Selecting Subpopulations for Causal Inference in Regression Discontinuity Designs

Alessandra Mattei

mattei@disia.unifi.it

University of Florence

Joint work with Laura Forastiere (University of Florence) and Fabrizia Mealli (University of Florence)

Workshop on
The Regression Discontinuity Design:
Methodological Issues and Applications in Economics, Statistics and Epidemiology
Institute for Fiscal Studies, London, UK

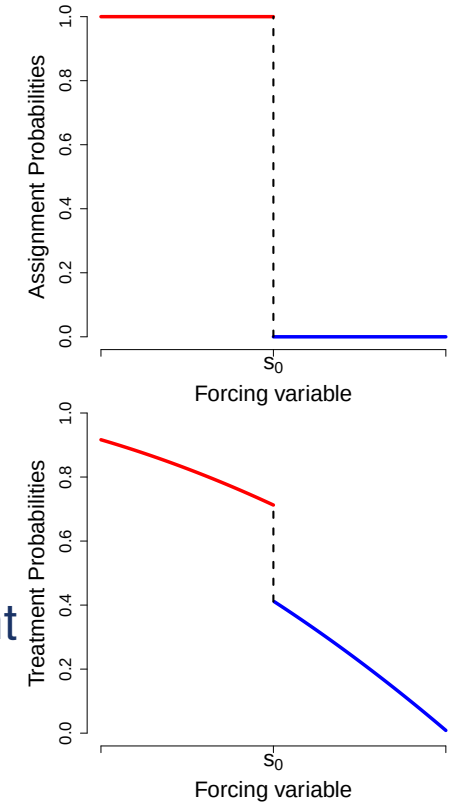
June 27th, 2017

Causal inference in regression-discontinuity (RD) designs

- The RD design is a quasi-experimental design where the treatment status changes discontinuously according to some underlying variable (*forcing variable*)
- Extracting causal information from RD designs is challenging
- Several methods for selecting suitable subpopulations for causal inference have been developed in the literature depending on the underlying assumptions *(e.g., Imbens and Kalyanaraman, 2012; Calonico et al., 2014, 2016; Cattaneo et al. 2015; Keele et al. 2015, Li et al., 2015)*
- Purpose: Drawing causal inference from RD designs for subpopulations of units for which the RD design defines a local randomized experiment

Regression discontinuity designs

- Two general setups: Sharp and Fuzzy RD designs
- In sharp RD designs:
  - ✓ Treatment assignment (eligibility) and treatment received are completely confounded
  - ✓ Treatment assignment and treatment status are deterministic step function of the forcing variable
- In fuzzy RD designs:
  - ✓ Treatment assignment (eligibility) still depends deterministically on the forcing variable, but
  - ✓ Treatment receipt does not coincide with treatment assignment



- RD designs are often exploited to identify causal effects of interventions
- Basic idea: Compare units with very similar values for the forcing variable, but different levels of treatment

Traditional approaches to the analysis of RD designs

Traditionally, RD designs are viewed as quasi-experimental designs with a non-probabilistic assignment mechanism

- The forcing variable is viewed as a pretreatment covariate
- Smoothness assumptions for the relationship between the potential outcomes and the forcing variable
- Focus on local causal effects at the threshold
  - ✓ A jump at the threshold, s_0, is interpreted as causal effect

(e.g., Thistlethwaite and Campbell, 1960; Imbens and Lemieux, 2008; Lee and Lemieux, 2010)

RD designs as local randomized experiments

- RD designs as local randomized experiments in a neighbourhood of the threshold *(e.g., Cattaneo et al., 2015; Li, Mattei and Mealli, 2015; Sales and Hansen, 2015; Mattei and Mealli, 2016)*

- Probabilistic formulation of the assignment mechanism underlying RD designs within the potential outcome approach *(Li, Mattei and Mealli, 2015; Mattei and Mealli, 2016)*

  - ✓ The forcing variable is viewed as a random variable

  - ✓ Local randomization: there exists at least a subpopulation, $\mathcal{U}_{s_0}$, around the threshold where the forcing variable, and therefore the treatment/eligibility status, can be seen as randomly assigned

  - ✓ Focus on local causal effects for units in $\mathcal{U}_{s_0}$

Our contribution

Following Li, Mattei and Mealli (2015), we use a probabilistic formulation of the assignment mechanism underlying RD designs within the potential outcome approach, proposing to

