* (1998a) develop in an experimental set-up. In both cases there is a benchmark estimate of the intervention mean impact - the RDD estimate in the former, the experimental one in the latter - to which the analyst is ready to attach credibility. Then, the analyst tests non-experimental estimators against the benchmark to discover whether the assumptions they rest upon are met.

The similarity between the two approaches stops here. On the one hand, the availability of an experimental set-up as in Heckman *et al.* (1998a) allows to fully characterize the selection bias and to test non-experimental estimators with reference to the population of participants. If a RDD is

available, this is feasible only with reference to the population of participants at $S = \bar{s}$.

On the other hand, it is very often the case that an intervention is targeted to a population of eligible units among which it is actually delivered only to those showing up to participate while it is much less frequent to have available an experimental set-up. Then, the three groups of units needed to implement the results in this paper in principle should be available¹. This opens the door to a routinely application of the specification test based on the RDD as a tool to validate non-experimental estimators of the mean impact on participants.

Rosenbaum (1987) in his discussion of the role of a second control group in an observational study gives an example (example 2 on p. 294) which resembles very closely the set-up we refer to. The Advanced Placement (AP) Program provides high school students with the opportunity to earn college credits for work done in high school. Not all high schools offer the AP program, and in those that do, only a small minority of students participate. Two comparison groups naturally arise in this context, (i) students enrolled in high school not offering the program and (ii) students enrolled in high schools offering the program who did not participate.

Then, Rosenbaum (1987) goes on discussing how the availability of two comparison groups can be exploited to test the strong ignorability condition needed to believe the results of a matching estimator.

Apparently, the first comparison group resembles our pool of non-eligible units while the second comparison group resembles our pool of eligible non-participant units. The main difference between the Rosenbaum's example and our set-up is that in the former case the rule according to which high schools decide whether to offer the AP program or not is unknown to the analyst while in our set-up the eligibility rule is known. It is exactly this feature to allow identifying the mean impact on participants as well as the selection bias at $S = \bar{s}$.

5 Conclusions

The main message from this paper is that every time an intervention is targeted to a population of eligible units but is actually administered to a sub-set of self-selected eligible units, it is worth collecting information separately on *three* groups of units: non-eligibles, eligible non-participants and eligible participants. Also, the variables with respect to which eligibility is established have to be recorded.

¹Whether they are actually available it depends on the design of the data collection.

The relevance of distinguishing between non-eligibles and eligible non-participants to improve the comparability between the treated and the comparison groups has already been stressed in the literature (see, amongst others, Heckman *et al.*, 1998a).

As a complementary result here we have shown that provided the eligibility rule is based on a continuous variable, *jointly* exploiting the two comparison groups the mean impact on participants on the margin between eligibility and non-eligibility is identifiable no matter for how the self-selection of participants takes place.

Then, we have shown that as a straightforward consequence of the previous result also the selection bias for units on the margin between eligibility and non-eligibility is identifiable. This opens up the door to a specification test in a neighborhood of the threshold for eligibility so that the properties of non-experimental estimators can be assessed. By design, such a test is informative on their performance only for a particular subgroup of units, thus results cannot be generalized to the whole population (unless we are willing to impose further identifying restrictions). The value of the specification test is that if it rejects the estimator locally then this is enough to reject it altogether.

References

- [1] Angrist, J.D. and Imbens, G.W. (1991), *Sources of Identifying Information in Evaluation Models*, NBER Technical Working Paper 117
- [2] Battistin, E. and Rettore, E. (2002), *Testing for programme effects in a regression discontinuity design with imperfect compliance*, Journal of the Royal Statistical Society A, Vol. 165, No. 1, 1-19
- [3] Bell, S.H. Orr, L.L. Blomquist, J.D. and Cain, G.G. (1995), *Program Applicants as a Comparison Group in Evaluating Training Programs: Theory and a Test*, Kalamazoo, MI: W.E. Upjohn Institute for Employment Research
- [4] Bloom, H.S. (1984), *Accounting for No-Shows in Experimental Evaluation Designs*, Evaluation Review, Vol. 8, 225-246
- [5] Blundell, R. and Costa Dias, M. (2000), *Evaluation methods for non-experimental data*, Fiscal Studies, Vol. 21, No. 4, 427-468
- [6] Campbell, D.T. (1964), Reforms as experiment,

