Manipulation Tests in Regression Discontinuity Design: The Need for Equivalence Testing

Jack Fitzgerald

Vrije Universiteit Amsterdam and Tinbergen Institute

June 25, 2025

RV Manipulation at the Cutoff

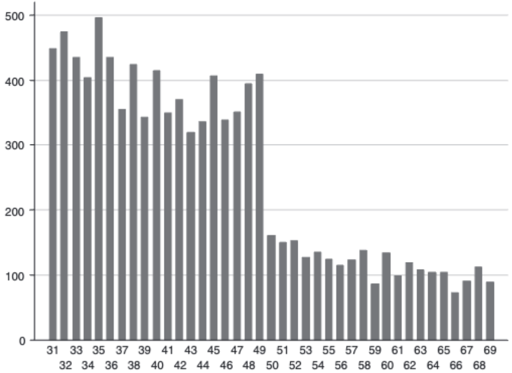


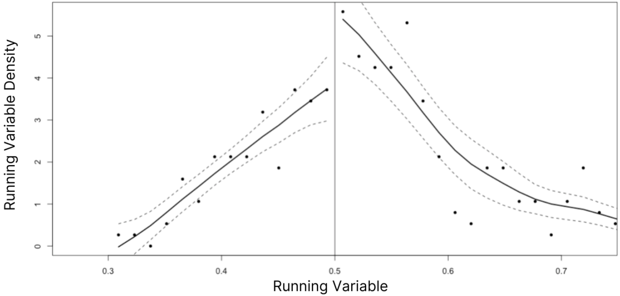
FIGURE 2. NUMBER OF FIRMS BY EMPLOYMENT SIZE IN FRANCE

Source: Garicano, Lelarge, & van Reenen (2016)

Endogenous manipulation of running variable (RV) values near the cutoff induces selection biases

- Agents can often effectively select into/out of treatment

RV Manipulation Tests



RV manipulation tests estimate and assess discontinuities in the RV's density at the cutoff

- Well-known versions include `DCdensity` and `rddensity` (McCrary 2008; Cattaneo, Jansson, & Ma 2018; Cattaneo, Jansson, & Ma 2020)
- Per Web of Science, these tests have over 2100 citations between them

… and How They're Misused

R Code | Stata Code

```
* If necessary, findit rddensity and install the rddensity package
causaldata gov_transfers_density.dta, use clear download

* Limit to the bandwidth ourselves
keep if abs(income_centered) < .02
* Run the discontinuity check
rddensity income_centered, c(0)
```

As expected, we find no statistically significant break in the distribution of income at the cutoff. Hooray!

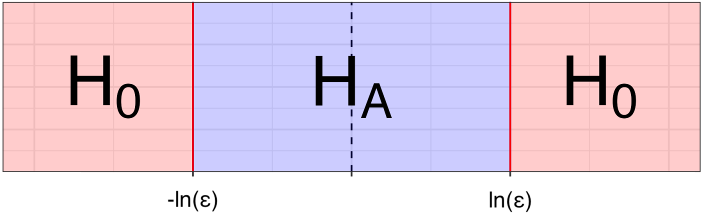
Source: Huntington-Klein (2022)

Unfortunately, researchers (mis)interpret *stat. insig.* manipulation as evidence of *negligible* manipulation

- This is a well-known fallacy (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

Meaningful manipulation may go undetected if these tests are underpowered

An Alternative Testing Framework



Ideal: Stat. sig. evidence that RV manipulation ≈ 0. We can get this using **equivalence testing**:

1. Define the smallest practically/economically significant RV density discontinuities at the cutoff for our given research setting
2. Use interval tests to assess whether the RV density discontinuity at the cutoff is bounded beneath this effect size

This Project

Novel equivalence testing procedure for RV manipulation tests

- Can provide sig. evidence that RV manipulation ≈ 0, which is what applied researchers usually want to show

Empirical evidence of its necessity in applied RDD research

- Replicating 36 published RDD papers shows that $> 44\%$ of RV density discontinuity magnitudes can't be stat. sig. bounded beneath a 50% upward jump