- Goal 1: Select suitable subpopulations around the cutoff point, $\mathcal{U}_{s_0}$, using a model-based finite mixture approach to clustering in a Bayesian framework

- Goal 2: Conduct exact Bayesian causal inference properly accounting for the uncertainty about the subpopulations, $\mathcal{U}_{s_0}$

Motivating Study: The Brazil's Bolsa Família (BF) Program

- Bolsa Família is a social welfare program of the Brazilian government, that started in 2003 and it is still ongoing
- Objective: Reducing short-term poverty by direct cash transfers and fighting long-term poverty by increasing human capital among poor Brazilian people through conditional cash transfers
  - ✓ Benefits are paid over time only to beneficiaries that comply with health and education conditionalities
- Causal question: Assessing causal effects of the Bolsa Família program on leprosy

The BF study: A (fuzzy) RD design

- BF benefit allocation rule: A family must (1) meet eligibility criteria; and (2) apply for the Bolsa Família benefits
  - ✓ Focus on families who applied for Bolsa Família benefits
- Eligibility: Per capita household income (forcing variable) falling below or above a pre-fixed threshold (120 Brazil Real $\simeq$ 36.5 USD per month)
- Eligible families who applied for BF benefits may not receive BF benefits due to budget constraints
- The Bolsa Família study defines a fuzzy RD design: Eligibility for BF benefits does not correspond with the receipt of BF benefits
  - ✓ Focus on intention-to-treat effects of eligibility statuses, not of the receipt of BF benefits
  - ✓ Intention-to-treat effects may be interesting for policy purposes

The potential outcome approach to causal inference

(Rubin, 1974, 1978)

- $i =$ Unit/Family ($i = 1, \ldots, N$)
- $\mathbf{X}_i =$ Vector of covariates
- $Z_i =$ BF benefit eligibility status:

$$Z_i = z \in \{0, 1\} = \{\text{Ineligible}, \text{Eligible}\}$$

- $S_i =$ Per capita household income: The forcing variable

$$Z_i = \mathbf{1}\{S_i \leq s_0\} \qquad s_0 = 120 \text{ Brazil Real (threshold)}$$

- $Y_i(\mathbf{s}) =$ Potential outcomes for the indicator of the presence of at least a leprosy case (after 2009) given the vector of values of the forcing variable, $\mathbf{s} \equiv (s_1, \ldots, s_N)'$

$$Y_i(\mathbf{s}) = \begin{cases} 1 & \text{If there is at least a leprosy case in family } i \text{ given } \mathbf{s} \\ 0 & \text{If there is no leprosy case in family } i \text{ given } \mathbf{s} \end{cases}$$

Local overlap, local RD-SUTVA and local estimands

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

- ✓ Under Local RD-SUTVA for each $i \in \mathcal{U}_{s_0}$, there are only two potential outcomes for the indicator of the presence of at least a leprosy case: $Y_i(0)$ and $Y_i(1)$

Causal Estimand. Local relative risk

$$RR_{\mathcal{U}_{s_0}} \equiv \frac{Pr\{Y_i(1) = 1; i \in \mathcal{U}_{s_0}\}}{Pr\{Y_i(0) = 1; i \in \mathcal{U}_{s_0}\}}$$

Probabilistic treatment assignment mechanism for RD designs

Assumption 3. Local Randomization (LR). For each $i \in \mathcal{U}_{s_0}$,

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

✓ Under local randomization, for each $i \in \mathcal{U}_{s_0}$,

$$\Pr(Z_i = 1) = \Pr(S_i \leq s_0)$$

Assumption 3'. Local Unconfoundedness (LU).
For each $i \in \mathcal{U}_{s_0}$,

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

✓ Under local unconfoundedness, for each $i \in \mathcal{U}_{s_0}$,

$$\Pr(Z_i = 1 \mid \mathbf{X}_i) = \Pr(S_i \leq s_0 \mid \mathbf{X}_i)$$

Selection of subpopulations $\mathcal{U}_{s_0}$: State of the art

- Local randomization based methods *(Cattaneo et al., 2015; Li, Mattei, Mealli, 2015; Licari, 2016)*
  - ✓ Assume LR and select subpopulations where pre-treatment variables are well balanced in the two subsamples defined by the assignment
  - ✓ Randomization or model-based Bayesian tests, possibly with adjustment for multiplicities
  - × These methods usually rely on assumptions on the shape of the subpopulations and are not immediately applicable when LU rather than LR is assumed
- Local unconfoundedness based methods
  - ✓ Assume LU and construct a subpopulation conditioning on observables and the discontinuity using penalized matching methods *(Keele et al., 2015)*
  - × The selected subpopulation depends on the penalty on the forcing variable distance between treated and control units
- LR and LU based methods do not directly account for the uncertainty about a selected subpopulation

