Bertanha, Marinho; Moreira, Marcelo J.

Working Paper

Impossible inference in econometrics: Theory and applications

cemmap working paper, No. CWP02/19

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Bertanha, Marinho; Moreira, Marcelo J. (2019) : Impossible inference in econometrics: Theory and applications, cemmap working paper, No. CWP02/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.0219

This Version is available at:
https://hdl.handle.net/10419/211095

cemmap
centre for microdata methods and practice

Impossible inference in econometrics: theory and applications

Marinho Bertanha
Marcelo J. Moreira

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP02/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Impossible Inference in Econometrics: Theory and Applications*

Marinho Bertanha† *University of Notre Dame*

Marcelo J. Moreira‡ *FGV EPGE*

This version: November 14, 2018

First version: October 11, 2016

This paper studies models in which hypothesis tests have trivial power, that is, power smaller than size. This testing impossibility, or impossibility type A, arises when any alternative is *not distinguishable* from the null. We also study settings where it is impossible to have almost surely bounded confidence sets for a parameter of interest. This second type of impossibility (type B) occurs under a condition weaker than the condition for type A impossibility: the parameter of interest must be *nearly unidentified*. Our theoretical framework connects many existing publications on impossible inference that rely on different notions of topologies to show models are not distinguishable or nearly unidentified. We also derive both types of impossibility using the weak topology induced by convergence in distribution. Impossibility in the weak topology is often easier to prove, it is applicable for many widely-used tests, and it is useful for robust hypothesis testing. We conclude by demonstrating impossible inference in multiple economic applications of models with discontinuity and time-series models.

Keywords: hypothesis tests, confidence intervals, weak identification, regression discontinuity

JEL Classification: C12, C14, C31

*We thank Tim Armstrong, Leandro Gorno, and anonymous referees for helpful comments and suggestions. Bertanha gratefully acknowledges support from CORE-UcLouvain, and Moreira acknowledges the research support of CNPq and FAPERJ. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

†Department of Economics, 3060 Jenkins Nanovic Halls, Notre Dame, IN 46556. Email: mbertanha@nd.edu. Website: www.nd.edu/∼mbertanh.

‡Praia de Botafogo, 190 11th floor, Rio de Janeiro - RJ 22250-040, Brazil. Email: mjmoreira@fgv.br. Website: epge.fgv.br/en/professor/marcelo-moreira.

1 Introduction

The goal of most empirical studies is to estimate parameters of a population statistical model using a random sample of data. The difference between estimates and population parameters is uncertain because sample data do not have all the information about the population. Statistical inference provides methods for quantifying this uncertainty. Typical approaches include hypothesis testing and confidence sets. In a hypothesis test, the researcher divides all possible population models into two sets of models. The null set includes the models which the researcher suspects to be false. The alternative set includes all other likely models. It is desirable to control the size of the test, that is, the error probability of rejecting the null set when the null set contains the true model. A powerful test has a small error probability of failing to reject the null set when the true model is outside the null set. Another approach is to use the data to build a confidence set for the unknown value of parameters of the true model. The researcher needs to control the error probability that the confidence set excludes the true value. Error probabilities must be controlled uniformly over the entire set of likely models. This paper studies necessary and sufficient conditions for the impossibility of controlling error probabilities of hypothesis tests and confidence sets.

Previous work demonstrates the impossibility of controlling error probabilities of tests and confidence sets in specific settings. There are essentially two types of impossibility found in the literature. The first type of impossibility says that any hypothesis test has power limited by size. That is, it is impossible to find a powerful test that controls size. We call this impossibility type A. The second type of impossibility states that any confidence set that is almost surely (a.s.) bounded has error probability arbitrarily equal to one (i.e. zero confidence level). In other words, it is impossible for finite bounds to contain the true value of parameters with high probability. We call this impossibility type B. Despite being related, both types of impossibility often appear disconnected in the existing literature.

The first contribution of this paper is to connect the literature on impossible inference and study the relationships between type A and type B impossibility. Figure 1 at the end of this introduction summarizes the literature along with novel relationships derived in this paper. To the best of our knowledge, impossibility type A dates back to the 1950s. In a classic paper, Bahadur and Savage (1956) show both types of impossibility in the population mean case. Any test for distinguishing zero mean from non-zero mean distributions has power limited by size; and any a.s. bounded confidence interval for the population mean has error probability equal to one.[1] Bahadur and Savage (1956) employ the Total Variation

[1]The impossibility typically arises due to the richness of models in the class of all likely models. Impossibility does not arise if we restrict the class to have only one model, which is the same as pointwise inference. Uniform inference over a larger class of models is important because the researcher typically does not know

(TV) metric to measure the distance between any two distributions. We refer to this notion of distance as *strong distance*. They show that the null set of distributions with a certain mean is dense with respect to (wrt) the TV metric in the set of distributions with all possible means.

In fact, impossibility type A is very much related to the density of the convex hull of the null set in the set of all likely models wrt the TV metric. Kraft (1955) targets the problem of testing any two sets of distributions and arrives at an important generalization of the theory of Bahadur and Savage (1956). Kraft's Theorem 5 gives a necessary and sufficient condition for the existence of a test whose minimum power is strictly greater than its size. Such tests exist if, and only if, the minimal TV distance between the convex hulls of the null and alternative sets is bounded away from zero. Kraft attributes the theorem to Le Cam, and an analogous version of his theorem appears in Theorem 2.1 of Ingster and Suslina (2003). Romano (2004) demonstrates that the null set being dense in the set of all likely models wrt the TV metric is a sufficient condition for impossibility A. We derive a corollary of Kraft's Theorem 5 that says that the convex hull of the null set being dense in the set of all likely models wrt the TV metric is a necessary and sufficient condition for impossibility type A. The null set being dense implies that the convex hull of the null set is dense. Our corollary connects the literature on impossibility type A wrt the TV metric.

A different branch of the econometrics literature focuses on impossibility type B of confidence sets for a given parameter of interest, e.g. mean or regression slope. In the population mean case, Bahadur and Savage (1956) arrive at impossibility type B by demonstrating the following fact. For any mean value m, the set of distributions with mean equal to m is dense in the set of all likely models wrt the TV metric. This is stronger than the sufficient condition for impossibility type B used by Gleser and Hwang (1987). Gleser and Hwang (1987) consider classes of models indexed by parameters in a Euclidean space. They obtain impossibility type B whenever there exists one distribution P^* such that, for every value of the parameter of interest, P^* is approximately equal to distributions with that value of the parameter of interest wrt the TV metric.[2] As with impossibility type A, impossibility type B also holds if the condition of Gleser and Hwang (1987) holds over the convexified space of distributions, which is a weaker sufficient condition. Donoho (1988) also provides

all aspects of the model at hand. For example, instruments could be weak, and if we incorrectly assume they are always strong, pointwise inference conclusions are quite misleading.

[2]Gleser and Hwang (1987) restrict their analysis to distributions that have parametric density functions wrt the same sigma-finite measure. Two distributions are indistinguishable if their density functions are approximately the same pointwise in the data. In their setting, pointwise approximation in density functions is the same as approximation in the TV metric. However, pointwise approximation in density functions is still stronger than convergence in distribution. See Proposition 2.29 and Corollary 2.30, Van der Vaart (2000).

type B impossibility for parameters of interest that are functionals of distributions satisfying a dense graph condition in the TV metric. One example of such a functional is the derivative of a probability density function (PDF). Althought it is impossible to obtain a.s. bounded confidence sets, Donoho (1988) shows that it is possible to build valid one-sided lower-bounded confidence intervals in some cases. Similarly, Low (1997) demonstrates the impossibility of adaptation gains for the length of confidence intervals on linear functionals of non-parametric functions. Low's lower bound on the expected length of confidence intervals grows to infinity as the class of possible models increases.

Dufour (1997) generalizes Gleser and Hwang (1987) to classes of models indexed by parameters in general metric spaces. Dufour (1997) also notes that impossibility type B implies that tests constructed from a.s. bounded confidence sets fail to control size. Unlike all authors mentioned thus far, Dufour (1997) relies on a notion of distance much weaker than the TV metric, which is the notion of distance behind weak convergence or convergence in distribution. He obtains impossibility type B whenever there exists one distribution P^* such that, for every value of the parameter of interest, there exists a sequence of distributions with that value of the parameter of interest that converges in distribution to P^*. The weaker notion of distance restricts the analysis to confidence sets whose boundary has zero probability under P^*. The Lévy-Prokhorov (LP) metric is known to metrize weak convergence. We refer to this notion of distance as *weak distance*. We demonstrate the impossibility type B of Dufour also holds after convexifying the space of distributions.

We revisit impossibility type A when distributions are indistinguishable in the LP metric as opposed to the TV metric. We find that impossibility type A applies to all tests that are a.s. continuous under alternative distributions. A sufficient condition is that the convex hull of the null set is dense in the set of all likely models wrt the LP metric. On the one hand, the LP metric does not yield impossibility type A for every test function. On the other hand, the class of a.s. continuous tests includes the vast majority of tests used in empirical studies. Convergence in the TV metric always implies convergence in the LP metric. The converse is not true, except in more restricted settings. For example, if convergence in distribution implies uniform convergence of probability density functions (PDF), then Scheffé's Theorem implies convergence in the TV metric (Corollary 2.30 of Van der Vaart (2000)).

The second contribution of this paper is to note that a weaker notion of distance, such as the LP metric, brings further insights into the problem of impossible inference. First, it is often easier to prove convergence of models in terms of the weak distance than it is in the strong distance. Application of arguments similar to Portmanteau's theorem immediately yields the LP version of impossible inference in an important class of models in economics that rely on discontinuities. Second, the use of the LP metric helps researchers look for tests with

non-trivial power. If models are indistinguishable wrt the LP metric, but distinguishable wrt the TV metric, we show that a useful test must necessarily be a.s. discontinuous. Third, the LP metric can be a sensible choice of distance to study hypothesis tests that are robust to small model departures. For example, consider the null set of continuous distributions versus the alternative set of discrete distributions with finite support in the rational numbers. It is possible to approximate any such discrete distribution by a sequence of continuous distributions in the LP metric. Hence, it is impossible to powerfully test these sets with a.s. continuous tests. On the other hand, a positive TV distance between null and alternative leads to a perfect test that rejects the null if observations take rational values. Robustness leads us to ask whether observing rational numbers is indeed evidence against the null hypothesis, or simply a matter of rounding or measurement error. The same problem may arise in reduced-form or structural econometric models, even when the degree of misspecification is small. Depending on the problem at hand, we may want to look for tests that separate the closure of each hypothesis wrt the LP metric.

The third contribution of this paper is to point out impossible inference in microeconometric models based on discontinuities and macroeconometric models of time series. Numerous microeconometric analyses identify parameters of interest by relying on natural discontinuities in the distribution of variables. This is the case of Regression Discontinuity Designs (RDD), an extremely popular identification strategy in economics. In RDD, the assignment of individuals into a program changes discontinuously at a cutoff point in a variable such as age or test score, as for Hahn, Todd, and Van der Klaauw (2001) and Imbens and Lemieux (2008). For example, Schmieder, von Wachter, and Bender (2012) study individuals whose duration of unemployment insurance jumps wrt age. Jacob and Lefgren (2004) analyze the effect of students' participation in summer school, which changes discontinuously wrt test scores. Assuming all other characteristics vary smoothly at the cutoff, the effect of the summer school on future performance is captured by a discontinuous change in average performance at the cutoff. A fundamental assumption for identification is that performance varies smoothly with test scores, after controlling for summer school. Models with continuous effects are well-approximated by models with discontinuous effects. Kamat (2018) uses the TV metric to show that the current practice of tests in RDD suffers from impossibility type A. We revisit his result using the LP metric, and we show that impossibility type B also holds in RDD.

A Monte Carlo experiment shows that the usual implementation of Wald tests in RDD, as suggested by Calonico, Cattaneo, and Titiunik (2014), may have size above the desired significance level, even under sensible model restrictions. We rely on data-generating processes that are consistent with the empirical example of Lee (2008). Moreover, the simulations

show that the Wald test has very little power, even after artificially controlling size. Slope restrictions on the conditional mean functions do not correct the finite sample failure of the typical Wald test.

In other applications, researchers assume a discontinuous change in unobserved characteristics of individuals at given points. This is the idea of bunching, widely exploited in economics. Bunching may occur because of a discontinuous change in incentives or a natural restriction on variables. For example, the distribution of reported income may display a non-zero probability at points where the income tax rates change, as in Saez (2010); or, the distribution of average smoking per day has a non-zero mass at zero smoking. We show that the problem of testing for existence of bunching in a scalar variable suffers from type A impossibility for a.s. continuous tests but not for discontinuous tests.

Caetano (2015) uses the conditional distribution of variables with bunching and proposes an exogeneity test without instrumental variables. The key insight is that bunching in the distribution of an outcome variable given a treatment variable constitutes evidence of endogeneity. For example, consider the problem of determining the effect of smoking on birth weight. A crucial assumption is that birth weight varies smoothly with smoking while controlling for all other factors. Under this assumption, bunching is equivalent to the observed average birth weight being discontinuous at zero smoking. The exogeneity test looks for such discontinuity as evidence of endogeneity. Our point is that models in which birth weight is highly sloped or even discontinuous based on smoking are indistinguishable from smooth models. Therefore, we find the exogeneity test has power limited by size. The current implementation of tests for the size of discontinuity leads to bounded confidence sets, so it also fails to control size.

