Declaring and Diagnosing Research Designs

Book Proposal Sample Sections

Graeme Blair Jasper Cooper Alexander Coppock and Macartan Humphreys

October 20, 2017

This document includes three sample sections of *Declaring and Diagnosing Reseach Designs*. The first two are entries in the Design Library (Part B) and the third is an entry in the Principles section (Part C).

Contents

Simple Random Sampling

Often we are interested in features of a population, but data on the entire population is prohibitively expensive to collect. Instead, researchers obtain data on a small fraction of the population and use measurements taken on that sample to draw inferences about the population.

Imagine we seek to estimate the average political ideology of residents of the small town of Portola, California, on a left-right scale that varies from 1 (most liberal) to 7 (most conservative). We draw a simple random sample in which residents have an equal chance of inclusion in the study. It's a straightforward design, but formally declaring it will make it easy to assess its properties.

Design Declaration

M Even for this most basic of designs, researchers bring to bear a background model of the world. As described in Chapter 1, the three elements of a model are the signature, probability distributions over variables, and functional equations among variables. The signature here is a specification of the variable of interest, Y, with a well defined domain (seven possible values between 1 and 7). In the code declaration below, we assume a uniform distribution over these 7 values. This choice is a speculation about the population distribution of Y; some features of the design diagnosis will depend on the choice of distribution. The functional equations seem absent here as there is only one variable in the model. We could consider an elaboration of the model that includes three variables: the true outcome, Y; the decision to measure the outcome, M; and the measured outcome, Y^M. We ignore this complication for now under the assumption that $Y = Y^M$, i.e., that Y is measured perfectly. Finally, the model also includes information about the size of the population. Portola, California, has a population of approximately 2100 people as of 2010, so $N = 2100$.

I Our inquiry is the population mean of Y: $\frac{1}{N}\sum_1^N Y_i = \bar{Y}$.

D In simple random sampling, we draw a random sample without replacement of size n, where every member of the population has an equal probability of inclusion in the sample, $\frac{n}{N}$. When N is very large relative to n, units are drawn approximately independently. In this design we measure Y for $n = 100$ units in the sample; the other $N - n$ units are not measured.

A We estimate the population mean with the sample mean estimator: $\widehat{\bar{Y}} = \frac{1}{n}\sum_1^n Y_i$. Even though our inquiry implies our answer should be a single number, an answer strategy typically also provides statistics that help us assess the uncertainty around that single number. To construct a 95% confidence interval around our estimate, we calculate the standard error of the sample mean, then approximate the sampling distribution of the sample mean estimator using a formula that includes a finite population correction. In particular, we approximate the estimated sampling distribution by a t distribution with $n - 1$ degrees of freedom. In the code for our answer strategy, we spell out each step in turn.

```
# Model ------------------------------------------------------------------------
N <- 2100
population <- declare_population(N = N, Y = sample(1:7, N, replace = TRUE))
fixed_population <- population()

# Inquiry ----------------------------------------------------------------------
estimand <- declare_estimand(Ybar = mean(Y))

# Data Strategy ----------------------------------------------------------------
n <- 100
sampling <- declare_sampling(n = n)

# Answer Strategy --------------------------------------------------------------
estimator <- declare_estimator(
  estimator_function = function(data) {
      est <- mean(data$Y)
      se  <- sd(data$Y) * sqrt((N / n - 1) / (N - 1))
      critical_value <- qt(0.975, df = n - 1)
      ci_lower <- est - critical_value * se
      ci_upper <- est + critical_value * se
      data.frame(est, ci_lower, ci_upper)},
  estimand = estimand,
  label = "Sample Mean Estimator")

# Design -----------------------------------------------------------------------
design <- declare_design(fixed_population, estimand, sampling, estimator)
diagnosands <- declare_diagnosands(
  bias = mean(est - estimand),
  mean_est = mean(est),
  coverage = mean(ci_lower <= estimand & estimand <= ci_upper)
  )
```

Takeaways

With the design declared we can run a diagnosis and plot results from Monte Carlo simulations of the design:

```
diagnosis <- diagnose_design(
  design, sims = 10000, bootstrap_sims = 1000, diagnosands = diagnosands)
```

Mean Estimate	Bias	SE(Bias)	Coverage	SE(Coverage)
3.92	-0.00	0.00	0.95	0.00

The diagnosis indicates that under simple random sampling, the sample mean estimator of the population mean is unbiased. The graph on the left shows the sampling distribution of the