Selection of subpopulations $\mathcal{U}_{s_0}$: Our proposal

- The problem of selecting suitable subpopulations, $\mathcal{U}_{s_0}$, as a clustering problem
- Sample units in a RD study come from (at least) three subpopulations:

$$\mathcal{U}_{s_0}^{-} = \{i \notin \mathcal{U}_{s_0} : S_i < s_0\} \qquad \mathcal{U}_{s_0} = \{i : S_i \in \mathcal{I}_{s_0}\} \qquad \mathcal{U}_{s_0}^{+} = \{i \notin \mathcal{U}_{s_0} : S_i > s_0\}$$

  where $\mathcal{I}_{s_0}$ is a neighborhood around s_0
- Crucial issue: We have some information on each subpopulation but we do not know which subpopulation each unit belongs to
- What do we know about the three subpopulations?
  - ✓ Each unit belongs to only one of the 3 subpopulations
  - ✓ For units who belong to $\mathcal{U}_{s_0}$ the RD assumptions hold
  - ✓ For units who belong to either $\mathcal{U}_{s_0}^{-}$ or $\mathcal{U}_{s_0}^{+}$ some RD assumptions may fail to hold
- Idea: Use clustering methods to ascertain, on the basis of the information we have, which subpopulation each unit belongs to
  - ✓ How can we include this information in the clustering algorithm?

Selection of subpopulations $\mathcal{U}_{s_0}$: A finite mixture model approach

- A finite mixture model-based approach *(e.g., McLachlan and Basford, 1988; Titterington, Smith, and Markov, 1985)*

$$
\begin{aligned}
p\left(S_i, Y_i(\mathbf{s}) \mid \mathbf{X}_i; \boldsymbol{\theta}\right) = & \\
& \pi_i(\mathcal{U}_{s_0}^-)\, p(S_i \mid \mathbf{X}_i; i \in \mathcal{U}_{s_0}^-; \boldsymbol{\eta}^-)\, p(Y_i(\mathbf{s}) \mid S_i, \mathbf{X}_i; i \in \mathcal{U}_{s_0}^-; \boldsymbol{\gamma}^-) + \\
& \pi_i(\mathcal{U}_{s_0})\, p(S_i \mid \mathbf{X}_i, i \in \mathcal{U}_{s_0}; \boldsymbol{\eta})\, p(Y_i(0), Y_i(1) \mid \mathbf{X}_i, i \in \mathcal{U}_{s_0}; \boldsymbol{\gamma}) + \\
& \pi_i(\mathcal{U}_{s_0}^+)\, p(S_i \mid \mathbf{X}_i, i \in \mathcal{U}_{s_0}^+; \boldsymbol{\eta}^+)\, p(Y_i(\mathbf{s}) \mid S_i, \mathbf{X}_i, i \in \mathcal{U}_{s_0}^+; \boldsymbol{\gamma}^+)
\end{aligned}
$$

where $\pi_i(\mathcal{U}_{s_0}^-) = Pr(i \in \mathcal{U}_{s_0}^- \mid \mathbf{X}_i; \boldsymbol{\alpha}) \geq 0$, $\pi_i(\mathcal{U}_{s_0}) = Pr(i \in \mathcal{U}_{s_0} \mid \mathbf{X}_i; \boldsymbol{\alpha}) \geq 0$, and $\pi_i(\mathcal{U}_{s_0}^+) = Pr(i \in \mathcal{U}_{s_0}^+ \mid \mathbf{X}_i; \boldsymbol{\alpha}) \geq 0$ are the mixing probabilities, with

$$\pi_i(\mathcal{U}_{s_0}^-) + \pi_i(\mathcal{U}_{s_0}) + \pi_i(\mathcal{U}_{s_0}^+) = 1,$$

