Statistical issues in the application of the regression discontinuity design for causal inference from clinical administrative databases

Gianluca Baio
University College London
Department of Statistical Science
g.baio@ucl.ac.uk

Joint work with Sara Geneletti (LSE), Aidan O’Keeffe & Federico Ricciardi (UCL), Sylvia Richardson (MRC-BSU), Linda Sharples (Leeds) & other collaborators — funded by UK MRC-MRP grant MR/K014838/1

ISBA World Meeting 2016
Forte Village Sardinia, Friday 17 June 2016
1. **Regression discontinuity design**
   - **Basic structure & assumptions**
   - **Sharp vs fuzzy**

2. It's good to be Bayesians (1)
   - Link to evidence synthesis
   - Stabilise estimates via prior information

3. It’s good to be Bayesians (2)
   - Binary outcomes
   - Wacky frequentist- vs regolarised Bayesian-estimates

4. How close is close enough?
   - Optimal bandwidths
   - No bandwiths?

5. Conclusions
1. Regression discontinuity design
   - Basic structure & assumptions
   - Sharp vs fuzzy

2. It’s good to be Bayesians (1)
   - Link to evidence synthesis
   - Stabilise estimates via prior information

3. It’s good to be Bayesians (2)
   - Binary outcomes
   - Wacky frequentist- vs regularised Bayesian-estimates

4. How close is close enough?
   - Optimal bandwidths
   - No bandwidths?

5. Conclusions
1. Regression discontinuity design
   - Basic structure & assumptions
   - Sharp vs fuzzy

2. It’s good to be Bayesians (1)
   - Link to evidence synthesis
   - Stabilise estimates via prior information

3. It’s good to be Bayesians (2)
   - Binary outcomes
   - Wacky frequentist- vs regularised Bayesian-estimates

4. How close is close enough?
   - Optimal bandwidths
   - No bandwidths?

5. Conclusions
1. Regression discontinuity design
   - Basic structure & assumptions
   - Sharp vs fuzzy

2. It’s good to be Bayesians (1)
   - Link to evidence synthesis
   - Stabilise estimates via prior information

3. It’s good to be Bayesians (2)
   - Binary outcomes
   - Wacky frequentist- vs regolarised Bayesian-estimates

4. How close is close enough?
   - Optimal bandwidths
   - No bandwidths?

5. Conclusions
Outline

1. Regression discontinuity design
   – Basic structure & assumptions
   – Sharp vs fuzzy

2. It’s good to be Bayesians (1)
   – Link to evidence synthesis
   – Stabilise estimates via prior information

3. It’s good to be Bayesians (2)
   – Binary outcomes
   – Wacky frequentist- vs regularised Bayesian-estimates

4. How close is close enough?
   – Optimal bandwidths
   – No bandwidths?

5. Conclusions
The Regression Discontinuity Design (RDD) was first introduced in the econometrics literature during the 1960s. The original idea was to exploit policy thresholds to estimate the causal effect of an educational intervention.

- Antiretroviral HIV drugs might be prescribed when a patient’s CD4 count is less than 200 cells/mm$^3$;
- Statins might be prescribed when a patient's 10-year risk of a cardiovascular event (10-year CVD risk score) exceeds a certain threshold (e.g. in the UK previously 20% and now 10%)

The key idea is that the threshold acts like a randomizing device. This is possible if we consider the units close to the threshold as they come from the same population in which the assignment variable has its own natural variability ⇒ (conditional) exchangeability.

Thistle & Campbell (1960) [9]
The Regression Discontinuity Design (RDD) was first introduced in the econometrics literature during the 1960s. The original idea was to exploit policy thresholds to estimate the causal effect of an educational intervention.