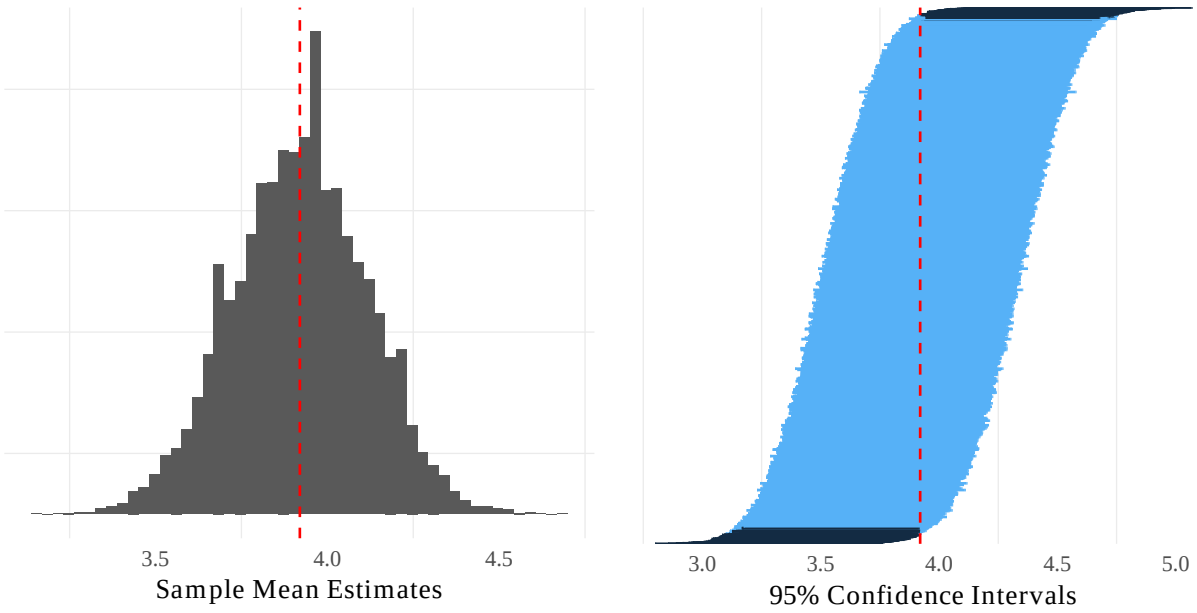


Figure 1: Sampling Distributions of the Sample Mean and 95% Confidence Interval

estimator: it's centered directly on the true value of the inquiry. Confidence intervals *also* have a sampling distribution – they change depending on the idiosyncrasies of each sample we happen to draw. The figure on the right shows that the 95% of the time the confidence intervals cover the true value of the estimand, as they should. As sample size grows, the sampling distribution of the estimator gets tighter, but the coverage of the confidence intervals stays at 95% – just the properties we would want out of our answer strategy.

Things work well here it seems. In the exercises we suggest some small modifications of the design that point to conditions under which things might break down.

Exercises

1. Modify the declaration to change the distribution of Y from being uniform to something else: perhaps imagine that more extreme ideologies are more prevalent than moderate ones. Is the sample mean estimator still unbiased? Interpret your answer.

2. Change the sampling procedure to favor units with higher values of ideology. Is the sample mean estimator still unbiased? Interpret your answer.

3. Modify the estimation function to use this formula for the standard error: $\widehat{se} \equiv \frac{\widehat{\sigma}}{\sqrt{n}}$. This equation differs from the one used in our declaration (it ignores the total population size N). Check that the coverage of this new design is incorrect when $N = n$. Assess how large N has to be for the difference between these procedures not to matter.

Regression Discontinuity

Regression discontinuity designs exploit substantive knowledge that treatment is assigned in a particular way: everyone above a threshold is assigned to treatment and everyone below it is not. Even though researchers do not control the assignment, substantive knowledge about the threshold serves as a basis for a strong identification claim.