$(\boldsymbol{\eta}^-, \boldsymbol{\gamma}^-)$, $(\boldsymbol{\eta}, \boldsymbol{\gamma})$ and $(\boldsymbol{\eta}^+, \boldsymbol{\gamma}^+)$ are parameter vectors defining each mixture component, and $\boldsymbol{\theta} = \left(\boldsymbol{\alpha}, \boldsymbol{\eta}^-, \boldsymbol{\gamma}^-, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\eta}^+, \boldsymbol{\gamma}^+\right)$ is the complete set of parameters specifying the mixture

- Bayesian approach to inference: Posterior computation via a Gibbs sampler with data augmentation (to impute missing subpopulation membership for each unit) *(e.g., Diebolt and Robert, 1994; Green and Richardson, 1997)*

BF study: Mixture-model specification

- Model for the mixing probabilities: conditional probit

$$\pi_i(\mathcal{U}_{s_0}^-) = \Pr(G_i^*(-) \leq 0) \qquad \pi_i(\mathcal{U}_{s_0}^+) = \Pr(G_i^*(-) > 0 \text{ and } G_i^*(+) \leq 0)$$

$$\pi_i(\mathcal{U}_{s_0}) = 1 - \pi_i(\mathcal{U}_{s_0}^-) - \pi_i(\mathcal{U}_{s_0}^+)$$

  where $G_i^*(-) = \alpha_0^- + \mathbf{X}_i'\boldsymbol{\alpha}^- + \epsilon_i^-$ and $G_i^*(+) = \alpha_0^+ + \mathbf{X}_i'\boldsymbol{\alpha}^+ + \epsilon_i^+$, with $\epsilon_i^- \sim N(0,1)$ and $\epsilon_i^+ \sim N(0,1)$, independently

- Models for the forcing variable (per capita household income): Log-normal models

$$\begin{aligned}
\log(S_i) \mid \mathbf{X}_i, i \in \mathcal{U}_{s_0}^- &\sim N\left(\beta_0^- + \mathbf{X}_i'\boldsymbol{\beta}^-; \sigma_-^2\right) \\
\log(S_i) \mid \mathbf{X}_i, i \in \mathcal{U}_{s_0}^+ &\sim N\left(\beta_0^+ + \mathbf{X}_i'\boldsymbol{\beta}^+; \sigma_+^2\right) \\
\log(S_i) \mid \mathbf{X}_i, i \in \mathcal{U}_{s_0} &\sim N\left(\beta_0 + \mathbf{X}_i'\boldsymbol{\beta}; \sigma^2\right)
\end{aligned}$$

- Models for the outcome (probit link):

$$\begin{aligned}
\Pr(Y_i(\mathbf{s}) = 1 | S_i = s, \mathbf{X}_i, i \in \mathcal{U}_{s_0}^-) &= \Phi\left(\gamma_0^- + \log(s)\gamma_1^- + \mathbf{X}_i'\boldsymbol{\gamma}^-\right) \\
\Pr(Y_i(\mathbf{s}) = 1 | S_i = s, \mathbf{X}_i, i \in \mathcal{U}_{s_0}^+) &= \Phi\left(\gamma_0^+ + \log(s)\gamma_1^+ + \mathbf{X}_i'\boldsymbol{\gamma}^+\right) \\
\Pr(Y_i(z) = 1 | \mathbf{X}_i, i \in \mathcal{U}_{s_0}) &= \Phi\left(\gamma_{0,z} + \mathbf{X}_i'\boldsymbol{\gamma}\right) \qquad z = 0, 1
\end{aligned}$$

BF study: Bayesian inference

- We assume that parameters are a priori independent
- We use weakly informative priors
  - ✓ Multivariate normal priors for the coefficients
  - ✓ Scaled inverse-χ^2 priors for the variances
- Finite sample estimands
- MCMC algorithm: For $\ell = 1 \ldots, L$
  - ✓ Impute missing subpopulation membership for each unit using a data augmentation step
  - ✓ Update the model parameters using Gibbs sampling
  - ✓ For each unit i in $\mathcal{U}_{s_0}$, draw the missing potential outcome, $Y_i^{mis} = Z_i Y_i(0) + (1 - Z_i) Y_i(1)$ from its posterior predictive distribution and calculate

$$RR^{\ell}_{\mathcal{U}_{s_0}} = \frac{\sum_{i:i\in\mathcal{U}_{s_0}} [Z_i Y_i^{obs} + (1 - Z_i) Y_i^{\ell}(1)] \Big/ N^{\ell}_{\mathcal{U}_{s_0}}}{\sum_{i:i\in\mathcal{U}_{s_0}} [(1 - Z_i) Y_i^{obs} + Z_i Y_i^{\ell}(0)] \Big/ N^{\ell}_{\mathcal{U}_{s_0}}}$$