The RDD has proven to be very useful when treatment is assigned based on a pre-specified rule linked to a continuous variable. For example:
- Antiretroviral HIV drugs might be prescribed when a patient's CD4 count is less than 200 cells/mm$^3$;
- Statins might be prescribed when a patient's 10-year risk of a cardiovascular event (10-year CVD risk score) exceeds a certain threshold (e.g. in the UK previously 20% and now 10%)

The key idea is that the threshold acts like a randomizing device. This is possible if we consider the units close to the threshold as they come from the same population in which the assignment variable has its own natural variability ⇒ (conditional) exchangeability

Thistle & Campbell (1960) [9]; NICE (2008) [6]; NICE (2014) [7]
Sharp vs Fuzzy RDDs

- **Sharp design: Risk Score vs. LDL**
  - 10-year CVD Risk Score vs. LDL cholesterol (mmol/l)
  - Treated vs. Unreated

- **Sharp design: Risk Score vs. p(T=1)**
  - 10-year CVD Risk Score vs. p(T=1)
  - Treated vs. Unreated

Real data: Risk Score vs. p(T=1)

Gianluca Baio (UCL)
Sharp vs Fuzzy RDDs

Sharp design: Risk Score vs. LDL

Risk Score vs. LDL Chol Level (mmol/l)

Real data: Risk Score vs. p(T=1)

Gianluca Baio (UCL)
Notation & assumptions

- $X =$ continuous forcing/assignment variable;
- $Z =$ threshold indicator;
- $T =$ treatment administered;
- $C \equiv (O, U) =$ observed and unobserved covariates;
- $Y =$ outcome.

(Main) Assumptions

1. **Unconfoundedness:** $Y \perp \perp Z \mid (T, C, X)$ guarantees that the units just above and below the threshold are “similar”.

2. **Independence of Guidelines:** $Z \perp \perp C \mid X$
   the threshold is set by an external body, e.g. a governmental agency.

3. **Monotonicity:**
   No decision-maker systematically defies the guidelines.

---

- Denote \( x^c = x - x_0 \) to be the centered forcing variable
- Consider the linear model

\[
E(Y) = \mu_{il} = \beta_{0l} + \beta_{1l}x^c_{il} \quad l = \text{above, below}
\]

- **NB:** “close” to the threshold, the covariates \( C \) are balanced, so no need to control for them (kind of...) — **but:** how close is close? (more on this later)
The Causal Effect: Fuzzy RDD

- Denote \( x^c = x - x_0 \) to be the centered forcing variable
- Consider the linear model

\[
E(Y) = \mu_{il} = \beta_0 l + \beta_1 l x^c_{il} \quad l = \text{above, below}
\]

- **NB**: “close” to the threshold, the covariates \( C \) are balanced, so no need to control for them (kind of...) — **but**: how close is close? (more on this later)

- The formula for the fuzzy causal effect estimator is

\[
\text{LATE} = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(T|Z = 1) - E(T|Z = 0)} = \frac{\Delta_\beta}{\Delta_\pi} = \frac{\beta_{0a} - \beta_{0b}}{\pi_a - \pi_b}
\]

where \( \pi_l \) is an estimate of \( \Pr(T = 1|Z = z) \), e.g. the chance of being treated when above or below the threshold.

- **NB**: The RDD can be linked to instrumental variables (IVs)
Case study: prescription of statins

- **Statins** are a class of drugs used to lower cholesterol and prescribed to prevent heart disease
  - Trials show an average reduction of LDL cholesterol of \( \approx 2 \text{ mmol/l} \)
  - UK NHS guidelines are to prescribe statins to individuals without previous CVD if their 10 year CVD score exceeds 20%(10%)
Case study: prescription of statins

- **Statins** are a class of drugs used to lower cholesterol and prescribed to prevent heart disease
  - Trials show an average reduction of LDL cholesterol of $\approx 2$ mmol/l
  - UK NHS guidelines are to prescribe statins to individuals without previous CVD if their 10 year CVD score exceeds 20%(10%)

- **Data**: Simulation + real clinical practice database containing routine GP prescriptions as well as information on the variables that determine them (THIN: [www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database](http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database))
  - 587 general practices in the UK, covering 5.2% of the (2013) UK population — over 10 million individuals living in the UK and fairly representative of the general population
  - Individual characteristics (sex, date of birth, date of registration, proxies of socioeconomic status)
  - Medical history (GP visits, prescriptions, exams)
  - Relevant clinical outcomes (LDL level, CHD events, deaths)
• Stabilising the estimators
  – The denominator of LATE can be very small (i.e. $\pi_a \approx \pi_b$)
  – Informative priors on the relevant parameters can encode knowledge and assumptions about these two probabilities so that the resulting estimator does not explode to $\infty$

The importance of being a Bayesian (1)