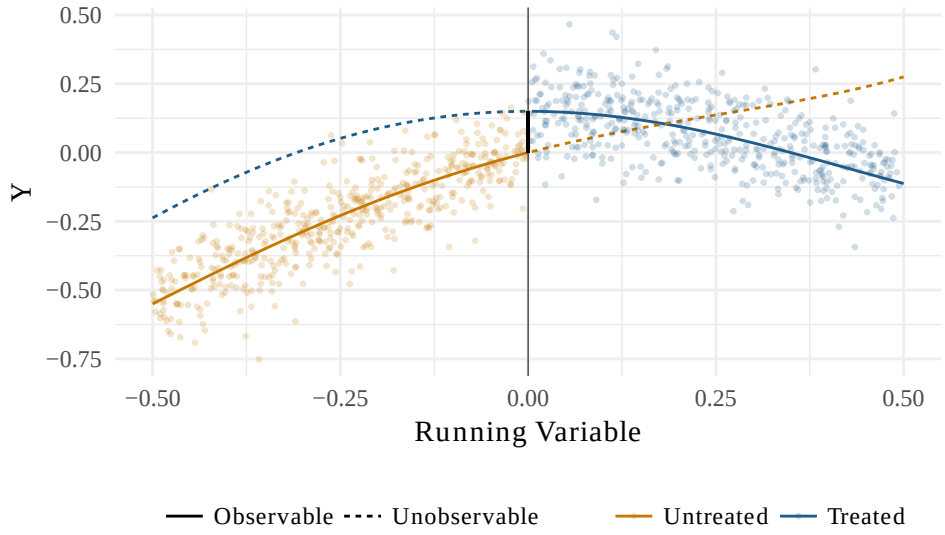
Thistlewhite and Campbell introduced the regression discontinuity design in the 1960s to study the impact of scholarships on academic success. They claim that students with a test score just above a scholarship cutoff were plausibly comparable to students with scores just below the cutoff, so differences in future academic success could be attributed to the scholarship alone.

Regression discontinuity designs identify a *local* average treatment effect: the average effect of treatment *exactly at the cutoff*. The main trouble with the design is that there is vanishingly little data exactly at the cutoff, so any answer strategy needs to use data that is some distance away from the cutoff. The further away from the cutoff we move, the larger the threat of bias.

We consider a regression discontinuity design application that examines party incumbency advantage: the effect of a party winning an election on its vote margin in the next election.

Design Declaration

M Regression discontinuity designs have four components: A running variable, a cutoff, a treatment variable, and an outcome. The cutoff determines which units are treated depending on the value of the running variable. In our example, the running variable X is the Democratic party's margin of victory at time $t-1$; and the treatment, Z, is whether the Democratic party won the election in time $t-1$. The outcome, Y, is the Democratic vote margin at time t. We'll consider a population of 1,000 of these pairs of elections.



A major assumption required for regression discontinuity is that the conditional expectation functions for both treatment and control potential outcomes are continuous at the cutoff.[1] To

[1] An alternative motivation for some designs that do not rely on continuity at the cutoff is "local randomization".

satisfy this assumption, we specify two smooth conditional expectation functions, one for each potential outcome. The figure plots Y (the Democratic vote margin at time t) against X (the margin at time $t-1$). We've also plotted the true conditional expectation functions for the treated and control potential outcomes. The solid lines correspond to the observed data and the dashed lines correspond to the unobserved data.

I Our estimand is the effect of a Democratic win in an election on the Democratic vote margin of the next election, when the Democratic vote margin of the first election is zero. Formally, it is the difference in the conditional expectation functions of control and treatment potential outcomes when the running variable is exactly zero. The plot's black vertical line shows this difference.

D We collect data on the Democratic vote share at time $t-1$ and time t for all 1,000 pairs of elections. There is no sampling or random assignment.

A We will approximate the treated and untreated conditional expectation functions to the left and right of the cutoff using a flexible regression specification estimated via OLS. In particular, we fit each regression using a fourth-order polynomial. Much of the literature on regression discontinuity designs focuses on the tradeoffs among answer strategies, with many analysts recommending against higher-order polynomial regression specifications. We use one here to highlight how well such an answer strategy does when it matches the functional form in the model. We discuss alternative estimators in the exercises.

```
# Model ------------------------------------------------------------------------
cutoff <- .5
control <- function(X) { as.vector(poly(X, 4, raw = T) %*% c(.7, -.8, .5, 1))}
treatment <- function(X) { as.vector(poly(X, 4, raw = T) %*% c(0, -1.5, .5, .8)) + .15}
population <- declare_population(N = 1000,
  X = runif(N,0,1) - cutoff,
  noise = rnorm(N,0,.1),
  Z = 1 * (X > 0))
potential_outcomes <- declare_potential_outcomes(
  Y_Z_0 = control(X) + noise,
  Y_Z_1 = treatment(X) + noise)

# Inquiry ----------------------------------------------------------------------
estimand <- declare_estimand(LATE = treatment(0) - control(0))

# Answer Strategy --------------------------------------------------------------
estimator <- declare_estimator(Y ~ poly(X, 4) * Z,
                               model = lm,
                               estimand = estimand)

# Design -----------------------------------------------------------------------
design <- declare_design(
    population, potential_outcomes, estimand, reveal_outcomes, estimator)
```