Guidelines and statistical software commands for credible implementation

- `lddtest` command in Stata (available on SSC) and in the `eqtesting` R package (available on CRAN)

Setup (1/2)

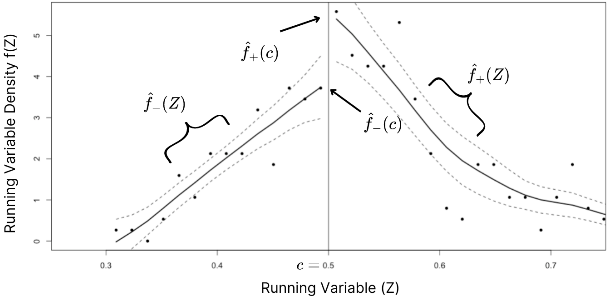
Standard cross-sectional RDD setup (panel setup possible via bootstrap)

- Agents i have some running variable Z_i
- Agents are assigned to treatment if Z_i crosses cutoff c:

$$D_i = \begin{cases} 1 \text{ if } Z_i \geq c \\ 0 \text{ if } Z_i < c \end{cases} \quad \text{or} \quad D_i = \begin{cases} 1 \text{ if } Z_i \leq c \\ 0 \text{ if } Z_i > c \end{cases}$$

- Z_i exhibits probability density function $f(Z_i)$

Setup (2/2)



We'll test for RV manipulation by testing a continuity assumption: $\lim_{Z_i \to c^-} f(Z_i) = \lim_{Z_i \to c^+} f(Z_i)$

- RV manipulation tests estimate density functions on each side of the cutoff, $\hat{f}_-(Z_i)$ and $\hat{f}_+(Z_i)$
- Our estimates of the LHS and RHS density limits are respectively $\hat{f}_-(c)$ and $\hat{f}_+(c)$

The Wrong Hypotheses: Standard NHST

Standard RV manipulation tests effectively assess the hypotheses

$$H_0 : \lim_{Z_i \to c^-} f(Z_i) = \lim_{Z_i \to c^+} f(Z_i)$$
$$H_A : \lim_{Z_i \to c^-} f(Z_i) \neq \lim_{Z_i \to c^+} f(Z_i).$$

There are many problems with this standard NHST approach

- **No burden of proof**: Researchers assume in the null hypotheses that what they want to show is true
- For most researchers, **imprecision is 'good'**
- **Negligible manipulation can be 'significant'** in high-powered research settings

Creates perverse incentives for **'reverse *p*-hacking'** by setting restrictive bandwidths or not reporting RV manipulation tests (see Dreber, Johanneson, & Yang 2024)

The Right Hypotheses: Equivalence Testing

We'll fix these problems by 1) flipping the hypotheses and 2) relaxing the constratins. As a reminder, **standard NHST hypotheses**:

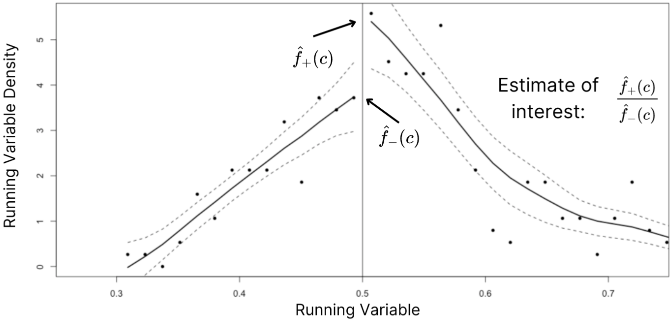
$$H_0 : \lim_{Z_i \to c^-} f(Z_i) = \lim_{Z_i \to c^+} f(Z_i)$$
$$H_A : \lim_{Z_i \to c^-} f(Z_i) \neq \lim_{Z_i \to c^+} f(Z_i).$$

And now **equivalence testing hypotheses**:

$$H_0 : \lim_{Z_i \to c^-} f(Z_i) \not\approx \lim_{Z_i \to c^+} f(Z_i)$$
$$H_A : \lim_{Z_i \to c^-} f(Z_i) \approx \lim_{Z_i \to c^+} f(Z_i).$$