In addition to these applications with discontinuities, we verify the existence of impossible inference in a macroeconometrics example where data are continuously distributed. We first show that the choice of the weak versus the strong distance connects to the work of Peter J. Huber on robust statistics and leads us to look at the closure of the set of covariance-stationary time-series processes wrt the LP metric. This closure contains error-duration models and Compound Poisson models. Our theory implies that it is impossible to robustly distinguish these models from covariance-stationary models, even with discontinuous tests.

It is important to emphasize that our goal with these applications is not to say that valid inference is never possible. Rather, we point practitioners to the need of either restricting the class of models under consideration or the null hypothesis being tested. In the RDD case, impossibility vanish if we restrict the variation of conditional mean functions on either side of the cutoff. Kamat (2018) demonstrates the asymptotic validity of Wald tests under uniform bounds on the derivatives of the conditional mean functions. Armstrong and Kolesár

(2018) derive minimax optimal-length confidence intervals in the case of a convex class of conditional mean functions, which covers most smoothness or shape assumptions used in econometrics. Alternatively, instead of limiting the whole class of models, researchers may consider null hypotheses that restrict other aspects of the model, beyond simply the effect at the threshold. For example, the null of smooth models with zero effect, or the null of no treatment spillover, do not suffer from type A impossibility. We expand this discussion in Section 4.1 with empirical examples in RDD.

The rest of this paper is divided as follows. Section 2 sets up a statistical framework for testing and building confidence sets. It presents necessary and sufficient conditions for impossible inference in general non-parametric settings. Section 3 connects the LP metric to robust hypothesis testing. Section 4 gives multiple economic applications where both types of impossibility arise. Section 5 presents a Monte Carlo simulation for an empirical application of RDD. Section 6 concludes. An appendix contains all formal proofs. Figure 1 (on the next page) summarizes the literature on impossible inference, along with implications of this paper.

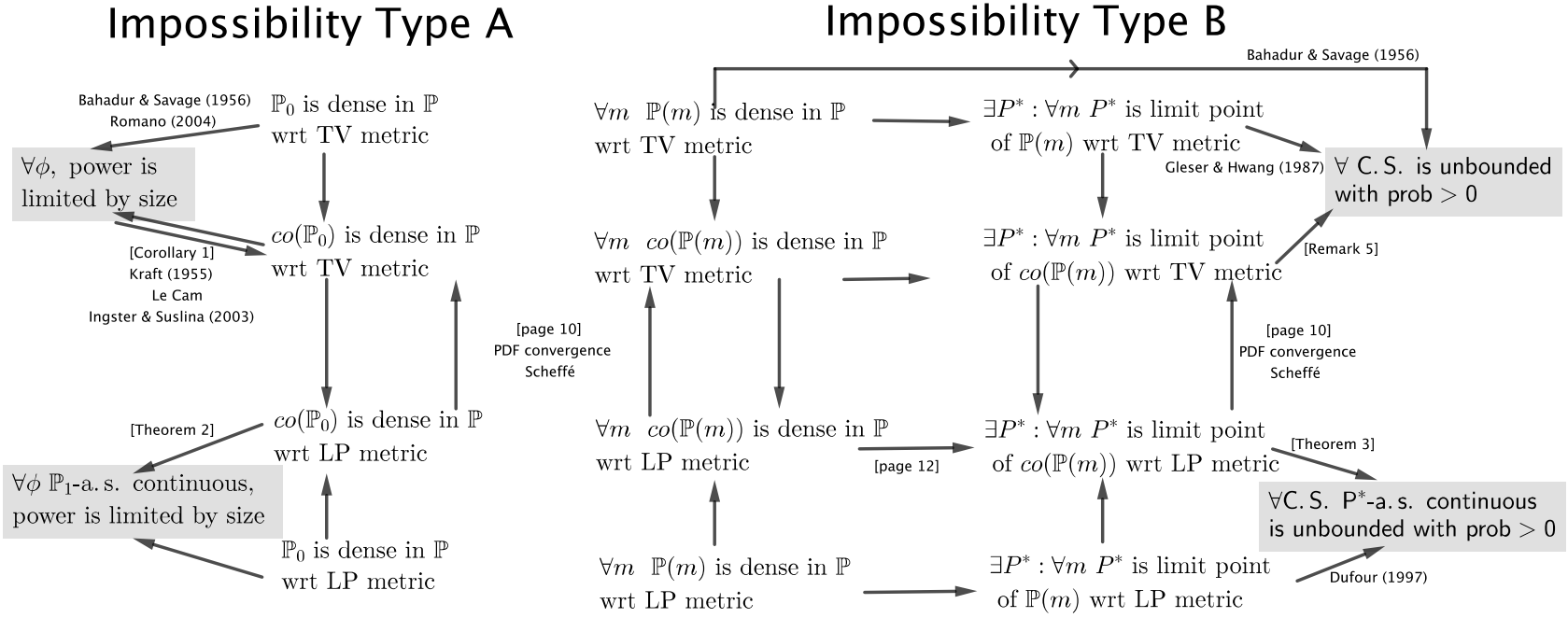
2 Impossible Inference

The researcher has a sample of n observations $Z = (Z_1, \ldots, Z_n)$ that take values in $\mathcal{Z}$, a subset of the Euclidean space $\mathbb{R}^{n \times l}$. The data Z follow a distribution P, and the set of all possible distributions considered by the researcher is $\mathbb{P}$. Every probability distribution $P \in \mathbb{P}$ is defined on the same sample space $\mathcal{Z}$ with Borel sigma-algebra $\mathcal{B}$. It is assumed that all distributions in $\mathbb{P}$ are absolutely continuous wrt the same sigma-finite measure.[3] We are interested in testing the null hypothesis $H_0 : P \in \mathbb{P}_0$ versus the alternative hypothesis $H_1 : P \in \mathbb{P}_1$ for a partition $\mathbb{P}_0, \mathbb{P}_1$ of $\mathbb{P}$. We characterize a hypothesis test by a function of the data $\phi : \mathcal{Z} \to [0, 1]$. If ϕ takes on only the values 0 and 1, the test is said to be *non-randomized*, but said to be *randomized* otherwise. Given a sample Z, we reject the null H_0 if the function $\phi(Z)$ equals one, but we fail to reject H_0 if $\phi(Z) = 0$. If the function $\phi(Z)$ is between 0 and 1, we reject the null with probability $\phi(Z)$ conditional on Z. The unconditional probability of rejecting the null hypothesis under distribution $P \in \mathbb{P}$ is denoted $\mathbb{E}_P[\phi]$.

The size of the test ϕ is $\sup_{P \in \mathbb{P}_0} \mathbb{E}_P[\phi]$. The power of the test under distribution $Q \in \mathbb{P}_1$ is given by $\mathbb{E}_Q[\phi]$. We say a test ϕ has power limited by size when $\sup_{Q \in \mathbb{P}_1} \mathbb{E}_Q[\phi] \leq$

[3]Examples include Lebesgue measure for continuous distributions; counting measure for discrete distributions; and sum of Lebesgue and counting measures for mixed continuous-discrete distributions.

Figure 1: Impossibility Literature Diagram



Notes: the diagram illustrates the relationships between the different versions of impossibility found in the literature. Arrows without labels or arrows with references in square brackets are relationships made explicit by this paper. Impossibility type A says that every test function ϕ has maximum power less than or equal to size. The set of all likely models is $\mathbb{P}$, which is the union of models under the null $\mathbb{P}_0$ and alternative hypothesis $\mathbb{P}_1$. A test ϕ is said to be $\mathbb{P}_1$-a.s. continuous if the set of discontinuity points of ϕ has zero probability under every $Q \in \mathbb{P}_1$. The set $co(\mathbb{P}_0)$ denotes the convex hull of $\mathbb{P}_0$. The TV metric is the Total Variation metric (Equation (2.2) in Section 2). The LP metric is the Lévy-Prokhorov metric (Equation (2.3) in Section 2). The set $\mathbb{P}_0$ is dense in $\mathbb{P}$ wrt a metric $d(\cdot, \cdot)$ if, for every $Q \in \mathbb{P}_1$, there exists a sequence $\{P_k\}_k \subseteq \mathbb{P}_0$ such that $d(P_k, Q) \to 0$. Impossibility type B says that every confidence set (C.S.) is unbounded with positive probability for some distributions in $\mathbb{P}$. The subset $\mathbb{P}(m)$ denotes all models P such that a parameter of interest $\mu(P) = m$. A confidence set is a function $C(\cdot)$ of the data Z. The C.S. is said to be P^*-a.s. continuous if the boundary of the set $\{m \in C(Z)\}$ has zero probability under P^* for every value of m in the range of $\mu(\cdot)$. The model P^* is a limit point of $\mathbb{P}(m)$ wrt a metric $d(\cdot, \cdot)$ if there exists a sequence $\{P_k\}_k \subseteq \mathbb{P}(m)$ such that $d(P_k, P^*) \to 0$. If $\mathbb{P}_0$ is dense in $\mathbb{P}$, then $co(\mathbb{P}_0)$ is dense in $\mathbb{P}$ because $\mathbb{P}_0 \subseteq co(\mathbb{P}_0)$. Similarly, if $P^\star$ is limit point of $\mathbb{P}(m)$, then $P^\star$ is also limit point of $co(\mathbb{P}(m))$. Convergence in TV implies convergence in LP. The converse is not generally true. See page 10 for sufficient conditions for the converse to hold.

$\sup_{P\in\mathbb{P}_0}\mathbb{E}_P[\phi]$. Define $co(\mathbb{P}')$ to be the convex hull of an arbitrary subset $\mathbb{P}'\subseteq\mathbb{P}$. That is,

$$co(\mathbb{P}')=\left\{P^*:\ P^*=\sum_{i=1}^{N}\alpha_iP_i,\ \text{for some } N\in\mathbb{N}, P_i\in\mathbb{P}'\ \forall i,\right.$$
$$\left.\alpha_i\in[0,1]\ \forall i, \sum_{i=1}^{N}\alpha_i=1\right\}. \tag{2.1}$$

A small distance between models in $\mathbb{P}_0$ and $\mathbb{P}_1$ determines testing impossibility. There exist various notions of distance to measure the difference between two distributions P and Q. A common choice in the literature on testing impossibility is the Total Variation (TV) metric $d_{TV}(P,Q)$:

$$d_{TV}(P,Q)=\sup_{B\in\mathcal{B}}|P(B)-Q(B)|. \tag{2.2}$$

Theorem 5 of Kraft (1955) says that there exists a test ϕ with minimum power strictly greater than size if, and only if, there exists $\varepsilon>0$ such that $d_{TV}(P,Q)\geq\varepsilon$ for every $P\in co(\mathbb{P}_0)$ and $Q\in co(\mathbb{P}_1)$. We restate his theorem below for convenience.

Theorem 1. ***(Kraft (1955))*** *Fix $\varepsilon>0$. The following statements are equivalent:*

(a) $\exists\phi\ :\ \inf_{Q\in\mathbb{P}_1}\mathbb{E}_Q\phi\geq\varepsilon+\sup_{P\in\mathbb{P}_0}\mathbb{E}_P\phi$, *and*

(b) $\forall P\in co(\mathbb{P}_0),\ \forall Q\in co(\mathbb{P}_1),\ d_{TV}(P,Q)\geq\varepsilon$.

An important implication of Theorem 1 for impossible inference is that it gives a necessary and sufficient condition in terms of the convex hull of the null set being dense in the set of all likely models wrt the TV metric. In other words, the convex hull $co(\mathbb{P}_0)$ is *indistinguishable* from (or dense in) the set of all likely models wrt the TV metric if, for any $Q\in\mathbb{P}_1$, there exists a sequence $\{P_k\}_{k=1}^{\infty}$ in $co(\mathbb{P}_0)$ such that $d_{TV}(P_k,Q)\to0$. We demonstrate this fact in the corollary below.

Corollary 1. *The following statements are equivalent:*

(a) for every $Q\in\mathbb{P}_1$, there exists a sequence $\{P_k\}_k\subseteq co(\mathbb{P}_0)$ such that $d_{TV}(P_k,Q)\to0$, and

(b) for every ϕ and $Q\in\mathbb{P}_1$, $\mathbb{E}_Q\phi\leq\sup_{P\in co(\mathbb{P}_0)}\mathbb{E}_P\phi$.

The proof of this corollary, as well as all other proofs for the paper, is included in the appendix. The striking result of Kraft (1955) stated in Theorem 1 makes the type A impossibility found by Bahadur and Savage (1956) and Romano (2004) special cases of

Corollary 1. In particular, Theorem 1 of Romano (2004) says that Corollary 1-(a) without convexification is a sufficient condition for Corollary 1-(b). Notably, Romano (2004) finds a positive result for testing population means. He demonstrates that the t-test uniformly controls size in large samples with a very weak uniform integrability type of condition, and that the t-test is also asymptotic minimax optimal.

