Supplementary Material:
Equivalence Testing for Regression Discontinuity Designs

Erin Hartman

Assistant Professor of Political Science and Statistics, University of California, Los Angeles.
Email: ekhartman@ucla.edu, URL: www.erinhartman.com

Regression discontinuity (RD) designs are increasingly common in political science. They have many advantages, including a known and observable treatment assignment mechanism. The literature has emphasized the need for "falsification tests" and ways to assess the validity of the design. When implementing RD designs, researchers typically rely on two falsification tests, based on empirically testable implications of the identifying assumptions, to argue the design is credible. These tests, one for continuity in the regression function for a pre-treatment covariate, and one for continuity in the density of the forcing variable, use a null of no difference in the parameter of interest at the discontinuity. Common practice can, incorrectly, conflate a failure to reject evidence of a flawed design with evidence that the design is credible. The well known equivalence testing approach addresses these problems, but how to implement equivalence tests in the RD framework is not straightforward. This paper develops two equivalence tests tailored for RD designs that allow researchers to provide statistical evidence that the design is credible. Simulation studies show the superior performance of equivalence-based tests over tests-of-difference, as used in current practice. The tests are applied to the close elections RD data presented in Eggers et al. (2015) and Caughey and Sekhon (2011a).

SI-1 Simulation Details

SI-1.1 Data-Generating Process Details

Step 1 For $N \in \{50, 100, 1000, 5000, 10000\}$

Step 2

Scenario 1: $z_\tau \in \{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$ and $d_\tau = 1$

Scenario 2: $z_\tau = 0$ and $d_\tau \in \{1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7\}$

Step 3 Repeat 1000 times:

- Draw:

$$x \sim N(0, 1)$$

Construct discontinuity in density (d_τ):

$$x[x < 0] = x[x < 0] * d_\tau$$

Construct Z with discontinuous jump (z_τ):

$$\phi \sim N(0, 1.295)$$

$$y = \begin{cases} 48 + 12.7(\frac{x}{10}) + 71.8(\frac{x}{10})^2 + 202.1(\frac{x}{10})^3 + 215.4(\frac{x}{10})^4 + 73.3(\frac{x}{10})^5 + \phi, & x < 0 \\ \tau + 48 + 8.4(\frac{x}{10}) - 30(\frac{x}{10})^2 + 79.9(\frac{x}{10})^3 - 90.1(\frac{x}{10})^4 + 35.6(\frac{x}{10})^5 + \phi, & x \geq 0 \end{cases}$$

- Conduct the following test for sorting:
  - Equivalence density test ($\epsilon_L = 2/3, \epsilon_U = 1.5$)
  - Cattaneo, Jansson, and Ma (2019) difference-based density test
  - McCrary (2008) difference-based density test
- Conduct the following test for continuity:
  - Equivalence continuity test ($\epsilon_L = -2.5, \epsilon_U = 2.5$)
  - Interval inclusion continuity test ($\epsilon_L = -2.5, \epsilon_U = 2.5$)
  - Calonico, Cattaneo, and Titiunik (2014) difference-based test for continuity

SI-1.2 Example of Data-Generating Process

Figure SI-1 shows an example of the simulated data generating process. Details can be found in the appendix. The left column shows scenarios in which there is no sorting, where as the right column shows sorting in which individuals can increase their probability of treatment. The upper row exhibits continuity in variable Z, whereas the lower row has a discontinuity of 2, amounting to about 20% of the range of Z.

When there is a discontinuous jump, you can see that the conditional expectation function for the treated units jumps at the cutoff. Changes in density are visible as a discontinuous density of points just above the cutoff.