If we can set a range of values wherein the RV's density jump at the cutoff ≈ 0, then we can get stat sig. evidence for H_A with a simple interval test

Step 1: Set the Effect Size Threshold



Set largest practically/economically insignificant RTL density ratio $\epsilon > 1$ for our research setting

- RTL density ratios are useful effect sizes because they are always comparable across datasets
- This threshold can be credibly set by surveying other researchers for their judgments Details

Step 2: Estimate the Logarithmic Density Discontinuity

McCrary's (2008) `DCdensity` procedure estimates **logarithmic density discontinuities**:

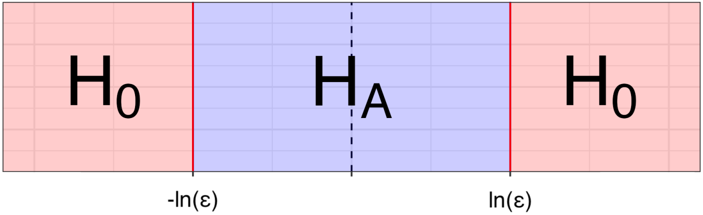
$$\hat{\theta} \equiv \ln\left(\hat{f}_+(c)\right) - \ln\left(\hat{f}_-(c)\right)$$
$$= \ln\left(\frac{\hat{f}_+(c)}{\hat{f}_-(c)}\right)$$

McCrary (2008) also shows that $\hat{\theta}$ is consistent and asymptotically normal

- We can thus use $\hat{\theta}$ and SE $\left(\hat{\theta}\right)$ from `DCdensity` for standard Gaussian inference

I also develop (cluster) bootstrap procedures for finite-sample (cluster-)robust inference

Step 3: Equivalence Testing

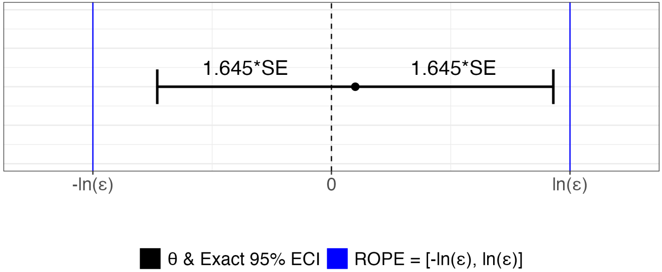


We'll test whether $\hat{\theta}$ is stat. sig. bounded between $-\ln(\epsilon)$ and $\ln(\epsilon)$ w/ two one-sided tests of the form

$$H_0 : \theta < -\ln(\epsilon) \qquad H_0 : \theta > \ln(\epsilon)$$
$$H_A : \theta \geq -\ln(\epsilon) \qquad H_A : \theta \leq \ln(\epsilon)$$

If both tests are stat. sig. at level α, then there's size-α stat. sig. evidence that RV manipulation at the cutoff is practically equal to zero (see Schuirmann 1987; Berger & Hsu 1996) Visualization

Equivalence Confidence Interval (ECI) Approach



$\hat{\theta}$'s $(1-\alpha)$ **equivalence confidence interval (ECI)** is just its $(1-2\alpha)$ CI

- If $\hat{\theta}$'s $(1-\alpha)$ ECI is entirely bounded in $[-\ln(\epsilon), \ln(\epsilon)]$, then we have size-$\alpha$ evidence under the TOST procedure that RV manipulation at the cutoff ≈ 0 (Berger & Hsu 1996)

We can use this for (percentile) bootstrap inference by constructing $(1-\alpha)$ bootstrap ECIs

Replication Data

I leverage replication data from Stommes, Aronow, & Sävje (2023), who run robustness checks on 36 published RDD papers in *AJPS*, *APSR*, and *JOP* from 2009-2018

- Some papers use multiple datasets; I run RV manipulation tests in each dataset (45 in total)

Designs in this dataset include close election designs, spatial discontinuities, and age discontinuities