Dufour (1997) uses the notion of distance associated with weak convergence to derive impossibility type B. We say a sequence $\{P_k\}_{k=1}^{\infty}$ converges in distribution to Q, if, for every $B \in \mathcal{B}$ such that $Q(\partial B) = 0$, $P_k(B) \to Q(B)$. Here, ∂B is the boundary of a Borel set B, that is, the closure of B minus the interior of B. We denote convergence in distribution by $P_k \xrightarrow{d} Q$. Convergence in distribution is equivalent to convergence in the Lévy-Prokhorov (LP) metric (Dudley (1976), Theorem 8.3) :

$$\begin{aligned} d_{LP}(P,Q) = \inf\{\varepsilon > 0 \ : \ P(A) \leq Q(A^{\varepsilon}) + \varepsilon \text{ for } A \in \mathcal{B}\} \\ \text{where } A^{\varepsilon} = \{x : \|x - a\| < \varepsilon \text{ for } a \in A\}, \\ \text{and } \|\cdot\| \text{ is the Euclidean norm on } \mathbb{R}^{n \times l}. \end{aligned} \tag{2.3}$$

Convergence of P_k to Q in the TV metric implies convergence in distribution. The converse does not hold, in general.[4] It is necessary to restrict the class of distributions in order for convergence in the TV metric to imply convergence in distribution. For example, suppose that $P_k \xrightarrow{d} Q$, that these distributions have common support $[a, b]$ and PDFs f_{P_k}, f_Q. Assume further that f_{P_k} converges uniformly over $[a, b]$. Then, f_{P_k} converges uniformly to f_Q (Theorem 7.17 of Rudin (1976)). Convergence of PDFs implies convergence in the TV metric (Scheffé's Theorem, see Corollary 2.30 of Van der Vaart (2000)). For a counter-example where these conditions do not hold, consider the bunching example of Section 4.2. The null is the set of distributions with a continuously differentiable CDF. The alternative is the set of distributions with a mass point at x_0 but continuously differentiable CDF otherwise. For any CDF F_Q in the alternative, there exists a sequence of CDFs F_{P_k} in the null that converges pointwise to F_Q, so that convergence in distribution holds. Convergence in TV does not hold because x_0 has positive probability under Q but zero probability under P_k for every k. It must be the case that the PDFs f_{P_k} do not converge uniformly. In fact, F_Q has a jump discontinuity at x_0, and the derivative of F_{P_k} at x_0 grows without limit as $k \to \infty$.

[4]For example, a standardized binomial variable converges in distribution to a standard normal as the number of trials goes to infinity and the probability of success is fixed. It does not converge in the TV metric because the distance between these two distributions is always equal to one. In fact, consider the event equal to the entire real line minus the support of the binomial distribution. This event has unit probability under the normal distribution, but zero probability under the binomial distribution.

On the one hand, it is true that the zero TV distance provides a necessary and sufficient condition for testing impossibility. On the other hand, there are examples of models with non-zero TV distance where it seems sensible that no powerful test should exist. Section 3 below formalizes this idea, but we start with a simple example for now. Consider the null set of continuous distributions versus the alternative set of discrete distributions with finite support in the rational numbers. It is possible to approximate any such discrete distribution by a sequence of continuous distributions in the LP metric. We are led to think the data generated by a null model is observationally equivalent to data generated by an alternative model. This motivates us to revisit impossibility type A when distributions are indistinguishable in the LP metric.

Assumption 1. *For every $Q \in \mathbb{P}_1$, there exists a sequence $\{P_k\}_{k=1}^{\infty}$ in $co(\mathbb{P}_0)$ such that $P_k \xrightarrow{d} Q$. In other words, the convex hull $co(\mathbb{P}_0)$ is indistinguishable from (or dense in) the set of all likely models wrt the LP metric.*

Assumption 1 is a sufficient condition for impossibility type A, as described in Theorem 2.

Theorem 2. *If Assumption 1 holds, then any hypothesis test $\phi(Z)$ that is a.s. continuous under any $Q \in \mathbb{P}_1$ has power limited by size.*

Remark 1. *As noted by Canay, Santos, and Shaikh (2013), the topology induced by the LP metric is not fine enough to guarantee convergence of integrals of any test function ϕ. Nevertheless, the class of tests that are a.s. continuous under any $Q \in \mathbb{P}_1$ can be very large. For example, take a test that rejects the null when a test statistic is larger than a critical value: $\phi(Z) = I(\psi(Z) > c)$. This test is a.s. continuous if the function ψ is continuous and $Q \in \mathbb{P}_1$ is absolutely continuous wrt the Lebesgue measure. Theorem 2 only requires a.s. continuity under the alternative $\mathbb{P}_1$, and the null $\mathbb{P}_0$ may still contain discrete distributions.*

Remark 2. *We do not need to restrict Theorem 2 to the class of a.s. continuous tests for every case of $\mathbb{P}$. For example, consider $\mathbb{P}$ to be a subset of the parametric exponential family of distributions with parameter θ of finite dimension. Then, for any test ϕ, the power function of ϕ is continuous in θ, and Theorem 2 applies under Assumption 1 (Theorem 2.7.1, Lehmann and Romano (2005)).*

Remark 3. *In many instances, Assumption 1 holds in both directions. That is, $\mathbb{P}_1$ is indistinguishable from $\mathbb{P}_0$, and $\mathbb{P}_0$ is indistinguishable from $\mathbb{P}_1$ in the weak distance. For example, Bahadur and Savage (1956) find that any distribution with mean m is well-approximated by distributions with mean $m' \neq m$, and vice-versa. Section 4 finds the same bidirectionality*

for models with discontinuities. If Assumption 1 holds in both directions, switching the roles of $\mathbb{P}_0$ and $\mathbb{P}_1$ in Theorem 2 shows that power is equal to size.

It is useful to connect our LP version of testing impossibility with the impossibility of controlling error probability of confidence sets found by Gleser and Hwang (1987) and Dufour (1997). Define a real-valued function $\mu : \mathbb{P} \to \mathbb{R}$, for example, mean, variance, median, and so on. The set of distributions $\mathbb{P}$ is implicitly chosen such that μ is well-defined. We consider real-valued functions for simplicity, and results for μ with more general ranges are straightforward to obtain. The range of μ is $\mu(\mathbb{P})$. Suppose we are interested in a confidence set for $\mu(P)$ when the true model is $P \in \mathbb{P}$. A confidence set takes the form of a function $C(Z)$. For a model $P \in \mathbb{P}$, the *coverage probability* of $C(Z)$ is given by $P[\mu(P) \in C(Z)]$. The confidence region $C(Z)$ has *confidence level* $1-\alpha$ (i.e. error probability α) if $C(Z)$ contains $\mu(P)$ with probability at least $1-\alpha$:

$$\inf_{P \in \mathbb{P}} P[\mu(P) \in C(Z)] = 1 - \alpha. \tag{2.4}$$

For any value $m \in \mu(\mathbb{P})$, we define the subset $\mathbb{P}(m)$ by

$$\mathbb{P}(m) = \{P \in \mathbb{P} \ : \ \mu(P) = m\}. \tag{2.5}$$

Impossibility type B says that confidence sets that are a.s. bounded under some distributions in $\mathbb{P}$ have zero confidence level. The next assumption gives a sufficient condition for impossibility type B in terms of the LP metric.

Assumption 2. *There exists a distribution P^* (not necessarily in $\mathbb{P}$) such that for every $m \in \mu(\mathbb{P})$ there exists a sequence $\{P_k\}_k$ in $co(\mathbb{P}(m))$ such that $P_k \xrightarrow{d} P^*$.*

If Assumption 1 holds with $\mathbb{P}_0 = \mathbb{P}(m)$ for every $m \in \mu(\mathbb{P})$, then Assumption 2 holds. In fact, if $\mathbb{P}(m)$ is dense in $\mathbb{P}$ for every m, then Assumption 2 is satisfied for $P^* = Q$ for any $Q \in \mathbb{P}_1$. Some models satisfy Assumption 1 with $\mathbb{P}_0 = \mathbb{P}(m)$ for every $m \in \mu(\mathbb{P})$ and suffer from both types of impossibility. Examples of this case include the problem of testing the mean (Bahadur and Savage (1956)), or the problem of testing the size of the discontinuity in RDD (Section 4.1). Nevertheless, some other models satisfy Assumption 2 but not Assumption 1 for every m. These models suffer from impossibility type B. Examples include the problem of ratio of regression parameters (Gleser and Hwang (1987)), and the problem of weak instruments (Dufour (1997)).

The next theorem encapsulates the impossibility of controlling coverage probabilities found by Gleser and Hwang (1987) and Dufour (1997). It differs from Gleser and Hwang

(1987) because Assumption 2 uses the LP distance. It differs slightly from Dufour (1997) because Assumption 2 is stated in terms of the convex hull of $\mathbb{P}(m)$ rather than simply $\mathbb{P}(m)$.

Theorem 3. *Suppose Assumption 2 holds with P^*. Assume the confidence set $C(Z)$ of Equation (2.9) has confidence level $1-\alpha$, and $P^*(\partial\{m \in C(Z)\}) = 0$ for every $m \in \mu(\mathbb{P})$. Then,*

$$\forall m \in \mu(\mathbb{P}) \ : \ P^*\left[m \in C(Z)\right] \geq 1-\alpha. \tag{2.6}$$

For a set $A \subset \mathbb{R}$, define $U[A] = \sup\{c : c \in A\}$, $L[A] = \inf\{c : c \in A\}$, and $D[A] = U[A] - L[A]$. Assume $\{U[C(Z)] \geq x\}$, $\{L[C(Z)] \leq -x\}$, and $\{D[C(Z)] \geq x\}$ are measurable events for every $x \in [0,\infty]$. If $D[\mu(\mathbb{P})] = \infty$, then

$$P^*\left[D[C(Z)] = \infty\right] \geq 1-\alpha. \tag{2.7}$$

In addition, if $P^*\left[\partial\{D[C(Z)] = \infty\}\right] = 0$, then

$$\forall \varepsilon > 0 \ : \ \sup_{P \in B_\varepsilon(P^*) \cap \mathbb{P}} P\left[D[C(Z)] = \infty\right] \geq 1-\alpha \tag{2.8}$$

where $B_\varepsilon(P^*) = \{P \ : \ d_{LP}(P, P^*) < \varepsilon\}$.

Remark 4. *Part (2.8) above implies the following. If $1-\alpha > 0$, then the confidence set $C(Z)$ is unbounded with strictly positive probability for some $P \in \mathbb{P}$. Alternatively, the contrapositive of part (2.8) says the following. Any confidence set that is a.s. bounded under distributions in $\mathbb{P}$ in a neighborhood of P^* has $1-\alpha = 0$ confidence level.*

Remark 5. *It is possible to obtain a slightly more general version of Theorem 3 using Assumption 2 stated in terms of the TV metric as opposed to the LP metric. In that case, Theorem 3 would be true for confidence sets that do not necessarily satisfy $P^*(\partial\{m \in C(Z)\}) = 0$ and $P^*\left[\partial\{D[C(Z)] = \infty\}\right] = 0$.*

A common way of obtaining confidence sets is to invert hypothesis tests. The function $C(Z)$ is constructed by inverting a test in the following manner. For a given $m \in \mu(\mathbb{P})$, define $\mathbb{P}_{0,m} = \mathbb{P}(m)$ and $\mathbb{P}_{1,m} = \mathbb{P} \setminus \mathbb{P}(m)$, where $A \setminus B$ denotes the remainder of set A after we remove the intersection of set B with set A. If $\phi_m(Z)$ is a test for $\mathbb{P}_{0,m}$ vs $\mathbb{P}_{1,m}$, then

$$C(Z) = \{m \in \mu(\mathbb{P}) \ : \ \phi_m(Z) = 0\}. \tag{2.9}$$

For every $m \in \mu(\mathbb{P})$, the test $\phi_m(Z)$ has size $\alpha(m) = \sup_{P \in \mathbb{P}_{0,m}} \mathbb{E}_P\left[\phi_m(Z)\right]$. The confidence level of $C(Z)$ is equal to one minus the supremum of $\alpha(m)$ over $m \in \mu(\mathbb{P})$. The proof

of this claim is found in Lemma 1 in the appendix.

Theorem 3 along with Lemma 1 imply that tests that invert into a.s. bounded confidence sets fail to control size.