Takeaways

We now diagnose the design:

```
diagnosis <- diagnose_design(
  design, sims = 10000, bootstrap_sims = 1000, diagnosands = diagnosands)
```

Bias	SE(Bias)	Power	SE(Power)	Coverage	SE(Coverage)
-0.057	0.009	0.055	0.002	0.945	0.002

We highlight three takeaways. First, the power of this design is very low: with 1,000 units we do not achieve even 10% statistical power. However, our estimates of the uncertainty are not too wide: the coverage probability indicates that our confidence intervals indeed contain the estimand 95% of the time as they should. Our answer strategy is highly uncertain because the fourth-order polynomial specification in regression model gives weights to the data that greatly increase the variance of the estimator (Gelman and Imbens, 2017). In the exercises we explore alternative answer strategies that perform better.

Second, the design is biased because polynomial approximations of the average effect at exactly the point of the threshold will be inaccurate in small samples (Sekhon and Titiunik, 2017), especially as units farther away from the cutoff are incorporated into the answer strategy. We know that the estimated bias is not due to simulation error by examining the bootstrapped standard error of the bias estimates.

Finally, from the figure, we can see how poorly the average effect at the threshold approximates the average effect for all units. The average treatment effect among the treated (to the right of the threshold in the figure) is negative, whereas at the threshold it is positive. This clarifies that the estimand of the regression discontinuity design, the difference at the cutoff, is only relevant for a small – and possibly empty – set of units very close to the cutoff.

Further Reading

1. Since its rediscovery by social scientists in the late 1990s, the regression discontinuity design has been widely used to study diverse causal effects such as: prison on recidivism (Mitchell et al., 2017); China's one child policy on human capital (Qin, Zhuang and Yang, 2017); eligibility for World Bank loans on political liberalization (Carnegie and Samii, 2017); and anti-discrimination laws on minority employment (Hahn, Todd and Van der Klaauw, 1999).

2. We've discussed a "sharp" regression discontinuity design in which all units above the threshold were treated and all units below were untreated. In fuzzy regression discontinuity designs, some units above the cutoff remain untreated or some units below take treatment. This setting is analogous to experiments that experience noncompliance and may require instrumental variables approaches to the answer strategy (see **Compliance is a Potential Outcome**).

3. Geographic regression discontinuity designs use distance to a border as the running variable: units on one side of the border are treated and units on the other are untreated. Keele

and Titiunik (2016) use such a design to study whether voters are more likely to turn out when they have the opportunity to vote directly on legislation on so-called ballot initiatives. A complication of this design is how to measure distance to the border in two dimensions.

Exercises

1. Gelman and Imbens (2017) point out that higher order polynomial regression specifications lead to extreme regression weights. One approach to obtaining better estimates is to select a bandwidth, h, around the cutoff, and run a linear regression. Declare a sampling procedure that subsets the data to a bandwidth around the threshold, as well as a first order linear regression specification, and analyze how the power, bias, RMSE, and coverage of the design vary as a function of the bandwidth.

2. The 'rdrobust' estimator in the 'rdrobust' package implements a local polynomial estimator that automatically selects a bandwidth for the RD analysis and bias-corrected confidence intervals. Declare another estimator using the 'rdrobust' function and add it to the design. How does the coverage and bias of this estimator compare to the regression approaches declared above?

3. Reduce the number of polynomial terms of the the 'treatment()' and 'control()' functions and assess how the bias of the design changes as the potential outcomes become increasingly linear as a function of the running variable.

4. Redefine the population function so that units with higher potential outcome are more likely to locate just above the cutoff than below it. Assess whether and how this affects the bias of the design.

Questions Should Have Answers

A basic requirement of a good research design is that the question it seeks to answer does in fact *have* an answer, at least under plausible models of the world. In our framework, this means that an inquiry I must have an associated answer a^M, which refers to the answer under the model. Interestingly, we sometimes might not be conscious that the questions we ask do not have answers. Fortunately, when we ask a computer to answer such a question, it complains.

How could a question not have an answer? Answerless questions can arise when inquiries depend on variables that do not exist or are undefined for some units. In other words, when there is a mismatch between the model and the inquiry, we're asking a question about something that doesn't exist.