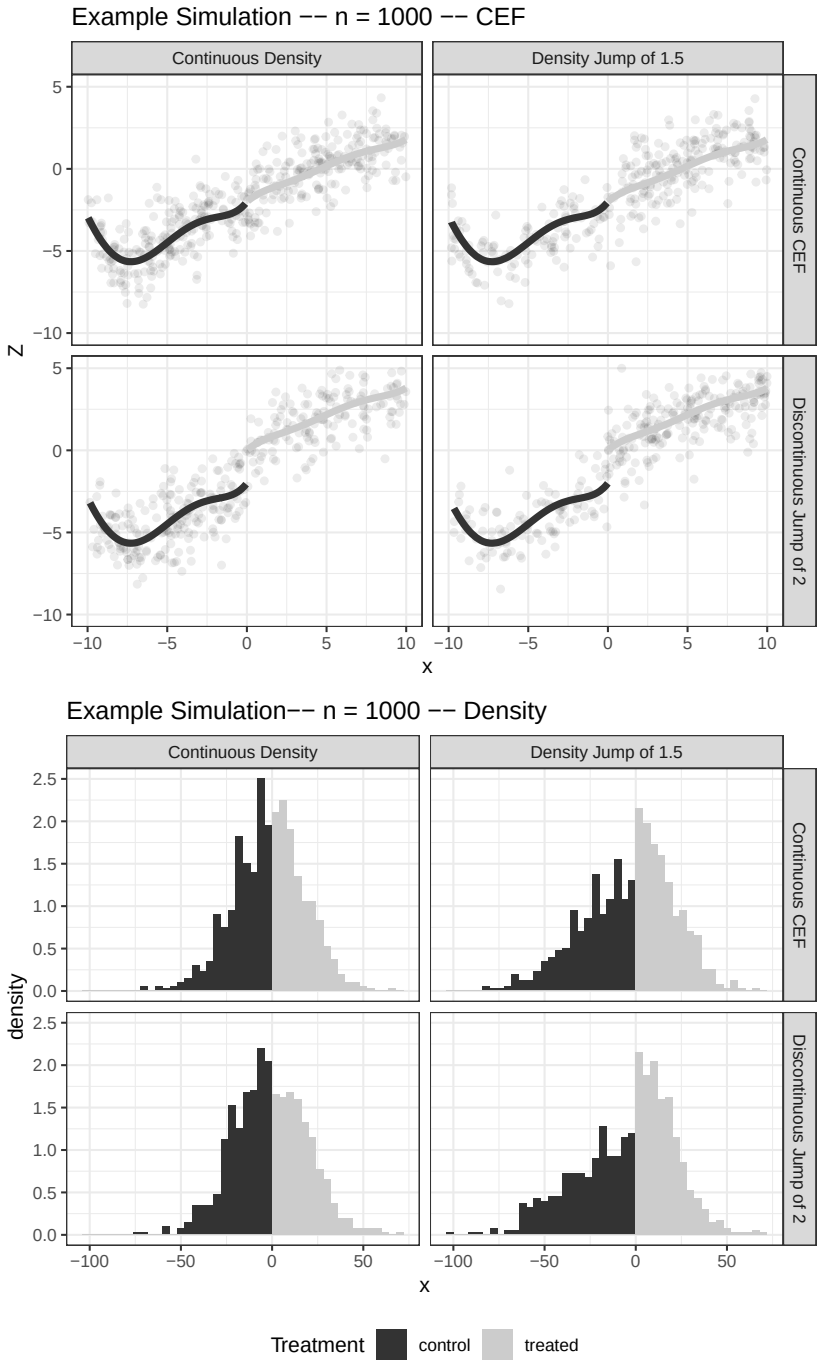

Figure SI-1: Sample draws from the simulated data-generating process. The top panels show the estimation of the regression function of a pre-treatment covariate, and the lower panels show the density of observations. The left columns have no sorting, and the right columns have sorting. Top rows have a continuous regression function, whereas the bottom rows have a discontinuous jump in the regression function at the cut-point.

SI-1.3 Re-analysis of Caughey and Sekhon, 2011a Falsification Tests

As an example of testing across numerous pre-treatment outcomes, I conduct a re-analysis of the 25 pre-treatment covariates originally analyzed in Caughey and Sekhon (2011a) for US House races from 1942-2008.[1] Figure SI-2 presents the results using the equivalence-based tests. Following De la Cuesta and Imai (2016), I standardize all non-binary variables. Following the recommendations in Wellek (2010), I use the conservative equivalence range of ± 0.36 standard deviations for the standardized non-binary variables, and a range of ± 0.1 for binary variables. The observed mean difference and equivalence confidence interval are presented on the scale of the original variable. Overall, using the local linear regression, similar to De la Cuesta and Imai (2016), I find less evidence of an invalid design. However, when applying the multiple testing correction, I find that there are still 8 outcomes that fail to reject the null of an invalid design, and it is concerning that this includes the probability of a democratic win at $t-1$ (e.g. incumbency).