Corollary 2. *Suppose Assumption 2 holds, and $\mu(\mathbb{P})$ is unbounded. Let the confidence set $C(Z)$ be constructed from tests $\phi_m(Z)$, as in Equation (2.9). Assume $C(Z)$ has confidence level $1-\alpha$ and satisfies the assumptions of Theorem 3. If $C(Z)$ is a.s. bounded under distributions in $\mathbb{P}$ in a neighborhood of P^*, then $\alpha = 1$. Consequently, for every $\varepsilon > 0$, there exists $m_\varepsilon \in \mu(\mathbb{P})$ such that $\sup_{P \in \mathbb{P}_{0,m_\varepsilon}} \mathbb{E}_P \phi_{m_\varepsilon} > 1 - \varepsilon$.*

Remark 6. *Moreira (2003) provides numerical evidence that Wald tests can have large null rejection probabilities for the null of no causal effect ($m = 0$) in the simultaneous equations model. To show that Wald tests have null rejection probabilities arbitrarily close to one, the hypothesized value m for the null would need to change as well. He also suggests replacing the critical value by a critical value function of the data. This critical value function depends on the hypothesized value m. Our theory shows that this critical value function is unbounded if we change m freely.*

3 Weak Convergence and Robustness

This section presents further motivation for using the LP metric to study impossible inference. It relates the weak topology induced by the LP metric to the theory developed by Peter J. Huber, who is the most prominent researcher in the area of robust statistics. We refer the reader to Huber and Ronchetti (2009) for more details. We start this section with a discussion of robust statistical procedures. An example of impossible robust hypothesis testing in time-series models appears in Section 4.4.

Several statistical procedures are susceptible to small model departures. This perception has led researchers to propose alternative procedures that are less sensitive to the break-down of usual assumptions. Huber studies different ways of defining a set of model departures $\mathbb{P}_\epsilon$. One possibility is to assume that the actual distribution of the data is a mixture of a distribution in $\mathbb{P}$ with a distribution from a more general set of models $\mathbb{M}$. In other words, $\mathbb{P}$ may be contaminated with probability ϵ:

$$\mathbb{P}_\epsilon = \{H \in \mathbb{M};\ \exists F \in \mathbb{P} \text{ and } \exists G \in \mathbb{M};\ H = (1-\epsilon)\, F + \epsilon\, G\}\,, \tag{3.1}$$

where $\mathbb{M}$ is larger than the original $\mathbb{P}$. Estimators or tests are said to be robust if they have minimax properties over the set of model departures $\mathbb{P}_\epsilon$. To highlight the importance of

robust procedures, we briefly discuss two examples.

The first example of a robust procedure involves point-estimation. The researcher has a sample of n iid observations $Z_i \in \mathbb{R}^l$, $i = 1, \ldots, n$. The set of joint probability distributions $\mathbb{P}$ is indexed by a parameter θ and admits marginal densities $p(Z_i; \theta)$ wrt the same dominating measure (e.g. Lebesgue). The maximum likelihood estimator (MLE) then minimizes

$$\sum\nolimits_{i=1}^{n} - \ln p(Z_i; \theta).$$

This estimator $\widehat{\theta}$ solves

$$\sum\nolimits_{i=1}^{n} - \frac{\partial p\left(Z_i; \widehat{\theta}\right)}{\partial \theta} \cdot \frac{1}{p\left(Z_i; \widehat{\theta}\right)} = 0.$$

Under the usual regularity conditions, $\widehat{\theta}$ is consistent, asymptotically normal, and efficient within the class of regular estimators.

A common choice for $\mathbb{M}$ is the set of distributions with symmetric, thick-tailed densities. It is well-known that optimal procedures derived under Gaussian distributions (sample drawn from $\mathbb{P}$) break down if there is a probability ϵ of observing outliers (sample drawn from $\mathbb{M}$). Huber (1964) suggests M-estimators. To give a specific example of a robust M-estimator, consider the regression model

$$Y_i = X_i'\theta + U_i,$$

where we observe $Z_i = (Y_i, X_i)$ but do not observe the zero-mean normal errors U_i. The MLE $\widehat{\theta}$ minimizes

$$\sum\nolimits_{i=1}^{n} (Y_i - X_i'\theta)^2,$$

and satisfies

$$\sum\nolimits_{i=1}^{n} X_i \left(Y_i - X_i'\widehat{\theta}\right) = 0.$$

More generally, a M-estimator $\widehat{\theta}$ minimizes

$$\sum\nolimits_{i=1}^{n} \rho(Y_i - X_i'\theta),$$

and satisfies

$$\sum\nolimits_{i=1}^{n} X_i \, \psi\left(Y_i - X_i'\widehat{\theta}\right) = 0$$

for choices of functions ρ and ψ. In the MLE case above, $\rho(u) = u^2$ and $\psi(u) = u$.

An M-estimator $\widehat{\theta}$ is said to be asymptotically minimax optimal among a class of estimators if it minimizes the maximal asymptotic variance over distributions in $\mathbb{P}_\varepsilon$. The

M-estimator associated with the functions

$$\rho_k(u) = \begin{cases} u^2/2 & \text{if } |u| \leq k \\ k|u| - u^2/2 & \text{if } |u| > k \end{cases} \quad \text{and } \psi_k(u) = \max\{-k, \min(k, u)\} \tag{3.2}$$

are known to be asymptotically minimax optimal for model contamination. The constant k depends on the deviations ϵ in (3.1). As $\epsilon \to 0$, the truncation parameter $k \to \infty$. As the model departure is small, the M-estimator approaches the MLE estimator. If $\epsilon \to 1$, the parameter $k \to 0$. As the contamination is arbitrarily large, the M-estimator approaches the least absolute deviation (LAD) estimator.

The second example of a robust procedure is in hypothesis testing. Consider the problem of testing a simple null P_0 against a simple alternative P_1. Assume both P_0 and P_1 have densities p_0 and p_1 wrt the Lebesgue measure. For a sample $X = (X_1, ..., X_n)$, the likelihood ratio (LR) test rejects the null if and only if

$$\prod_{i=1}^n \frac{p_1(X_i)}{p_0(X_i)} > c_\alpha,$$

where c_α is the $1-\alpha$ quantile of the distribution of the left-hand side under the null. The Neyman-Pearson Lemma asserts that the LR test is optimal, as it maximizes power within the class of tests with correct size α.

Similar to model departures in the point-estimation example above, we consider the possibility that the null and alternative hypotheses are misspecified. The ϵ-contaminated null and alternatives are

$$\mathbb{P}_{i,\epsilon} = \{H \in \mathbb{M}; \exists F \in \mathbb{P}_i \text{ and } \exists G \in \mathbb{M}; H = (1-\epsilon)F + \epsilon G\}, \tag{3.3}$$

for $i = 0, 1$. The new sets $\mathbb{P}_{0,\epsilon}$ and $\mathbb{P}_{1,\epsilon}$ allow for local departures for arbitrary distributions in $\mathbb{M}$. A minimax optimal hypothesis test maximizes the minimal power over $\mathbb{P}_{1,\epsilon}$ within the class of tests with correct size over $\mathbb{P}_{0,\epsilon}$.

Huber (1965) shows that the minimax test to these model departures rejects the null if and only if

$$\prod_{i=1}^n \pi_k\left(\frac{p_1(x)}{p_0(x)}\right) > c_\alpha,$$

where

$$\pi_k(w) = \max\{k_1, \min(k_2, w)\},$$

for constants $k = (k_1, k_2)$ that depend on the size of the departure ϵ. As $\epsilon \to 0$, the constant k_1 approaches zero, and k_2 diverges to infinity. Hence, as the departure decreases, the robust

test approaches the usual LR test.

In the two examples of robust procedures given above, arbitrarily small model departures ($\epsilon \to 0$) do not affect the solution to the minimax problem. That is, as ϵ approaches zero, the robust estimator converges to the MLE, and the robust test converges to the LR test. These limiting solutions remain the same, even if we ignore model departures ($\varepsilon = 0$). These inference procedures target a parameter which is a functional of the underlying distribution $P \in \mathbb{P}_\varepsilon$. These functionals vary smoothly wrt ε as $\varepsilon \to 0$. Robustness is associated with smoothness of the functional, but such smoothness may not always occur in other settings.

Our work on impossible inference and the different metrics connects to Huber's work on robustness when we look at the following definition of model departure. For a metric space $(\mathbb{M}, d)$, the set of model departures is defined as an ε-neighborhood of $\mathbb{P}$:

$$\mathbb{P}_\epsilon = \{H \in \mathbb{M}; \exists F \in \mathbb{P} \text{ s.t. } d(F, H) \leq \epsilon\}.$$

The set $\mathbb{P}_\epsilon$ is closed.[5] The set $\bigcap_{\epsilon>0} \mathbb{P}_\epsilon$ is also closed and coincides with $\overline{\mathbb{P}}$, the closure of $\mathbb{P}$. Hence, the set $\overline{\mathbb{P}}$ is the minimal set of the Huber-type model departures $\mathbb{P}_\epsilon$ containing $\mathbb{P}$.

The minimal set of model departures crucially depends on a choice for the metric d. Aside from the Lévy-Prokhorov (LP) and the Total Variation (TV) metrics, there are many choices of metrics for spaces of probability measures: Kolmogorov, Hellinger, and Wasserstein, among others. Gibbs and Su (2002) provide a review. Which metric shall we choose? The choice of the metric on the space of models $\mathbb{M}$ induces a topology $\mathcal{V}$ on that space. Parameters of interest are functionals $\mu : (\mathbb{M}, \mathcal{V}) \to (\mathbb{R}, \mathcal{U})$ where $\mathcal{U}$ is the topology on the $\mathbb{R}$ space. Robustness is about the continuity of the functional μ, which crucially depends on the choices of topologies $\mathcal{V}$ and $\mathcal{U}$. For the real line, it seems reasonable to work with the smallest topology involving all open sets of the form (a, b). However, there are many choices of topologies for the set of measures $\mathbb{M}$.

As the set of continuous functionals μ grows, the topology becomes finer on the domain of μ. Let us consider a simple example to illustrate this point. Take two topological spaces, $(\mathbb{R}, \mathcal{V})$ and $(\mathbb{R}, \mathcal{U})$, and a function $\psi(x) = x$. Continuity of this simple function requires $\psi^{-1}(U) \in \mathcal{V}$ for every open set $U \in \mathcal{U}$. Take $U = (0, 1)$, then $\psi^{-1}(U) = (0, 1)$. If we choose the coarsest topology $\mathcal{V} = \{\emptyset, \mathbb{R}\}$, then even this simple function is not continuous. It seems reasonable to require all linear functions to be continuous. If the topology $\mathcal{V}$ is generated by all open sets of the form (a, b), then all linear functions are continuous. Of course, other non-linear functions may be continuous as well; e.g., $\psi(x) = x^2$. This example makes the

[5] In fact, take an arbitrary convergent sequence $H_n \to H$, such that $H_n \in \mathbb{P}_\epsilon \; \forall n$. To show $H \in \mathbb{P}_\epsilon$, pick an arbitrary $F \in \mathbb{P}$. It is true that $d(H_n, F) \leq \epsilon \; \forall n$. Therefore, $d(F, H) \leq d(F, H_n) + d(H_n, H) \leq \epsilon + d(H_n, H)$. Taking the limit as $n \to \infty$ gives $d(F, H) \leq \epsilon$.

point that continuity depends heavily on the topology $\mathcal{V}$ associated to the domain of the function. If a function is continuous for a topology $\mathcal{V}$, then it is also continuous for a finer topology. This goes back to the discussion of robustness as continuity of a functional μ.

If we choose a fine topology, then many statistical procedures will be deemed robust, because many functionals will be continuous. If we choose a coarse topology, then fewer statistical procedures will be robust. However, if a statistical procedure is robust in the coarse topology, it is also robust in the fine topology. Among the commonly-used notions of distance in measure spaces, the notion of distance behind weak convergence or convergence in distribution induces the coarsest topology. The LP metric is a notion of distance that metrizes weak convergence (Dudley (1976), Theorem 8.3). To be conservative, if we were to choose one metric, we would choose one that metrizes weak convergence. After all, if a functional is continuous wrt the topology induced by the LP metric, it is also continuous wrt the stronger topologies induced by the Kolmogorov or TV metrics.

Another question is whether we should be stricter with robustness and look for an even weaker topology than the weak topology induced by the LP metric. In perfect analogy to the real line example, the weak topology is the coarsest topology that guarantees continuity for all functionals of the form

$$\mu(P) = \int g \, dP, \tag{3.4}$$

for g bounded and continuous. It seems reasonable, after all, to require μ to be continuous when g is a bounded and continuous function. If we choose a weaker topology, then not even μ of this form will be continuous.

In hypothesis testing, robustness over a minimal set of model departures motivates testing $\overline{\mathbb{P}}_0$ against $\overline{\mathbb{P}}_1$ instead of testing $\mathbb{P}_0$ against $\mathbb{P}_1$. Allowing for robustified hypotheses $\overline{\mathbb{P}}_0$ and $\overline{\mathbb{P}}_1$ potentially protects us against numerical approximation errors, misspecified models, measurement errors, and optimization frictions, among other deviations from the set of models we are testing. Robustness of inference procedures for μ that are as simple as (3.4) requires a topology no weaker than the topology induced by the LP metric. Therefore, we use the LP metric to define the closure of a set for robust hypothesis testing. We give two simple examples to strengthen the argument of why the LP metric may be a sensible choice.