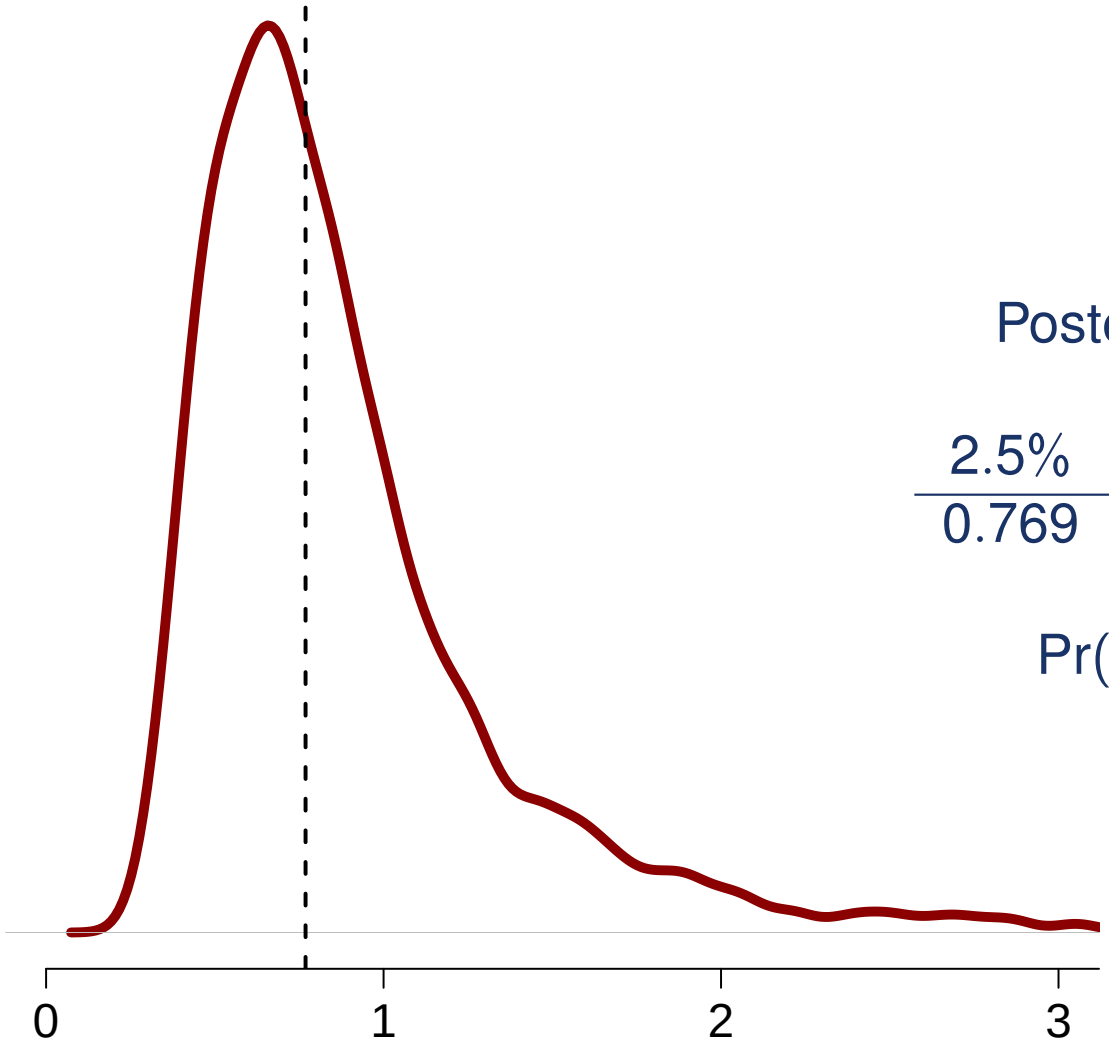
where $Y_i^{obs} = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ and $N^{\ell}_{\mathcal{U}_{s_0}}$ is the number of units in $\mathcal{U}_{s_0}$