- Historically popular RVs in economics research (Lee & Lemieux 2010)

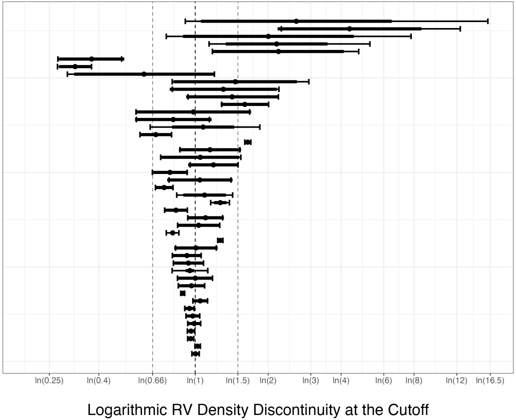
Equivalence Testing Performance

I re-examine these papers with my equivalence-based RV manipulation test, using a lenient threshold of $\epsilon = 1.5$ Why?

- I.e., each test asks: *Can we significantly bound RV manipulation at the cutoff beneath a 50% upward jump/33.3% downward jump?*
- Given the caliber of journals, these RVs should 'pass' this lenient equivalence test

I then compute **equivalence testing failure rates** – the proportion of these equivalence tests that are *not* significant at a 5% level

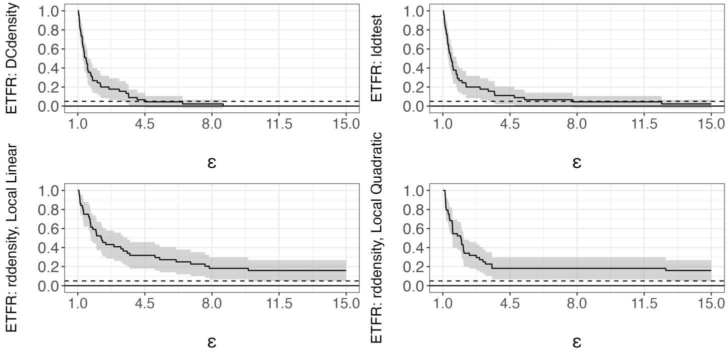
Main Equivalence Testing Failure Rate Estimates



Failure rates for my equivalence-based RV manipulation test range from 44-75%

- **Interpretation:** Over 44% of RV density discontinuity magnitudes at the cutoff can't be significantly bounded beneath a 50% upward jump

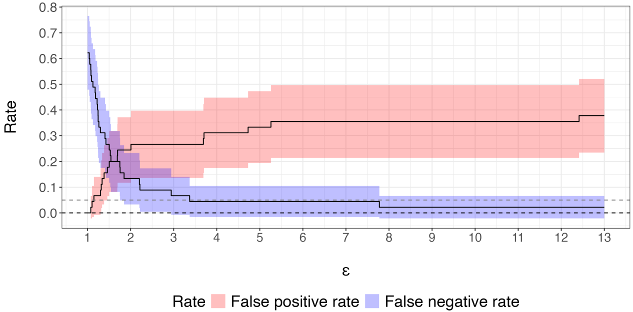
Failure Curves



To obtain 'equivalence testing failure rates' beneath 5%, we'd have to be willing to argue that a 350% upward density jump is practically equal to zero

- **Takeaway:** Meaningful RV manipulation at the cutoff is still a serious problem in RDD research

Confusion Curves



17.7% of LDD estimates at cutoff are **false positives**: Stat. sig., but sig. bounded within $\epsilon \in [2/3, 3/2]$

- Likewise, **26.6%** of LDD estimates at the cutoff are **false negatives**: Not stat. sig., but not sig. bounded within $\epsilon \in [2/3, 3/2]$

Takeaway: Standard NHST often misclassifies the practical significance of RV manipulation at cutoff

Practical Considerations

How do you set the threshold ϵ?