The first example of robust hypothesis testing using the LP metric compares extremely simple discrete distributions under both null and alternative hypotheses. Take X to be a Bernoulli random variable and $X_n = X + 1/(1+n)$ for $n \in \mathbb{N}$. Let P^X denote the distribution of X. The minimal TV distance between $\mathbb{P}_0 = \left\{P^{X_n} \text{ for } n \in \mathbb{N}\right\}$ and $\mathbb{P}_1 = \left\{P^X\right\}$ is equal to one. According to Theorem 5 of Kraft (1955), there exists a test for $\mathbb{P}_0$ vs $\mathbb{P}_1$ with non-trivial power. For example, define a test which rejects the null if we observe the values 0 or 1, but

fails to reject the null otherwise. This test has size equal to zero and power equal to one. Should we take the values 0 and 1 as evidence against the null? Or should we think instead that the null could have led to those same values, for all practical purposes? In this example, we note that $\mathbb{P}_1 \subset \overline{\mathbb{P}}_0$, where the closure is defined wrt the LP metric.[6] Hence, the minimal TV distance between $\overline{\mathbb{P}}_0$ and $\mathbb{P}_1$ is equal to zero. After we robustify the null set to $\overline{\mathbb{P}}_0$, it becomes impossible to find any test with power greater than size (Corollary 1). However, if we define the closure of $\mathbb{P}_0$ wrt the TV metric, say $\overline{\mathbb{P}}_0^{TV}$, the minimal TV distance between $\overline{\mathbb{P}}_0^{TV}$ and $\mathbb{P}_1$ is non-zero, which means it is still possible to powerfully distinguish these sets.

The second example of robust hypothesis testing using the LP metric uses the multinomial approximation to continuous distributions. Take $\mathbb{P}_0$ as the collection of multinomial distributions, with each support being a finite subset of rational numbers. Let $\mathbb{P}_1$ be the set of continuous distributions. The minimal TV distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ is one, and it is possible to powerfully distinguish these sets. Robustness leads us to ask whether observing rational numbers is indeed evidence of the null hypothesis, or simply a matter of rounding or measurement error. The closure $\overline{\mathbb{P}}_0$ wrt the LP metric contains continuous distributions, and the minimal TV distance between $\mathbb{P}_1$ and $\overline{\mathbb{P}}_0$ is zero. After we robustify the null set to $\overline{\mathbb{P}}_0$, it becomes impossible to powerfully test these hypotheses.

Both in the Bernoulli and multinomial examples, it becomes clear that the LP closure of the null set robustifies the testing procedure. The next step is to check the TV distance between the robustified null and alternative sets as a way to search for robust tests with non-trivial power. The use of the TV metric in the second step is justified by a corollary of Theorem 5 of Kraft (1955). Corollary 1 demonstrates that a necessary and sufficient condition for the existence of tests with non-trivial power is that the null set is not dense in the set of all distributions wrt the TV metric.

4 Applications

In this section, we apply our theory to multiple economic examples. The first three examples are of models with discontinuities: RDD, bunching in a scalar variable, and exogeneity tests based on bunching. In these settings, the proof of the LP version of impossible inference follows arguments similar to Portmanteau's Theorem. That is, the indicator functions are approximately the same as the steep continuous functions using the weak distance. The problem of testing for the existence of bunching in a scalar variable differs from the other applications with discontinuities because there exists a discontinuous powerful test. A fourth

[6] In fact, $F_n(x) = P(X_n \leq x) = P(X \leq x - 1/n)$, $F_n(x) \to P(X < x) = F(x^-)$, $F(x^-) \neq F(x) \Leftrightarrow x \in \{0, 1\}$ where $\{0, 1\}$ are the only discontinuity points of F, so P^X is a limit point of $\mathbb{P}_0$.

example is in time series; it connects the LP version of impossible inference to Huber's work on robust statistics. This connection leads to the conclusion that it is impossible to powerfully discriminate error-duration or Compound Poisson models from covariance-stationary models.

4.1 Regression Discontinuity and Kink Designs

The first example is the Regression Discontinuity Design (RDD), first formalized by Hahn, Todd, and Van der Klaauw (2001) (HTV01). RDD has had an enormous impact in applied research in various fields of economics. Applications of RDD started gaining popularity in economics in the 1990s. Influential papers include those of Black (1999), who studies the effect of quality of school districts on house prices, where quality changes discontinuously across district boundaries; Angrist and Lavy (1999), who measure the effect of class sizes on academic performance, where size varies discontinuously with enrollment; and Lee (2008), who analyzes US House of Representatives elections and incumbency, where election victory is discontinuous on the share of votes.

Recent theoretical contributions include the study of rate optimality of RDD estimators by Porter (2003) and the data-driven optimal bandwidth rules by Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014). RDD identifies causal effects local to a cutoff value; several authors develop conditions for extrapolating local effects farther away from the cutoff. These include estimation of derivatives of the treatment effect at the cutoff by Dong (2016) and Dong and Lewbel (2015); tests for homogeneity of treatment effects in fuzzy RDD by Bertanha and Imbens (2018); and estimation of average treatment effects in RDD with variation in cutoff values by Bertanha (2017). All these theoretical contributions rely on point identification and inference, and they are subject to both types of impossibility. The current practice of testing and building confidence intervals relies on Wald test statistics $(t(Z) - m)/s(Z)$, where $t(Z)$ and $s(Z)$ are a.s. continuous and bounded in the data. For a choice of critical value z, hypothesis tests $\phi(Z) = \mathbb{I}\{|(t(Z)-m)/s(Z)| > z\}$ are a.s. continuous when the data is continuously distributed. Confidence intervals $C(Z) = \{t(Z) - s(Z)z \leq m \leq t(Z) + s(Z)z\}$ have a.s. bounded length $2s(Z)z$.

The setup of RDD follows the potential outcome framework. For each individual $i = 1, \ldots, n$, define four primitive random variables $D_i, X_i, Y_i(0), Y_i(1)$. These variables are independent and identically distributed. The variable D_i takes values in $\{0, 1\}$ and indicates treatment status. The real-valued variables $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes, respectively, if untreated and treated. Finally, the forcing variable X_i represents a real-valued characteristic of the individual that is not affected by the treatment. The forcing

variable has a continuous PDF $f(x)$ with interval support equal to $\mathbb{X}$. The econometrician observes X_i, D_i, and only one of the two potential outcomes for each individual: $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. For simplicity, we consider the sharp RDD case, but it is straightforward to generalize our results to the fuzzy case. In the sharp case, agents receive the treatment if, and only if, the forcing variable is greater than or equal to a fixed policy cutoff c in the interior of support $\mathbb{X}$. Hence, $D_i = \mathbb{I}\{X_i \geq c\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function.

We focus on average treatment effects. In RDD settings, identification of average effects is typically obtained only at the cutoff value after assuming continuity of average potential outcomes, conditional on the forcing variable. In other words, we assume that $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$ are bounded continuous functions of x. HTV01 show that this leads to identification of the parameter of interest:

$$m = \mathbb{E}\left[Y_i(1) - Y_i(0)|X_i = c\right] = \lim_{x \downarrow c} \mathbb{E}\left[Y_i|X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i|X_i = x\right]. \tag{4.1}$$

Let $\mathcal{G}$ denote the space of all functions $g : \mathbb{X} \to \mathbb{R}$ that are bounded, and that are infinitely many times continuously differentiable in every $x \in \mathbb{X} \setminus \{c\}$. The notation $\mathbb{X} \setminus \{c\}$ represents the set with every point of $\mathbb{X}$ except for c. Continuity of functions in $\mathcal{G}$ suffices to show impossible inference in this section. Nevertheless, non-parametric estimators of the size of the discontinuity m typically assume that functions in g are continuously differentiable of first or second order. We impose that functions in $\mathcal{G}$ are continuously differentiable of infinite order, to demonstrate that both types of impossibility hold even in this more restricted class of functions. The size of the discontinuity m at the cutoff may take any value in $\mathbb{R}$.

Each individual pair of variables $Z_i = (X_i, Y_i)$ is iid as P. The family of all possible models for P is denoted as

$$\mathbb{P} = \{P : (X_i, Y_i) \sim P,\ \exists g \in \mathcal{G} \text{ s.t. } \mathbb{E}_P[Y_i|X_i = x] = g(x)\}. \tag{4.2}$$

The local average causal effect is the function of the distribution of the data $P \in \mathbb{P}$ given by (4.1), provided the identification assumptions of HTV01 hold. The parameter m of the size of the discontinuity is weakly identified in the set of possible true models $\mathbb{P}$. Intuitively, any conditional mean function $\mathbb{E}[Y_i|X_i = x]$ that is continuous except for a jump discontinuity at $x = c$ is well-approximated by a sequence of continuous conditional mean functions. The reasoning behind this approximation is similar to the proof of part of Portmanteau's theorem (Theorem 25.8, Billingsley (2008)). It is known that, if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for every bounded function f that is a.s. continuous under the distribution of X, then $X_n \xrightarrow{d} X$. The proof of Corollary 3 uses an infinitely continuously differentiable function f that is

approximately equal to an indicator function.

Corollary 3. *Assumption 1 is satisfied for $\mathbb{P}_{0,m}$ $\forall m \in \mathbb{R}$, and Theorems 2 and 3 apply to the RDD case. Namely, (i) a.s. continuous tests $\phi_m(Z)$ for the value of the discontinuity m have power limited by size; and (ii) confidence sets for the value of the discontinuity m and with finite expected length have zero confidence level.*

Remark 7. *Corollary 3 also applies to quantile treatment effects by simply changing the definition of the functional $\mu(\mathbb{P})$ to be the difference in side limits of a conditional τ-th quantile $Q_\tau(Y_i|X_i = x)$ at $x = c$. This contrasts with the problem of testing unconditional quantiles, which does not suffer from impossible inference. See Lehmann and D'Abrera (2006), Tibshirani and Wasserman (1988), and Coudin and Dufour (2009).*

Remark 8. *In the fuzzy RDD case, the treatment effect is equal to the discontinuity in $\mathbb{E}[Y_i|X_i]$ at $X_i = c$ divided by the discontinuity in $\mathbb{E}[D_i|X_i]$ at $X_i = c$. Corollary 3 applies to both of these conditional mean functions, and it leads to impossible inference in the fuzzy RDD case as well. Feir, Lemieux, and Marmer (2016) study weak identification in fuzzy RDD and propose a robust testing procedure. In contrast to Kamat (2018) and to this paper, their source of weak identification comes from an arbitrarily small discontinuity in $\mathbb{E}[D_i|X_i]$ at $X_i = c$.*

The most common inference procedures currently in use in applied research with RDD rely on Wald tests that are a.s. continuous in the data and produce confidence intervals of finite expected length. See Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014) for the most commonly-used inference procedures. Corollary 3 implies that it is impossible to control size of these tests and coverage of these confidence intervals.

Ours is not the first paper to show impossible inference in the RDD case. Kamat (2018) demonstrates the important fact that models with a discontinuity are similar to models without a discontinuity in the TV metric. He applies the testing impossibility of Romano (2004) and finds that tests have power limited by size. Using the graphical intuition of Figure 3, we provide a simpler proof of the same facts, using the weak distance instead of the TV metric. Moreover, we add that confidence intervals produced from Wald tests have zero confidence level. It is worth highlighting the statistics literature on the impossibility of adaptation gains for confidence intervals on linear functionals of non-parametric functions (Low (1997), Cai and Low (2004)). They take confidence intervals with correct coverage over a class of models $\mathbb{P}$ and derive a lower bound for the expected length of any confidence interval under a given model $P \in \mathbb{P}$. As the sample size increases, the rate at which these bounds shrink to zero does not depend on P. In other words, any confidence interval whose

expected length at $P \in \mathbb{P}$ shrinks to zero faster than the lower bound must have incorrect coverage over $\mathbb{P}$. Armstrong and Kolesár (2018) derive a lower bound for the expected length of any confidence set that has correct coverage over $\mathbb{P}$. The lower bound increases to infinity as $\mathbb{P}$ becomes more general which is our impossibility of type B.

On a positive note, the two types of impossibility vanish if we restrict the class of models $\mathbb{P}$. The approximation used to prove Corollary 3 fails if we assume that functions in $\mathcal{G}$ have absolute slopes bounded by a finite constant C on either side of the cutoff. Kamat (2018) shows that Wald tests have correct size asymptotically if the first three derivatives of $g(x)$, as well as conditional moments, are uniformly bounded across $\mathbb{P}$. Armstrong and Kolesár (2018) derive minimax optimal-length confidence intervals for a convex function class $\mathcal{G}$ covering most smoothness or shape assumptions used in econometrics. In the RDD case, they consider functions $g(x)$ such that the p^{th}-order Taylor approximation residual is bounded by Cx^p on either side of the cutoff. In summary, applied researchers should bear in mind that the validity of tests and confidence sets for the value of the discontinuity at the threshold relies heavily on restricting the variation of average outcomes wrt the forcing variable X. For example, consider the analysis of summer-school programs in Chicago by Jacob and Lefgren (2004). The forcing variable X is a standardized reading score determining eligibility for the program, and Y is a standardized test score in math or reading after the program. Looking at their Figures 6 and 7, it seems reasonable to assume that the slope of the conditional mean of Y given X is smaller than one. In other words, an increase in today's reading score by 1 point increases tomorrow's average scores by less than one point.