Consider an audit experiment (see **Audit Experiment Design**) that seeks to assess the effects of an email from a Latino name (versus a White name) on *whether* and *how well* election officials respond to requests for information. For example, do they use a positive or negative tone. These questions seem reasonable enough. The problem, however, is that if there are officials who don't send responses, tone is undefined. More subtly, if there is an official that does send an email but would not have sent it in a different treatment condition, then tone is undefined for one of their potential outcomes.

Design Declaration

M The model has two outcome variables, R_i and Y_i. R_i stands for "response" and is equal to 1 if a response is sent, and 0 otherwise. Y_i is the tone of the response and is normally distributed when it is defined. Z_i is the treatment and equals 1 if the email is sent using a Latino name and 0 otherwise. The table below shows the potential outcomes for four possible types of subjects, depending on the potential outcomes of R_i. A types always respond regardless of treatment and D types never respond, regardless of treatment. B types respond if and only if they are treated, whereas C types respond if and only if they are *not* treated. The table also includes columns for the potential outcomes of Y_i, showing which potential outcome subjects would express depending on their type. The key thing to note is that for the B, C, and D types, the effect of treatment on Y_i is *undefined* because messages never sent have no tone. The last (and very important) feature of our model is that the outcomes Y_i are possibly correlated with subject type. Even though both $E[Y_i(1)|\text{Type} = A]$ and $E[Y_i(1)|\text{Type} = B]$ exist, there's no reason to expect that they are the same. In the design we assume a distribution of types with 40% A, 5% B, 10% C, and 45% D.

Type	$R_i(0)$	$R_i(1)$	$Y_i(0)$	$Y_i(1)$
A	1	1	$Y_i(0)$	$Y_i(1)$
B	0	1	NA	$Y_i(1)$
C	1	0	$Y_i(0)$	NA
D	0	0	NA	NA

Table 1: Causal Types

I We have two inquiries. The first is straightforward: $E[R_i(1) - R_i(0)]$ is the Average Treatment Effect on response. The second inquiry is the undefined inquiry that does not have an answer:

$E[Y_i(1) - Y_i(0)]$. We will also consider a third inquiry, which *is* defined: $E[Y_i(1) - Y_i(0)|\text{Type} = A]$, which is the average effect of treatment on tone among A types.

D The data strategy will be to use complete random assignment to assign 250 of 500 units to treatment.

A We'll try to answer all three inquiries with the difference-in-means estimator, but as the diagnosis will reveal, this strategy works well for some inquiries but not others.

```
# Model ------------------------------------------------------------------------
population <- declare_population(
  N = 500,
  type = sample(c("A", "B", "C", "D"), size = N,
                replace = TRUE, prob = c(.40, .05, .10, .45)))

potential_outcomes <- declare_potential_outcomes(
  R_Z_0 = type %in% c("A", "C"),
  R_Z_1 = type %in% c("A", "B"),
  Y_Z_0 = ifelse(
    R_Z_0, rnorm(n = sum(R_Z_0), mean = .1*(type == "A") - 2*(type == "C")), NA),
  Y_Z_1 = ifelse(
    R_Z_1, rnorm(n = sum(R_Z_1), mean = .2*(type == "A") + 2*(type == "B")), NA)
)

# Inquiry ----------------------------------------------------------------------
estimand_1 <- declare_estimand(ATE_R = mean(R_Z_1 - R_Z_0))
estimand_2 <- declare_estimand(ATE_Y = mean(Y_Z_1 - Y_Z_0))
estimand_3 <- declare_estimand(
  ATE_Y_for_As = mean(Y_Z_1[type == "A"] - Y_Z_0[type == "A"]))

# Data Strategy ----------------------------------------------------------------
assignment <- declare_assignment(m = 250)

# Answer Strategy --------------------------------------------------------------
estimator_1 <- declare_estimator(R ~ Z, estimand = estimand_1, label = ATE_R)
estimator_2 <- declare_estimator(Y ~ Z, estimand = estimand_2, label = ATE_Y)
estimator_3 <- declare_estimator(Y ~ Z, estimand = estimand_3, label = ATE_YA)

# Design -----------------------------------------------------------------------
design <- declare_design(
  population,
  potential_outcomes,
  assignment,
  estimand_1, estimand_2, estimand_3,
  reveal_outcomes(outcome_variable_names = c("R", "Y")),
  estimator_1, estimator_2, estimator_3)
```

Takeaways

We now diagnose the design:

```
diagnosis <- diagnose_design(
  design, sims = 10000, bootstrap_sims = 1000, diagnosands = diagnosands)
```

Estimand Label	Mean Estimand	Mean Estimate	SE(Mean Estimate)	Bias	SE(Bias)
ATE_R	-0.05	-0.05	0.00	-0.00	0.00
ATE_Y	NA	0.52	0.00	NA	NA
ATE_Y_for_As	0.22	0.52	0.00	0.30	0.00

We learn three things from the design diagnosis. First, as expected, our experiment is unbiased for the average treatment effect on response.