- **Stabilising the estimators**
  - The denominator of LATE can be very small (i.e. $\pi_a \approx \pi_b$)
  - Informative priors on the relevant parameters can encode knowledge and assumptions about these two probabilities so that the resulting estimator does not explode to $\infty$

- **Computational advantages**
  - Estimation of variances and intervals does not rely on asymptotics — just a byproduct of MCMC procedures + can naturally include more appropriate models (vs 2SLS)

---

The importance of being a Bayesian (1)

- **Stabilising the estimators**
  - The denominator of LATE can be very small (i.e. $\pi_a \approx \pi_b$)
  - Informative priors on the relevant parameters can encode knowledge and assumptions about these two probabilities so that the resulting estimator does not explode to $\infty$

- **Computational advantages**
  - Estimation of variances and intervals does not rely on asymptotics — just a byproduct of MCMC procedures + can naturally include more appropriate models (vs 2SLS)

- **Expand the model to include extra information & deal with the two levels of compliance (GP vs patients)**
  - For example, logistic regression models to explain the treatment assignment in terms of practice-level covariates
  - Mixture models to include individual level covariates to account for proxies of compliance with treatment

---

The importance of being a Bayesian (1)

- **Stabilising the estimators**
  - The denominator of LATE can be very small (i.e. $\pi_a \approx \pi_b$)
  - Informative priors on the relevant parameters can encode knowledge and assumptions about these two probabilities so that the resulting estimator does not explode to $\infty$

- **Computational advantages**
  - Estimation of variances and intervals does not rely on asymptotics — just a byproduct of MCMC procedures + can naturally include more appropriate models (vs 2SLS)

- **Expand the model to include extra information & deal with the two levels of compliance (GP vs patients)**
  - For example, logistic regression models to explain the treatment assignment in terms of practice-level covariates
  - Mixture models to include individual level covariates to account for proxies of compliance with treatment

- **Cooler! 😏

Informative prior on the slopes, based on clinical expert opinions

Estimated prior predictive distribution of LDL cholesterol for a patient whose risk score = 0

Estimated prior predictive distribution of LDL cholesterol for a patient whose risk score = 0.199
Bayesian modelling: \( \mu_{il} = \beta_{0l} + \beta_{1l}x_{il}^c \)

1. Informative priors on the intercepts:
   \( \beta_{1b} \sim \text{Normal}(m_1, s_1^2) \) and \( \beta_{1a} = \beta_{0b} + \phi \)

   - **Weakly informative prior**: \( \phi \sim \text{Normal}(0, 2) \)
     - "Skepical" prior on the effect of treatment, which is assumed to be null

   - **Strongly informative prior**: \( \phi \sim \text{Normal}(-2, 1) \)
     - "Enthusiastic" prior, strongly based on the available information coming from the RCTs (reduction of 2 mmol/l)
     - Relatively small variance to represent strong belief in the trials

---

Bayesian modelling: $\mu_{il} = \beta_0 l + \beta_1 l x_{il}^c$

2 Informative priors on the intercepts:
   $\beta_{1b} \sim \text{Normal}(m_1, s_1^2)$ and $\beta_{1a} = \beta_{0b} + \phi$

- **Weakly informative prior**: $\phi \sim \text{Normal}(0, 2)$
  - “Sketpical” prior on the effect of treatment, which is assumed to be null

- **Strongly informative prior**: $\phi \sim \text{Normal}(-2, 1)$
  - “Enthusiastic” prior, strongly based on the available information coming from the RCTs (reduction of 2 mmol/l)
  - Relatively small variance to represent strong belief in the trials

3 Informative prior on the probability of treatment:

   $\text{logit}(\pi_a) \sim \text{Normal}(2, 1), \quad \text{logit}(\pi_b) \sim \text{Normal}(-2, 1)$

- **NB**: implies that $\Delta_\pi = \pi_a - \pi_b$ is centered far from 0 but can vary
- Helps stabilise the denominator and thus the LATE