Restricting the class of models $\mathbb{P}$ is not the only way to construct valid tests in RDD. Another way to approach the problem is to consider null sets $\mathbb{P}_0$ different than those in Corollary 3, where the focus is on the jump discontinuity at the threshold. One example is the null hypothesis that an individual's outcome is solely affected by the treatment he receives and not by the effect of the treatment on neighboring individuals. In the summer-school application, the number of students attending classes in the summer is much smaller than during the school year. It is likely that students in the summer program interact much more with each other, which leads to spillover effects of the treatment. A researcher who desires to test for no spillovers specifies the null hypothesis of independence of Y_i and Y_j, conditional on $X_i = X_j = x$ for any $i \neq j$ and x near the threshold. The alternative hypothesis that outcomes exhibit dependence across treated individuals cannot be approximated by models in the null.

Another example of null hypothesis that is immune to testing impossibility is when absence of treatment effects is equivalent to a smooth conditional mean function. We may define the null hypothesis that g is Lipschitz continuous with some constant C, and the

alternative hypothesis that g is any other function as in Equation (4.2). Settings like this arise when the treatment variable D is a function of the forcing variable X, and this function changes at a known cutoff. This is the case of unemployment benefits in Austria, studied by Card, Lee, Pei, and Weber (2015) (CLPW15). For unemployed individuals that used to earn X less than a threshold c, the unemployment benefit grows with their earnings; otherwise, if they used to earn more than c, they simply receive a fixed benefit regardless of their earnings. CLPW15 find that the unemployment duration does not depend on past earnings for those whose benefit is fixed to the right of the cutoff (see their Figure 3). Moral hazard leads to unemployment duration that increases as benefits increase with income to the left of the cutoff. Therefore, the researcher may specify the null hypothesis of a smooth conditional mean to test for the lack of moral hazard. Rejections may occur because of a sudden change in slope or a jump discontinuity at the threshold, both of which are evidence of a change in behavior regarding job search. Note, however, that the null hypothesis of Lipschitz g is different than the null hypothesis in the so-called Regression Kink Design (RKD) studied by CLPW15. The RKD null states that the first derivative of g is continuous at the threshold, and such null suffers from testing impossibility.

RKD has recently gained popularity in economics. In addition to CLPW15, see Dong (2016), Nielsen, Sørensen, and Taber (2010), and Simonsen, Skipper, and Skipper (2016). The setup is the same as in the RDD case, except that the causal effect of interest is the change in the slope of the conditional mean of outcomes at the threshold. Continuity of the first derivatives $\nabla_x \mathbb{E}[Y_i(1)|X_i = x]$ and $\nabla_x \mathbb{E}[Y_i(0)|X_i = x]$ at the threshold $x = c$ guarantees identification of the average effect. The parameter of interest $m = \mu(P)$ is a function of the distribution of $Z_i = (X_i, Y_i)$:

$$\mu(P) = \nabla_x \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \lim_{x \downarrow c} \nabla_x \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \nabla_x \mathbb{E}[Y_i|X_i = x]. \tag{4.3}$$

The family of all possible distributions of Z_i is defined in a slightly different way than in Equation (4.2):

$$\mathbb{P} = \{P : (X_i, Y_i) \sim P,\ \exists g \in \mathcal{G} \text{ s.t. } \nabla_x \mathbb{E}[Y_i|X_i = x] = g(x)\}. \tag{4.4}$$

Weak identification of μ arises from the fact that any conditional mean function $\mathbb{E}[Y_i|X_i = x]$ with a discontinuous first derivative at $x = c$ is well-approximated by a sequence of continuously differentiable conditional mean functions. Assumption 1 is easily verified using this insight.

Corollary 4. *Assumption 1 is satisfied for $\mathbb{P}_{0,m}$ $\forall m \in \mathbb{R}$, and Theorems 2 and 3 apply to*

RKD. Namely, (i) a.s. continuous tests $\phi_m(Z)$ for the value of the kink discontinuity m have power limited by size; and (ii) confidence sets for the value of the kink discontinuity m and with finite expected length have zero confidence level.

The proof of Corollary 4 follows that of Corollary 3. Simply use the new definitions of $\mathbb{P}$ and $\mu(P)$, and construct the sequence P_k with $\nabla_x \mathbb{E}_{P_k}[Y_i|X_i = x] = g_k(x)$.

4.2 Testing for the Existence of Bunching

The second example applies Theorem 2 to the problem of testing for the existence of bunching in a scalar random variable. Bunching occurs when the distribution of X exhibits a non-zero probability at known point x_0, but it is continuous in a neighborhood of x_0. Bunching in the distribution of a single variable is the object of interest in many empirical studies. For example, Saez (2010) and Kleven and Waseem (2013) rely on the existence of bunching on "reported income" at the boundary of tax brackets to identify the elasticity of reported income wrt tax rates; Goncalves and Mello (2018) use bunching on "charged speed in traffic tickets" to separate lenient from non-lenient police officers and identify racial discrimination; and a standard practice in RDD analyses is to check if the distribution of the forcing variable has bunching at the cutoff, which would count as evidence against the design.

Suppose X is a scalar random variable. In the absence of bunching, assume the CDF of X is continuously differentiable. Testing for bunching amounts to testing whether X has positive probability mass at x_0. Let $\mathbb{P}_0$ be the set of distributions of X with a continuously differentiable CDF. The set $\mathbb{P}_1$ is all mixed continuous-discrete distributions, with one mass point at x_0, but continuously differentiable CDF otherwise.[7] Any distribution Q under the alternative is well-approximated in the LP metric by a sequence of distributions P_k under the null. Therefore, any a.s. continuous test has power limited by size.

Corollary 5. *Assumption 1 is satisfied in the problem of testing for the existence of bunching. Hence, any test $\phi(Z)$ that is a.s. continuous under $\mathbb{P}_1$ has power limited by size.*

There is one interesting feature about this example that is not shared by the RDD and RKD examples of the previous section. In this example, it is not possible to find a sequence P_k under the null that approximates a $Q \in \mathbb{P}_1$ using the TV metric. The event $X = x_0$ always has zero probability under the null, but strictly positive probability under the alternative.

[7] The assumption that the CDF is continuously differentiable is not necessary in this section. We impose this assumption because typical non-parametric density estimators assume a continuous density. The testing impossibility of this section occurs regardless of whether the CDF is assumed continuously differentiable, or simply continuous.

Therefore, $d_{TV}(P,Q) > 0$ for every $P \in \mathbb{P}_0, Q \in \mathbb{P}_1$. Theorem 1 suggests that there exists a test whose maximum power is bigger than size, but our Theorem 2 says this test cannot be a.s. continuous under $\mathbb{P}_1$.

The use of the LP metric, as opposed to the TV metric, leads us to search for tests that are discontinuous under $\mathbb{P}_1$. For a sample with n iid observations X_i, the test $\phi(X_1, \ldots, X_n) = \mathbb{I}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i = x_0\} > 0\right\}$ is discontinuous under $\mathbb{P}_1$. This test has size equal to zero, and power equal to $1-(1-\delta)^n$, where $\delta = \mathbb{P}[X_i = x_0]$.

4.3 Exogeneity Tests Based on Bunching

The third example comes from Caetano (2015), who uses the idea of bunching in a conditional distribution of Y given X to construct an exogeneity test that does not require instrumental variables. It applies to regression models where the distribution of unobserved factors are assumed to be discontinuous wrt an explanatory variable. Of interest is the impact of a scalar explanatory variable X on an outcome variable Y, after controlling for covariates W. For example, suppose we are interested in the effect of average number of cigarettes smoked per day X on birth weight Y, after controlling for mothers' observed characteristics W. Conditional on (X, W), the distribution of mothers' unobserved characteristics U is said to bunch at zero smoking if it changes drastically when we compare non-smoking mothers to mothers that smoke very little. If bunching occurs, then the variable X is endogenous because we cannot separate the effect of smoking on birth weight from the effect of unobserved characteristics on birth weight.

The population model that determines Y is written as $Y = h(X, W) + U$, where U summarizes unobserved confounding factors affecting Y. We are unable to infer bunching on U unless h is assumed continuous on (X, W). Bunching of U wrt X is evidence of local endogeneity of X at $X = 0$. Bunching at 0 implies discontinuity of $\mathbb{E}[U|X = 0, W] - \mathbb{E}[U|X = x, W]$ as $x \downarrow 0$. Continuity of h makes bunching equivalent to a discontinuity of $\mathbb{E}[Y|X = 0, W = w] - \mathbb{E}[Y|X = x, W = w]$ as $x \downarrow 0$ for every w. Caetano (2015) proposes testing

$$\forall w \ \lim_{x \downarrow 0} \mathbb{E}[Y|X = 0, W = w] - \mathbb{E}[Y|X = x, W = w] = 0 \tag{4.5}$$

as a means of testing for local exogeneity of X at $X = 0$. We argue that h may have a high slope on X, or even be discontinuous on X, which makes exogeneity untestable.

The observed data $Z = (Z_1, \ldots, Z_n)$, $Z_i = (X_i, W_i, Y_i)$ is iid with probability P. The support of (X_i, W_i) is denoted $\mathbb{X} \times \mathbb{W}$. The distribution of Y conditional on (X, W) is assumed to be continuous. The distribution of X has non-zero probability at $X = 0$, but it is continuous otherwise. Assume $\exists \delta > 0$ such that $[0, \delta) \subset \mathbb{X}$.

Let $\mathcal{G}$ denote the space of all functions $g : \mathbb{X} \times \mathbb{W} \to \mathbb{R}$ that are bounded and infinitely many times continuously differentiable wrt x over $\{\mathbb{X}\backslash\{0\}\} \times \mathbb{W}$. The size of the discontinuity at $X = 0$ may take any value in $\mathbb{R}$. The family of all possible distributions is denoted as

$$\mathbb{P} = \{P : Z_i \sim P,\ \exists g \in \mathcal{G} \text{ s.t. } \mathbb{E}_P[Y_i|X_i = x, W_i = w] = g(x, w)\}. \tag{4.6}$$

Under local exogeneity of X, the function $\tau_P(w) = \mathbb{E}_P[Y_i|X_i = 0, W_i = w] - \lim_{x \downarrow 0} \mathbb{E}_P[Y_i|X_i = x, W_i = w]$ must be equal to 0 $\forall w \in \mathbb{W}$. In practice, it is convenient to conduct inference on an aggregate of $\tau_P(w)$ over $w \in \mathbb{W}$ instead of on the entire function $\tau_P(w)$. Examples of aggregation include the average of $|\tau_P(W)|$, the square root of the average of $\tau_P(W)^2$, or the supremum of $|\tau_P(w)|$ over $w \in \mathbb{W}$. For the sake of brevity, we choose the second option. For a distribution $P \in \mathbb{P}$, define $\mu(P) = \left[\mathbb{E}_P\left(\tau_P(W)^2\right)\right]^{1/2}$. Local exogeneity corresponds to the test of $\mu(P) = 0$ versus $\mu(P) \neq 0$.

The parameter $\mu(P)$ is weakly identified in the class of models $\mathbb{P}$. Just as in the RDD case, any conditional mean function $\mathbb{E}[Y_i|X_i = x, W_i = w]$ with a discontinuity at $x = 0$ is well-approximated by a sequence of continuous conditional mean functions $\mathbb{E}[Y_i|X_i = x, W_i = w]$. Assumption 1 is verified using the same argument as in the RDD case.

Corollary 6. *Assumption 1 is satisfied for $\mathbb{P}_{0,m}$ $\forall m \in \mathbb{R}$, and Theorems 2 and 3 apply to the case of the local exogeneity test. Namely, (i) a.s. continuous tests $\phi_m(Z)$ for the value of the aggregate discontinuity m have power limited by size; and (ii) confidence sets for the value of the aggregate discontinuity m and with finite expected length have zero confidence level.*

The inference procedures suggested by Caetano (2015) rely on non-parametric local polynomial estimation methods. As in the RDD case, these procedures yield tests that are a.s. continuous in the data and confidence intervals of finite expected length. Corollary 6 implies lack of size control and zero confidence level.

4.4 Time-Series Models

The fourth example illustrates robust hypothesis testing wrt the LP metric, and it is of practical relevance to macroeconomists. Macroeconometrics often uses linear time-series processes. This is motivated by the Wold Representation Theorem, which asserts that every covariance-stationary process x_t can be written as an MA process plus some deterministic term:

$$x_t = B(L)\,\epsilon_t,$$

where L is the lag operator, $B(l) = \sum_{i=0}^{q} b_i \, l^i$, and ϵ_t is an uncorrelated error sequence. A caveat is that the order q needs to be too large to be useful for many applications. The features of MA processes with infinite lag order are well captured by ARMA models

$$A(L) x_t = B(L) \epsilon_t,$$

with small orders for $A(L)$ and $B(L)$, where $A(l) = \sum_{i=0}^{p} a_i \, l^i$. The closure of the set of stationary ARMA(p,q) models with finite order (p, q) does not necessarily contain only stationary models. The simplest example happens when $A(l) = 1 - al$ and $B(l) = 1$. The process is stationary when $|a| < 1$, but it is non-stationary when $a = 1$. This observation led to ARIMA models, which better capture the persistence in time series.