- [7] Cook, T.D. and Campbell, D.T. (1979), *Quasi-Experimentation. Design and Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company
- [8] Friedlander, D. Greenberg, D.H. and Robins, P.K. (1997), *Evaluating Government Training Programs for the Economically Disadvantaged*, Journal of Economic Literature, Vol. 35, No. 4, 1809-1855
- [9] Hahn, J. Todd, P. and Van der Klaauw, W. (2001), *Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design*, Econometrica, Vol. 69, No. 3, 201-209
- [10] Heckman, J.J. (1996), *Randomization as an instrumental variable*, The Review of Economics and Statistics, Vol. 78, No. 2, pp. 336-341
- [11] Heckman, J.J. and Hotz, V.J. (1989), *Choosing Among Alternative Non-experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training*, Journal of the American Statistical Association, Vol. 84, No. , 862-874
- [12] Heckman, J.J. and Smith, J. (1995), *Assessing the case for social experiments*, Journal of Economic Perspectives, 9 (2), pp. 85-110
- [13] Heckman, J.J. and Smith, J. (1998), *Accounting for Dropouts in Evaluations of social programs*, The review of Economics and Statistics
- [14] Heckman, J.J. Smith, J. and Clements, N. (1997), *Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts*, The Review of Economic Studies, Vol. 64, No. 4, 487-535
- [15] Heckman, J.J. Ichimura, H. Smith, J. and Todd, P. (1998a), *Characterizing Selection Bias Using Experimental Data*, Econometrica, Vol. 66, No. , 1017-1098
- [16] Heckman, J.J. Smith, J. and Taber, C. (1998b), *Accounting for Dropouts in Evaluations of Social Experiments*, The Review of Economics and Statistics, Vol. 80, No. 1, 1-14
- [17] Heckman, J.J. Lalonde, R. and Smith, J. (1999), *The Economics and Econometrics of Active Labor Market Programs*, Handbook of Labor Economics, Volume 3, Ashenfelter, A. and Card, D. (eds.), Amsterdam: Elsevier Science

- [18] Imbens, G.W. and Rubin, D.B. (1997), *Estimating Outcome Distributions for Compliers in Instrumental Variables Models*, Review of Economic Studies, Vol. 64, No. , 555-574
- [19] LaLonde, R. (1986), *Evaluating the econometric evaluations of training programs with experimental data*, American Economic Review, Vol. 76, No. , 604-20
- [20] Lechner, M. (2001), *A note on the common support problem in applied evaluation studies*, Discussion Paper 2001-01, Department of Economics, University of St. Gallen
- [21] Little, R.J.A. and Yau, L. (1998), *Statistical Techniques for Analyzing Data from Prevention Trials: Treatment of No-Shows Using Rubin's Causal Model*, Psychological Methods, Vol. 3, No. 2, 147-159
- [22] Rosenbaum, P.R. (1984), *From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment*, Journal of the American Statistical Association, Vol. 79, No. 385, 41-48
- [23] Rosenbaum, P.R. (1987), *The Role of a Second Control Group in an Observational Study*, Statistical Science, Vol. 2, No. 3, 292-306
- [24] Rosenbaum, P.R. and Rubin, D.B. (1983), *The central role of the propensity score in observational studies for causal effects*, Biometrika, Vol. 70, No. , 41-55
- [25] Rubin, D.B. (1974), *Estimating causal effects of treatments in randomized and nonrandomized studies*, Journal of Educational Psychology, Vol. 66, No. , 688-701
- [26] Rubin, D.B. (1977), *Assignment to Treatment Group on the Basis of a Covariate*, Journal of Educational Statistics, Vol. 2, 4-58
- [27] Trochim, W. (1984), *Research Design for Program Evaluation: the Regression-Discontinuity Approach*, Beverly Hills: Sage Publications