Bandwidth = 0.25 (fairly large!), Treatment Effect Size = 2

\[
\text{LATE estimates (Strong IV)}
\]

\[
\text{LATE estimates (Weak IV)}
\]

\[
\text{LATE}_{cnst} = \frac{\Delta_{sip}}{\beta \Delta_{fix} \pi}
\]

\[
\text{LATE}_{flex} = \frac{\Delta_{wip}}{\beta \Delta_{fdp} \pi}
\]

\[
\text{LATE}_{unct} = \frac{\Delta_{wip}}{\beta \Delta_{unc} \pi}
\]
• Most of the RDD literature focusses on continuous outcomes, but often in biostatistics, practitioners are interested in \textbf{binary} outcomes

• Can draw on the IV-based Multiplicative Structural Mean Models (MSMMs), which consider the causal \textbf{Risk Ratio for the Treated} (RRT)

\[
\text{RRT} = \frac{E[E_a(Y \mid Z) \mid T = 1]}{E[E_b(Y \mid Z) \mid T = 1]}
= 1 - \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(Y \bar{T} \mid Z = 1) - E(Y \bar{T} \mid Z = 0)}
\]

when a set of assumptions holds (log-linear in $t + \text{no } T-Z$ multiplicative interaction)

\[
\text{Geneletti et al (2016) [4]}
\]
The importance of being a Bayesian (2)

- Most of the RDD literature focusses on continuous outcomes, but often in biostatistics, practitioners are interested in **binary** outcomes.
- Can draw on the IV-based Multiplicative Structural Mean Models (MSMMs), which consider the causal **Risk Ratio for the Treated** (RRT)

\[
RRT = \frac{E[\mathbb{E}_a(Y | Z) | T = 1]}{E[\mathbb{E}_b(Y | Z) | T = 1]}
= 1 - \frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(Y \bar{T} | Z = 1) - E(Y \bar{T} | Z = 0)}
\]

when a set of assumptions holds (log-linear in $t +$ no $T$-$Z$ multiplicative interaction).

- Known issues of standard estimators (*e.g.* generalised method of moments):
  - May give absurd results (lower 95% interval estimate < 0)
  - The data for the product term $(Y \bar{T})$ are usually sparse ⇒ implausibly wide interval estimates
- Can “fix” it by using suitable constraints — fairly easy in a Bayesian setting

---

Constraining the models (1)

• The RRT is expressed as a function of a set of parameters (in the same spirit as the LATE)

\[ RRT = f(\exp(\alpha_a) - \exp(\alpha_b)) \]

where:
- \( \alpha_a \) and \( \alpha_b \) are the intercepts in the log-linear models for \( E(Y \mid Z = 1) \) and \( E(Y \mid Z = 0) \)
- For convenience, model \( y_{il} \sim \text{Poisson}(\mu_{il}) \) — consistent with MSMM assumptions

• Typically, we would put priors on \( \alpha_a \) and \( \alpha_b \), which would induce a prior on RRT
Constraining the models (1)

- The RRT is expressed as a function of a set of parameters (in the same spirit as the LATE)

\[ RRT = f(\exp(\alpha_a) - \exp(\alpha_b)) \]

where:

- \( \alpha_a \) and \( \alpha_b \) are the intercepts in the log-linear models for \( E(Y \mid Z = 1) \) and \( E(Y \mid Z = 0) \)
- For convenience, model \( y_{il} \sim \text{Poisson}(\mu_{il}) \) — consistent with MSMM assumptions

- Typically, we would put priors on \( \alpha_a \) and \( \alpha_b \), which would induce a prior on RRT

- **But**: can also put a prior on RRT to ensure that it is \( > 0 \) and, say, \( \alpha_a \) and then induce a prior on \( \alpha_b \), e.g.

\[ RRT \sim \text{Gamma}(3, 1) \quad \alpha_a \sim p(\alpha_a) \quad \text{and} \quad \alpha_b = g(RRT, \alpha_a) \]
NB: “Industry standard” methods (based on generalised method of moments) fail to give reasonable results in many scenarios.
Bandwidth selection for RDD is addressed in the literature in two main ways:
Bandwidth selection for RDD is addressed in the literature in two main ways:

1. Producing results using data within a number of different bandwidths, that may be selected also with the guidance of an expert in the field of study.

2. Selecting an “optimal” bandwidth aimed at minimizing an error term related to the estimation of the effect in a non-parametric fashion.
Bandwidth selection for RDD is addressed in the literature in two main ways:

1. Producing results using data within a number of different bandwidths, that may be selected also with the guidance of an expert in the field of study.