Starting in the 1990s, applied researchers began to realize that ARIMA models themselves have limitations. This led to the development of other stochastic processes, including error duration models, Markov switching models, threshold models, structural breaks, and fractionally integrated processes, among others. This is a vast literature and includes papers by Hamilton (1989), Parke (1999), and Bai and Perron (1998), just to name a few.

A number of authors point out that these different model extensions may not be too far from each other. For example, Perron (1989) shows that integrated processes with drift and stationary models with a broken trend can be easily confused; Parke (1999) points out that the error-duration model encapsulates fractionally integrated series; Granger and Hyung (1999) and Diebold and Inoue (2001) find that linear processes with breaks can be misinterpreted as long-memory models. In these papers, and in most of the related econometrics literature, the focus is on the autocovariance of the stochastic process.

Our discussion of robust hypothesis testing in Section 3 suggests looking at the closure of ARMA processes to distinguish these models from each other. For example, take the problem of testing the null that a process is covariance-stationary, against the alternative that it is an error-duration model or a Compound Poisson model. The existence of a test with non-trivial power requires us to look for the TV distance between these sets of processes. However, the ability to approximate theses processes in the TV distance is often based on quite stringent assumptions. For example, see Barbour and Utev (1999) for the TV approximation of Compound Poisson processes.

The problem of searching for tests for covariance-stationary vs error-duration or Compound Poisson becomes much easier if we focus on the LP metric. To solve this problem, we rely on Bickel and Bühlmann (1996), whose work has been largely ignored in the econometrics literature. They characterize the closure of AR and MA processes wrt the TV and the Mallows metric (also known as the Wasserstein metric). The TV metric is stronger than the

Mallows metric, which in turn is stronger than the LP metric. Indeed, convergence under the Mallows metric implies weak convergence and convergence in second moments; see Bickel and Freedman (1981) and Bickel and Bühlmann (1996). As a result, the closure of stochastic processes wrt the LP metric is larger than the closure wrt the Mallows metric. It turns out that error-duration and Compound Poisson models are in the closure wrt the LP metric. In other words, the robustified null set wrt the LP metric contains the alternative set, and the minimal TV distance between these sets is zero. Hence, all tests for the robustified null have power no larger than size.

Given that the closure of ARMA processes of infinite order is quite rich, we may wonder which hypotheses are testable. Bahadur and Savage (1956) and Romano (2004) point out it is hopeless to test population means, even in the iid case without further moment constraints. Could we try to test quantiles? Peskir (2000) and Shorack and Wellner (2009) provide sufficient conditions for uniform convergence of empirical processes under time dependence. A natural choice for quantile testing is the value at risk (VaR), which is commonly used in the finance literature. It would be interesting to establish the class of empirical processes for which hypotheses for the VaR are testable. We leave this example for future work.

5 Simulations

In this section, we provide Monte Carlo simulations to illustrate the impossibility of testing within the context of RDD. We find that the Wald test fails to control size uniformly under the null hypothesis. We use a data-generating process (DGP) based on an empirical example. Lack of size control occurs even for DGPs that are consistent with the data. Moreover, the simulations also show that the Wald test has very little power after artificially controlling size. For the sake of brevity, we focus on the RDD case, and we expect similar findings for the RKD and Exogeneity Test cases.

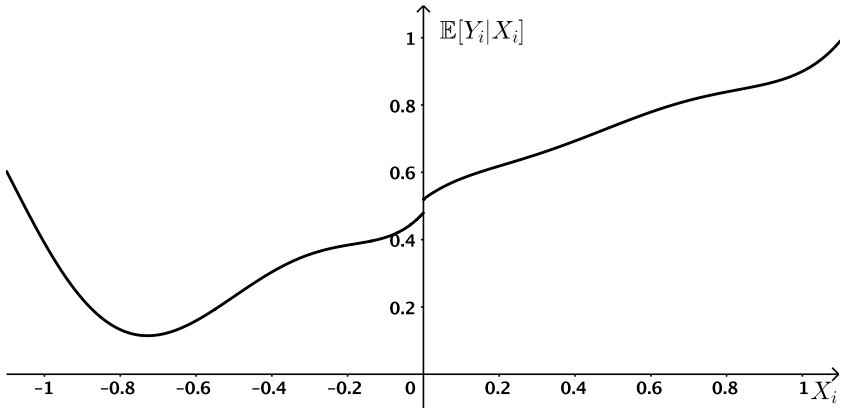
Our DGP is based on the incumbency data of Lee (2008). Lee studies incumbency advantage in the US House of Representatives. Districts where a party's candidate barely wins an election are, on average, comparable to districts where that party's candidate barely loses the election. The forcing variable X is the margin of victory of the Democratic party in percentage of votes. The target parameter is the effect of the Democrats winning the election at time t (incumbency) on the probability of the Democrats winning the election at time $t+1$. Lee's data have been used for simulation studies by several other econometricians, for example, by Imbens and Kalyanaraman (2012), Calonico, Cattaneo, and Titiunik (2014), and Armstrong and Kolesár (2018). We use the Monte Carlo DGP of Imbens and Kalyanaraman

(2012) and Calonico, Cattaneo, and Titiunik (2014), described in Equation (5.1).[8]

$$Y = \begin{cases} 0.48 + 1.27X + 7.18X^2 + 20.21X^3 \\ \qquad\qquad +21.54X^4 + 7.33X^5 + U \quad \text{if } X \in (-0.99, 0) \\ \\ 0.52 + 0.84X - 3X^2 + 7.99X^3 \\ \qquad\qquad -9.01X^4 + 3.56X^5 + U \quad \text{if } X \in [0, 0.99] \end{cases} \tag{5.1}$$

where X is distributed as Beta$(2, 4)$, U is zero-mean Gaussian with standard deviation 0.1295, and X is independent of U. Figure 2 depicts the conditional mean function of Equation (5.1).

Figure 2: Conditional Mean Function Based on Lee's (2008) Data



Notes: conditional mean function of Equation (5.1). The forcing variable X is the margin of victory of the Democratic party in percentage of votes in time t. The outcome variable Y is equal to one if Democrats win in time $t + 1$, but equal to zero otherwise.

Our simulation study uses variations of Equation (5.1) that are governed by two parameters: $\tau \in \mathbb{R}$ and $M \in \mathbb{R}_+$.

$$Y = \begin{cases} 0.48 + \tau\Lambda\left(4MX/\tau\right) + 1.27X + 7.18X^2 + 20.21X^3 \\ \qquad\qquad +21.54X^4 + 7.33X^5 + U \quad \text{if } X \in (-0.99, 0) \\ \\ 0.48 + \tau\Lambda\left(4MX/\tau\right) + 0.84X - 3X^2 + 7.99X^3 \\ \qquad\qquad -9.01X^4 + 3.56X^5 + U \quad \text{if } X \in [0, 0.99) \end{cases} \tag{5.2}$$

where $\Lambda\left(\cdot\right)$ is the logistic CDF function.

The conditional mean function of both Equations (5.1) and (5.2) are differentiable on

[8]The DGP in Equation (5.1) belongs to the class of functions that Armstrong and Kolesár (2018) study in their application to RDD. The set of functions $\mathcal{F}_{RDP,p}(C)$ on their page 658 contains Equation (5.1) with $p = 2$ and constant $C = 7.2$.

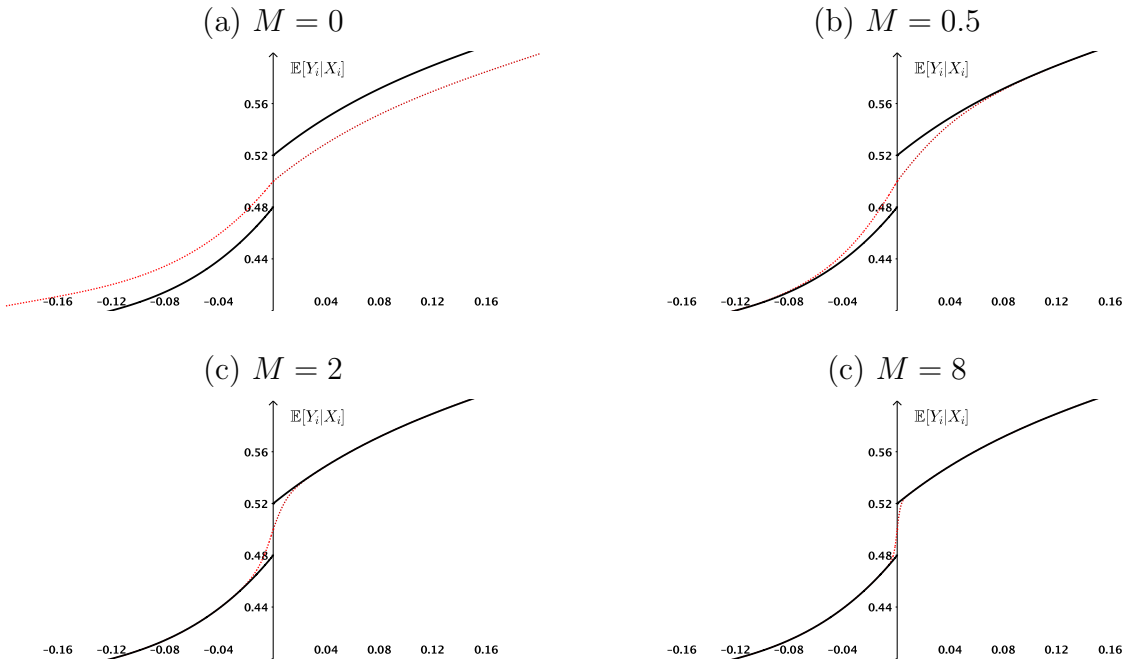
either side of the cutoff. The first is discontinuous at $X = 0$ with discontinuity of size 0.04, while the second is continuous at $X = 0$. For $\tau = 0.04$, Equation (5.2) approximates Equation (5.1) as $M \to \infty$. The parameter M is the derivative of $\tau\Lambda\left(4MX/\tau\right)$ wrt X at $X = 0$. As the slope M grows large, the continuous conditional mean function of Equation (5.2) approximates a discontinuous function with discontinuity of size τ. Figure 3 illustrates this approximation, as well as the proof of Corollary 3 in Section 4.1. For example, a model similar to Equation (5.2) with high values of M arises when districts manipulate the share of votes in order to win the election. Manipulation of the forcing variable has been extensively studied in the RDD literature. See, for example, McCrary (2008) and Gerard, Rokkanen, and Rothe (2016). Suppose the average causal effect of winning the election conditional on X is small for districts with small margin of victory, but large otherwise. In the absence of manipulation, $\mathbb{E}[Y|X]$ is continuous and very smooth to the right of the cutoff. The party in districts with low X has incentives to manipulate the election, and the researcher observes the manipulated margin of victory $\tilde{X}$, instead of X. Assume the probability that manipulation occurs conditional on $\tilde{X}$ increases continuously but sharply to the right of the cutoff. In this case, the researcher observes a conditional mean function that is continuous at the cutoff but that increases sharply to the right of the cutoff. In practice, one may falsely reject the null of zero effect simply because of manipulation, and not because of an actual causal effect. We provide a concrete example for this DGP in Section A.6 in the appendix.

The parameter of interest is m, the size of the jump discontinuity at $X = 0$. The null hypothesis is $m = 0$, which is the set of models in Equation (5.2) with $\tau \in \mathbb{R}$ and $M \in \mathbb{R}_+$. The alternative hypothesis is $m \neq 0$, which is the set of models with $\tau \neq 0$ and $M = \infty$. Section 4.1 shows that any model in the alternative is well-approximated in the LP metric by models under the null. The power of a.s. continuous tests is less than or equal to size.

The Monte Carlo experiment simulates 10,000 draws of an iid sample with 500 observations. The range of (τ, M) values for Model 5.2 in the experiment is consistent with the magnitudes of Lee's DGP. The maximum slope magnitude of the conditional mean graph in Figure 2 is 1.97, and we set $M \in \{0, 2, \ldots, 10\}$. The value of m for Lee's DGP is 0.04, and we vary τ in $\{0, 0.01, 0.04, 0.08\}$. We conduct a size and a power analysis. In the size analysis, we simulate rejection probabilities of the Wald test under each (τ, M)-model. The estimates of m and standard errors are obtained by the robust bias-corrected method of Calonico, Cattaneo, and Titiunik (2014) and implemented using the STATA package `rdrobust`. For each model (τ, M), the critical value of the test comes from the simulated distribution of the statistic under model $(\tau, 0)$. This ensures exact size of the test in the smoothest model under the null ($M = 0$).