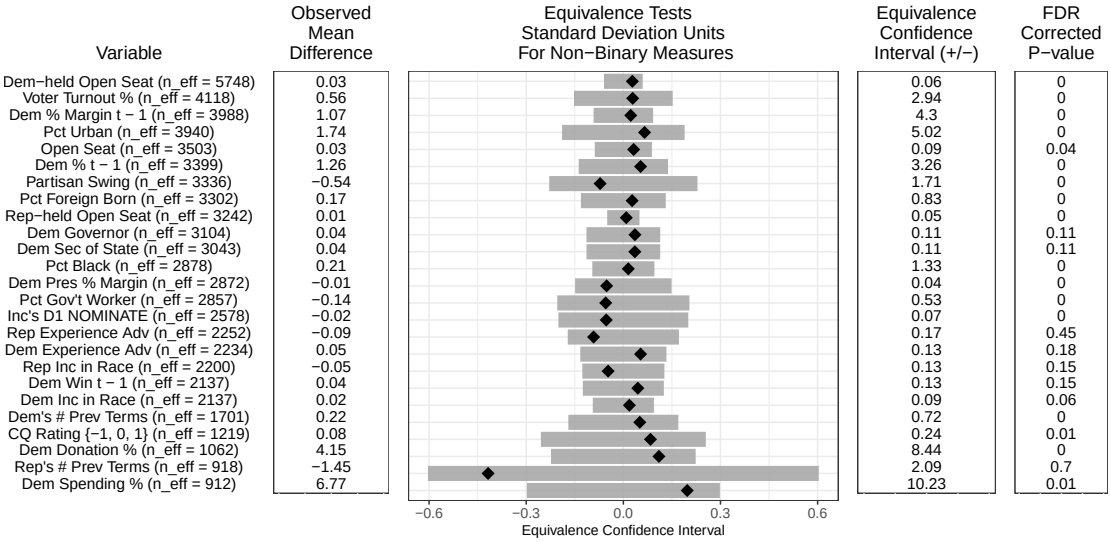

Figure SI-2: Equivalence test for continuity in the Caughey and Sekhon (2011a) data, sorted by effective sample size. Non-binary variables are standardized. The equivalence range is 0.36 standard deviations for non-binary variables, and 0.1 for binary variables. The observed mean difference and equivalence confidence interval are presented on the scale of the original variable. Black diamonds correspond to the point estimate. Gray bars indicate the equivalence confidence interval. The p-value includes a false discovery race correction.

Many of the variables that fail to reject the null of data inconsistent with a valid design are binary, indicating that, perhaps, the equivalence range for binary variables is too conservative, or the range for non-binary variables is too large. For example, if the equivalence range for non-binary variables is set at 0.2, an additional three variables would fail to reject the null of a discontinuity, before the multiple testing correction. Given the sensitivity of the p-values to the definition of the equivalence range, we can focus on the equivalence confidence range. A researcher should evaluate each variable to determine if the equivalence confidence range is sufficiently small to mitigate concerns for bias.

[1] Data available from Caughey and Sekhon (2011b).

References

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik (2014). "Robust nonparametric confidence intervals for regression-discontinuity designs". In: *Econometrica* 82.6, pp. 2295–2326.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma (2019). "Simple local polynomial density estimators". In: *Journal of the American Statistical Association*, pp. 1–7.

Caughey, Devin and Jasjeet S Sekhon (2011a). "Elections and the regression discontinuity design: Lessons from close US house races, 1942–2008". In: *Political Analysis* 19.4, pp. 385–408.

Caughey, Devin and Jasjeet S. Sekhon (2011b). "Replication data for: Elections and the Regression-Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008". Version DRAFT VERSION. In: DOI: 10.7910/DVN/8EYYA2.

De la Cuesta, Brandon and Kosuke Imai (2016). "Misunderstandings about the regression discontinuity design in the study of close elections". In: *Annual Review of Political Science* 19, pp. 375–396.

Eggers, Andrew C et al. (Jan. 2015). "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races". In: *American Journal of Political Science* 59.1, pp. 259–274.

McCrary, Justin (2008). "Manipulation of the running variable in the regression discontinuity design: a density test". In: *Journal Of Econometrics* 142.2, pp. 698–714.

Wellek, Stefan (June 2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority, Second Edition*. CRC Press.