Next, we see that our second inquiry, as well as our diagnostics for it, are undefined. The diagnosis tells us that our definition of potential outcomes produces a definition problem for the estimand. Note that the diagnosands that are defined, including power, depend only on the answer strategy and not on the estimand.

Finally, our third estimand – the average effects for the A types – is defined but our estimates are biased. The reason for this is that we cannot tell from the data which types are the A types: we are not conditioning on the correct subset. Indeed, we are unable to condition on the correct subset. If a subject responds in the treatment group, we don't know if she is an A or a B type; in the control group, we can't tell if a responder is an A or a C type. Our difference-in-means estimator of the ATE on Y among As will be off whenever As have different outcomes from Bs and Cs.

In some cases, the problem might be resolved by changing the inquiry. Closely related estimands can often be defined, perhaps by redefining Y (e.g., emails never sent have a tone of zero). Some redefinitions of the problem, as in the one we examine above, require estimating effects for unobserved subgroups which is a difficult challenge.

Applications

This kind of problem is surprisingly common. Here are three more distinct instances of the problem:

1. Y is the decision to vote Democrat ($Y = 1$) or Republican ($Y = 0$), R is the decision to turn out to vote and Z is a campaign message. The decision to vote may depend on treatment but if subjects do not vote then Y is undefined.

2. Y is the weight of infants, R is whether a child is born and Z is a maternal health intervention. Fertility may depend on treatment but the weight of unborn (possibly never conceived) babies is not defined.

3. Y is the charity to whom contributions are made during fundraising and R is whether anything is contributed and Z is an encouragement to contribute. The identity of beneficiaries is not defined if there are no contributions.

All of these problem exhibit a form of post treatment bias (see section *Post treatment bias*) but the issue goes beyond picking the right estimator. Our problem here is conceptual: the effect of treatment on the outcome just doesn't exist for some subjects.

Exercises

1. The amount of bias on the third estimand depends on both the distribution of types and the correlation of types with the potential outcomes of Y. Modify the declaration so that the estimator of the effect on Y is unbiased, changing only the distribution of types. Repeat the exercise, changing only the correlation of type with the potential outcomes of Y.

2. Try approaching the problem by redefining the inquiry, seeking to assess the effect of treatment on the share of responses with positive tone.

References

Carnegie, Allison and Cyrus Samii. 2017. "International Institutions and Political Liberalization: Evidence from the World Bank Loans Program." *British Journal of Political Science* pp. 1–23.

Gelman, Andrew and Guido Imbens. 2017. "Why high-order polynomials should not be used in regression discontinuity designs." *Journal of Business & Economic Statistics* .

Hahn, Jinyong, Petra Todd and Wilbert Van der Klaauw. 1999. "Evaluating the effect of an antidiscrimination law using a regression-discontinuity design." NBER Working Paper 7131.

Keele, Luke and Rocío Titiunik. 2016. "Natural experiments based on geography." *Political Science Research and Methods* 4(1):65–95.

Mitchell, Ojmarrh, Joshua C. Cochran, Daniel P. Mears and William D. Bales. 2017. "Examining Prison Effects on Recidivism: A Regression Discontinuity Approach." *Justice Quarterly* 34(4):571–596.

Qin, Xuezheng, Castiel Chen Zhuang and Rudai Yang. 2017. "Does the one-child policy improve children's human capital in urban China? A regression discontinuity design." *Journal of Comparative Economics* 45(2):287 – 303.

Sekhon, Jasjeet S. and Rocío Titiunik. 2017. On Interpreting the Regression Discontinuity Design as a Local Experiment. In *Regression Discontinuity Designs*, ed. Thomas B. Fomby, R. Carter Hill, Ivan Jeliazkov, Juan Carlos Escanciano and Eric Hillebrand. Emerald chapter 1, pp. 1–28.