Posterior distributions of the mixing probabilities

Estimand	Median	2.5%	97.5%
$\pi(\mathcal{U}_{s_0}^-)$	0.417	0.414	0.419
$\pi(\mathcal{U}_{s_0}^+)$	0.076	0.076	0.077
$\pi(\mathcal{U}_{s_0})$	0.507	0.504	0.510
$N_{\mathcal{U}_{s_0}}$	77 368	76 933	77 795
$\sum_{i \in \mathcal{U}_{s_0}} (1 - Z_i)$	2 724	2 647	2 804
$\sum_{i \in \mathcal{U}_{s_0}} Z_i$	74 644	74 231	75 041

- No assumption on the shape of the subpopulations
- Units with similar realized values of the forcing variable may belong to different subpopulations

Posterior distribution of $RR_{\mathcal{U}_{s_0}}$ (finite sample causal effect)



Posterior Median $= 0.769$

2.5%	5%	95%	97.5%
0.769	0.406	1.860	2.304

$\Pr(RR_{\mathcal{U}_{s_0}} < 1) = 0.723$

Concluding Remarks

- Crucial features of the model-based Bayesian mixture approach to the selection of subpopulations, $\mathcal{U}_{s_0}$, in RD designs
  - ✓ It explicitly accounts for the uncertainty about $\mathcal{U}_{s_0}$ membership
  - ✓ It imposes no constraint on the shape of $\mathcal{U}_{s_0}$
- We propose a model-based approach to causal inference, combining the selection of $\mathcal{U}_{s_0}$ and the inference on the local causal effects of interest for units belonging to $\mathcal{U}_{s_0}$ in a unique Bayesian framework
- Alternative approaches to causal inference, using the model-based mixture approach just as a tool to select suitable subpopulations
  - ✓ Multiple impute sub-population membership creating a set of complete membership datasets
  - ✓ For each complete membership dataset, use units belonging to $\mathcal{U}_{s_0}$ to draw inference on the causal effects of interest using a proper mode of causal inference
  - ✓ Combine the complete-data inferences on the local causal effects to form one inference that properly reflects missing $\mathcal{U}_{s_0}$-membership uncertainty (and possibly sampling variability)
- Extension to fuzzy RD designs

References

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

Calonico S., Cattaneo M.D., Farrell M.H., Titiunik R. (2016). Regression discontinuity designs using covariates. Working paper, University of Michigan.

Cattaneo M.D., Frandsen B.R., 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, 1-24.

Diebolt J., Christian P.R. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. JRSS-B 56, 363-375.

Imbens G.W., Lemieux T. (2008). Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics 142, 615-635.

Lee D.S., Lemieux T. (2010). Regression Discontinuity Designs in Economics. Journal of Economic Literature 485, 281-355.

Li F., Mattei A., Mealli F. (2015). Bayesian Inference for Regression Discontinuity Designs with Application to the Evaluation of Italian University Grants. AOAS 9, 1906-1931.
Imbens G.W., Kalyanaraman K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. Review of Economic Studies 79, 933-959.

Keele, L.J., 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. JRSS-A 178, 223-239.

Licari F. (2016). Randomization Based Inference in RD Designs with Multiple Outcomes: An Application to Italian University Grants. MA Thesis.

Mattei A., Mealli F. (2016). Regression Discontinuity Designs as Local Randomized. Observational Studies 2, 156-172.

McLachlan G., Basford K. (1988). Mixture Models: Inference and Application to Clustering. New York: Marcel Dekker.

Sales A., Hansen B.B. (2016). Limitless regression discontinuity: Causal inference for a population surrounding a threshold. ArXiv WP 1403.5478v3.

Richardson S., Green P.J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components. JRSS-B 59, 731-792.

Rubin D.B. (1974). Estimating Causal Effects of Treatments in Randomized and nonrandomized studies. Journal of Educational Psychology, 66: 688-701.

Rubin D.B. (1978). Bayesian Inference for Causal Effects. The Annals of Statistics 6, 34-58.

Thistlethwaite D., Campbell D. (1960). Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment. Journal of Educational Psychology 51, 309-317.

Titterington D., Smith A., Markov U. (1985). Statistical Analysis of Finite Mixture Distributions. Chichester, U.K.: John Wiley & Sons.