- If we set it ourselves, we'll likely get (reasonable) accusations of *p*-hacking
- But if others set it for us, the threshold is credibly independent of our data

I recommend setting ϵ by **surveying other researchers for their judgments** of the smallest practically/economically significant RV density jump at the cutoff

- Practical using online resources such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019)
- Data from these researcher surveys can be useful for reasons beyond this test

An alternative is partial identification robust to RV manipulation (Gerard, Rokkanen, & Rothe 2020)

- **Takeaway:** If you're going to decide whether RV manipulation is meaningful using a test, then use an equivalence test

Equivalence testing is likely useful for many econometric specification tests! Step 1

Thank You For Your Attention!

These Slides

Paper: https://doi.org/10.31222/osf.io/2dgrp_v1

I am on the job market in 2025-2026!

Website: https://jack-fitzgerald.github.io
Email: j.f.fitzgerald@vu.nl

References I

Berger, R. L. and J. C. Hsu (1996, Nov).
Bioequivalence trials, intersection-union tests and equivalence confidence sets.
Statistical Science 11(4).

Cattaneo, M. D., M. Jansson, and X. Ma (2018, Mar).
Manipulation testing based on density discontinuity.
The Stata Journal 18(1), 234–261.

Cattaneo, M. D., M. Jansson, and X. Ma (2020, Sep).
Simple local polynomial density estimators.
Journal of the American Statistical Association 115(531), 1449–1455.

Chen, H., P. Cohen, and S. Chen (2010, Apr).
How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies.
Communications in Statistics - Simulation and Computation 39(4), 860–864.

References II

Cohen, J. (1988).
Statistical power analysis for the behavioral sciences (2 ed.).
L. Erlbaum Associates.

Cunningham, S. (2021, Aug).
Causal inference: The mixtape (1 ed.).
Yale University Press.

DellaVigna, S., D. Pope, and E. Vivalt (2019, Oct).
Predict science to improve science.
Science 366(6464), 428–429.

Dreber, A., M. Johannesson, and Y. Yang (2024, Mar).
Selective reporting of placebo tests in top economics journals.
Economic Inquiry Forthcoming.

References III

Garicano, L., C. Lelarge, and J. Van Reenen (2016).
Firm size distortions and the productivity distribution: Evidence from france.
American Economic Review 106(11), 3439–3479.

Gerard, F., M. Rokkanen, and C. Rothe (2020, Jul).
Bounds on treatment effects in regression discontinuity designs with a manipulated running variable.
Quantitative Economics 11(3), 839–870.

Hartman, E. (2021, Oct).
Equivalence testing for regression discontinuity designs.
Political Analysis 29(4), 505–521.

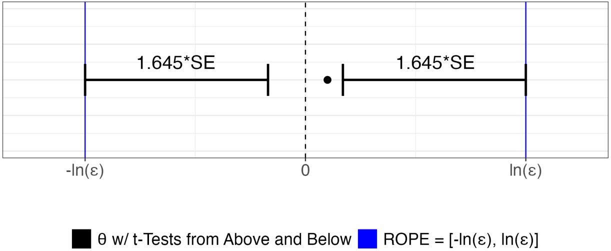
References IV

Huntington-Klein, N. (2022, Aug).
The Effect: An introduction to research design and causality (1 ed.).
CRC Press.

McCrary, J. (2008, Feb).
Manipulation of the running variable in the regression discontinuity design: A density test.
Journal of Econometrics 142(2), 698–714.

Schuirmann, D. J. (1987, Dec).
A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.
Journal of Pharmacokinetics and Biopharmaceutics 15(6), 657–680.

Two One-Sided Tests (TOST) Procedure



In other words, we have stat. sig. evidence at the 5% level that $\theta \approx 0$ if

1. $\hat{\theta}$ is 1.645 SEs above $-\ln(\epsilon)$, **and**
2. $\hat{\theta}$ is 1.645 SEs below $\ln(\epsilon)$

Step 3

Why $\epsilon = 1.5$?

- Chen, Cohen, & Chen (2010) show that an odds ratio of 1.5 corresponds closely w/ a Cohen's (1988) $d = 0.2$, the classic small effect size benchmark
- Same effect size proposed by Hartman (2021)
- Practically large in many research-relevant RDD settings (e.g., elections)

Back