2. Selecting an “optimal” bandwidth aimed at minimizing an error term related to the estimation of the effect in a non-parametric fashion.
Bandwidth selection for RDD is addressed in the literature in two main ways:

1. Producing results using data within a number of different bandwidths, that may be selected also with the guidance of an expert in the field of study.

2. Selecting an “optimal” bandwidth aimed at minimizing an error term related to the estimation of the effect in a non-parametric fashion.

   CV A Cross Validation based approach

---

Bandwidth selection for RDD is addressed in the literature in two main ways:

1. Producing results using data within a number of different bandwidths, that may be selected also with the guidance of an expert in the field of study.

2. Selecting an “optimal” bandwidth aimed at minimizing an error term related to the estimation of the effect in a non-parametric fashion.

   CV  A Cross Validation based approach
   IK  A Mean Square Error minimization based method, designed to give unbiased point estimator for the effect;

---

Bandwidth selection for RDD is addressed in the literature in two main ways:

1. Producing results using data within a number of different bandwidths, that may be selected also with the guidance of an expert in the field of study.

2. Selecting an “optimal” bandwidth aimed at minimizing an error term related to the estimation of the effect in a non-parametric fashion.

   - **CV** A Cross Validation based approach
   - **IK** A Mean Square Error minimization based method, designed to give unbiased point estimator for the effect;
   - **CCT** A bias-correction and robust inference method recently, focusing on getting an unbiased interval estimator for the effect.

---

Fail to account properly for the **real** issue — **exchangeability**

```r
set.seed(90) # sets the random number generator seed
tr <- 0.5 # sets the value for the threshold
X <- runif(2000,0,1) # generates assignment variable
Z <- as.numeric(X>tr) # generates treatment indicator
X.c <- X-tr # centers assignment variable
e <- rnorm(2000, 0, .05) # generates random error (white noise)
Y <- Z*(3+0.5*X.c) + (1-Z)*(2.5+0.7*X.c) + e # generates the outcome variable
```

**NB:** In this case

- The design is sharp
- No unobserved confounders
- The outcome is fully determined by the forcing variable

⇒ observations are (conditionally) exchangeable below and above the threshold!
⇒ shouldn’t we be able to use **all** (most?) the data to estimate the causal effect (and gain precision)?
Use them all?
Issues with the RDD for causal inference

Default optimisation (min MSE)
Use them all? Alternative optimisation (sum for regression estimates)
Alternative methods?

- Possibly use flexible regressions (e.g. splines) — but standard setting may not be flexible enough...
  - We may know that individuals at either extreme really are too different and do not want them to basically matter at all...

- Other avenues?
  - Reverse random walk priors — anchor priors to one extreme and "filter" irrelevant data [work in progress]
  - Spatially structured models — formally account for spatial distance from the threshold [work in progress?]

- More importantly, selection should happen according to balancing in the confounders above & below the threshold!
• Possibly use flexible regressions (e.g. splines) — but standard setting may not be flexible enough...
  – We may know that individuals at either extreme really are too different and do not want them to basically matter at all...

• Other avenues?
  – **Reverse random walk priors** — anchor priors to one extreme and “filter” irrelevant data [work in progress]
  – Spatially structured models — formally account for spatial distance from the threshold [work in progress?!]

• More importantly, selection should happen according to balancing in the confounders above & below the threshold!
Conclusions

- "Real World Evidence" (i.e. Electronic Health Record data) is increasingly popular in research
  - Causal estimates are still tricky because of issues with self-selection, confounding, etc
- Useful to (critically!) explore specific designs to balance characteristics
  - RDD
  - Interrupted time series
  - ...

Gianluca Baio (UCL)
Conclusions

• “Real World Evidence” (i.e. Electronic Health Record data) is increasingly popular in research
  – Causal estimates are still tricky because of issues with self-selection, confounding, etc

• Useful to (critically!) explore specific designs to balance characteristics
  – RDD
  – Interrupted time series
  – ...

• Bayesian modelling particularly helpful
  – Because data are available in registries, administrative databases, there are likely to be RCTs (may be on small samples/time frames) to base priors on
  – Design alone may not be sufficient to obtain balance — may need to impose constraints ⇒ explicit and typically relatively easy in a full Bayesian framework


Thank you!