The nominal size of the Wald tests in Table 1 is 5%, and the simulated rejection proba-

Figure 3: Approximating a Discontinuous Conditional Mean Function ($\tau = 0.04$)



Notes: the discontinuous conditional mean function $\mathbb{E}[Y|X]$ (solid line) is approximated by a sequence of continuous conditional mean functions (dotted lines). The solid line is the $\mathbb{E}[Y|X]$ of Model 5.1, and the dotted line is the $\mathbb{E}[Y|X]$ of Model 5.2 for $\tau = 0.04$ and $M \in \{0, 0.5, 2, 8\}$. The figure illustrates that model 5.2 approximates the DGP based on Lee (2008) as the slope at $X = 0$ grows large.

Table 1: Rejection Probability Under the Null - Size 5%

τ	$M = 0$	$M = 2$	$M = 4$	$M = 6$	$M = 8$	$M = 10$
.01	0.0500	0.0540	0.0557	0.0592	0.0585	0.0580
.02	0.0500	0.0665	0.0678	0.0649	0.0694	0.0685
.03	0.0500	0.0910	0.0942	0.0938	0.0941	0.1016
.04	0.0500	0.1005	0.1067	0.1071	0.1121	0.1139
.05	0.0500	0.1114	0.1264	0.1334	0.1464	0.1434
.06	0.0500	0.1292	0.1632	0.1680	0.1819	0.1819
.07	0.0500	0.1258	0.1617	0.1832	0.1906	0.2026
.08	0.0500	0.1320	0.1888	0.2142	0.2266	0.2427

Notes: the table displays the simulated rejection probability of the Wald test under various choices of (τ, M) for Model 5.2. Critical values of the test vary by row, but are constant across columns. For each (τ, M)-model, the critical value of the test comes from the simulated distribution of the statistic under model $(\tau, 0)$. The estimates of m and standard errors for the Wald test are obtained by the robust bias-corrected method of Calonico, Cattaneo, and Titiunik (2014) and implemented using the STATA package 'rdrobust'.

bility increases with τ and M. For the maximum slope of $M = 2$ observed from the model in Equation (5.1), the size of the test varies between 5.4% and 13.2%, depending on the choice of the model under the null. The true value of M is unknown, and a more conservative upper bound on the slope $M = 10$ distorts the size of the test up to 24%.

In the power analysis, we study rejection probabilities for models with $M = \infty$ and $\tau \in \{0, 0.01, \ldots, 0.08\}$. These models fall under the alternative because $m = \tau$ when $M = \infty$. For each (τ, ∞)-model, we would like the test to have correct size under the least favorable null model. Table 1 suggests that the least favorable model under the null is the one with the highest slope M. Figure 3 shows that null models can approximate any alternative (τ, ∞)-model arbitrarily well. If we restrict the slope at $X = 0$ to be at most M, the worst-case model under the null for the alternative (τ, ∞)-model is the (τ, M)-model. To evaluate the rejection probability under a (τ, ∞)-model, the critical value of the test comes from the simulated distribution of the statistic under a (τ, M)-model for various choices of (τ, M). That way, the test has correct size when $m = 0$ under all possibilities of least favorable (τ, M)-models.

Table 2: Rejection Probability Under the Alternative - Size 5%

τ	$M = 0$	$M = 2$	$M = 4$	$M = 6$	$M = 8$	$M = 10$
.01	0.0610	0.0508	0.0504	0.0501	0.0504	0.0500
.02	0.0763	0.0527	0.0524	0.0505	0.0513	0.0501
.03	0.1020	0.0574	0.0532	0.0525	0.0526	0.0527
.04	0.1204	0.0646	0.0571	0.0556	0.0536	0.0524
.05	0.1583	0.0770	0.0618	0.0597	0.0573	0.0544
.06	0.2013	0.0899	0.0682	0.0638	0.0605	0.0569
.07	0.2192	0.1023	0.0732	0.0677	0.0642	0.0590
.08	0.2781	0.1179	0.0839	0.0707	0.0654	0.0631

Notes: the entries of the table display the simulated rejection probability of the Wald test under Model 5.2 with various τ and $M = \infty$, so that the size of the discontinuity is $m = \tau$. Critical values of the test vary by row and column. For each (τ, M)-entry, the critical value comes from the simulated distribution of the statistic under a null (τ, M)-model. The estimates of m and standard errors for the Wald test are obtained by the robust bias-corrected method of Calonico, Cattaneo, and Titiunik (2014) and implemented using the STATA package 'rdrobust'.

The power of the tests in Table 2 increases with the size of discontinuity τ, but it decreases with the slope M of the least favorable model under the null. Intuitively, the higher M is, the harder it becomes to distinguish a (τ, M)-model from a (τ, ∞)-model. For the empirically relevant values of $\tau = 0.04$ and $M = 2$, we see that the power of the test is 6.5%, barely above its size. More conservative upper bounds on the slope of the model under the null essentially make power equal size. Section A.9 in the appendix contains versions of these tables for nominal levels 1% and 10%, as well as the simulated critical values used.

6 Conclusion

When drawing inference on a parameter in econometric models, some authors provide conditions under which tests have trivial power (impossibility type A). Others examine when confidence regions have error probability equal to one (impossibility type B). The motivation behind these negative results is that the parameter of interest may be nearly unidentified across models. Impossible inference relies on models being indistinguishable wrt some notion of distance. Some authors distinguish models using the Total Variation (TV) metric and others rely on the Lévy-Prokhorov (LP) metric, which is a weaker notion of distance. The ability to distinguish models in the TV metric is a necessary and sufficient condition for the existence of tests with non-trivial power. Impossible inference in terms of a weaker notion of distance is often easier to prove, it is applicable to the widely-used class of almost surely continuous tests, and it is useful for robust hypothesis testing.

Impossibility type A is stronger than type B. Dufour (1997) focuses on models in which tests based on bounded confidence regions fail to control size, but they could still have non-trivial power. Take the simultaneous equations model when instrumental variables may be arbitrarily weak. Moreira (2002, 2003) and Kleibergen (2005) propose tests that have correct size in models with type B impossibility. Furthermore, these tests have good power when identification is strong, being efficient under the usual asymptotics. Their power is not trivial, exactly because not every model under the alternative is approximated by models under the null.

The choice of the LP versus the TV metric connects our work to the work of Peter J. Huber on robust statistics. It leads us to look at the closure of model departures under the LP metric. In particular, robust hypothesis testing requires a non-zero TV distance between the closure of the null and alternative sets under the LP metric. For example, it is impossible to find a robust test that powerfully distinguishes covariance-stationary models from error-duration and Compound Poisson models, because the closure of the former contains the latter. This closure is quite rich, and we wonder what sort of hypotheses are testable. It is impossible to test the population mean, so one possibility may be quantiles such as value at risk (VaR). Peskir (2000) and Shorack and Wellner (2009) provide sufficient conditions for convergence of empirical processes under dependence. It would be interesting future work to build on these conditions to establish the class of processes in which quantile testing is possible.

References

Angrist, J., and V. Lavy (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114(2), 533–575.

Armstrong, T. B., and M. Kolesár (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86(2), 655–683.

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 27(4), 1115–1122.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66(1), 47–78.

Barbour, A., and S. Utev (1999): "Compound Poisson Approximation in Total Variation," *Stochastic Processes and Their Applications*, 82(1), 89–125.

Bertanha, M. (2017): "Regression Discontinuity Design with Many Thresholds," working paper, Department of Economics, University of Notre Dame.

Bertanha, M., and G. Imbens (2018): "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business and Economic Statistics, forthcoming*.

Bickel, P. J., and P. Bühlmann (1996): "What is a Linear Process?," *Proceedings of the National Academy of Sciences*, 93(22), 12128–12131.

Bickel, P. J., and D. A. Freedman (1981): "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, 9(6), 1196–1217.

Billingsley, P. (2008): *Probability and Measure*. John Wiley & Sons, New York.

Black, S. (1999): "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics*, 114(2), 577–599.

Caetano, C. (2015): "A Test of Exogeneity without Instrumental Variables in Models with Bunching," *Econometrica*, 83(4), 1581–1600.

Cai, T. T., and M. G. Low (2004): "An Adaptation Theory for Nonparametric Confidence Intervals," *The Annals of Statistics*, 32(5), 1805–1840.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

Canay, I. A., A. Santos, and A. M. Shaikh (2013): "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica*, 81(6), 2535–2559.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83(6), 2453–2483.

Coudin, E., and J.-M. Dufour (2009): "Finite-sample Distribution-free Inference in Linear Median Regressions under Heteroscedasticity and Non-linear Dependence of Unknown Form," *The Econometrics Journal*, 12, S19–S49.

Diebold, F. X., and A. Inoue (2001): "Long Memory and Regime Switching," *Journal of Econometrics*, 105(1), 131–159.

Dong, Y. (2016): "Jump or Kink? Regression Probability Jump and Kink Design for Treatment Effect Evaluation," working paper, University of California, Irvine.

Dong, Y., and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics*, 97(5), 1081–1092.

Donoho, D. L. (1988): "One-sided Inference About Functionals of a Density," *Annals of Statistics*, 16(4), 1390–1420.

Dudley, R. M. (1976): "Probabilities and Metrics," Lecture Notes Series No. 45, Matematisk Institut, Aarhus Universitet.

Dufour, J.-M. (1997): "Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65(6), 1365–1387.

Feir, D., T. Lemieux, and V. Marmer (2016): "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business and Economic Statistics*, 34(2), 185–196.

Gerard, F., M. Rokkanen, and C. Rothe (2016): "Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable," NBER Working Paper 22892.

Gibbs, A. L., and F. E. Su (2002): "On Choosing and Bounding Probability Metrics," *International Statistical Review*, 70(3), 419–435.

Gleser, L. J., and J. T. Hwang (1987): "The Nonexistence of 100(1-α)% Confidence Sets of Finite Expected Diameter in Errors-in-Variables and Related Models," *Annals of Statistics*, 15(4), 1351–1362.

Goncalves, F., and S. Mello (2018): "A Few Bad Apples?: Racial Bias in Policing," working paper, Crime Lab New York.

Granger, C. W., and N. Hyung (1999): "Occasional Structural Breaks and Long Memory," discussion paper 99-14, University of California, San Diego.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

Hamilton, J. D. (1989): "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57(2), 357–384.

Huber, P. J. (1964): "Robust Estimation of a Location parameter," *Annals of Mathematical Statistics*, 35(1), 73–101.

——— (1965): "A Robust Version of the Probability Ratio Test," *Annals of Mathematical Statistics*, 36(6), 1753–1758.

Huber, P. J., and E. M. Ronchetti (2009): *Robust Statistics*. John Wiley & Sons, Inc. Hoboken, NJ.

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79(3), 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142(2), 615–635.

Ingster, Y., and I. Suslina (2003): *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, vol. 169. Springer Science & Business Media, New York.

Jacob, B. A., and L. Lefgren (2004): "Remedial Education and Student Achievement: a Regression-Discontinuity Analysis," *Review of Economics and Statistics*, 86(1), 226–244.

Kamat, V. (2018): "On Nonparametric Inference in the Regression Discontinuity Design," *Econometric Theory*, 34(3), 694–703.

Kleibergen, F. (2005): "Testing Parameters in GMM Without Assuming That They Are Identified," *Econometrica*, 73(4), 1103–1123.

Kleven, H. J., and M. Waseem (2013): "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," *Quarterly Journal of Economics*, 128(2), 669–723.

Kraft, C. (1955): "Some Conditions for Consistency and Uniform Consistency of Statistical Procedures," in *University of California Publications in Statistics*, ed. by J. Neyman, L. LeCam, and H. Scheffé, vol. 2(6), pp. 125–142. University of California Press, Berkeley and Los Angeles.

Lee, D. S. (2008): "Randomized Experiments from Non-Random Selection in US House Elections," *Journal of Econometrics*, 142(2), 675–697.

Lehmann, E., and H. D'Abrera (2006): *Nonparametrics*. Springer-Verlag New York.

Lehmann, E., and J. Romano (2005): *Testing Statistical Hypotheses*. Springer-Verlag New York.

Low, M. G. (1997): "On Nonparametric Confidence Intervals," *Annals of Statistics*, 25(6), 2547–2554.

McCrary, J. (2008): "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142(2), 698–714.

Moreira, M. J. (2002): "Tests with Correct Size in the Simultaneous Equations Model," Ph.D. thesis, University of California, Berkeley.

——— (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027–1048.

Nielsen, H. S., T. Sørensen, and C. Taber (2010): "Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform," *American Economic Journal: Economic Policy*, 2(2), 185–215.

Parke, W. R. (1999): "What is Fractional Integration?," *Review of Economics and Statistics*, 81(4), 632–638.

Perron, P. (1989): "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, 57(6), 1361–1401.

Peskir, G. (2000): "From Uniform Laws of Large Numbers to Uniform Ergodic Theorems," Lecture Notes Series No. 66, Dept. Math. Univ. Aarhus.

Porter, J. (2003): "Estimation in the Regression Discontinuity Model," unpublished manuscript, University of Wisconsin, Madison.

Romano, J. P. (2004): "On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems," *Scandinavian Journal of Statistics*, 31(4), 567–584.

Rudin, W. (1976): *Principles of Mathematical Analysis*. McGraw-Hill New York.

Saez, E. (2010): "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy*, 2(3), 180–212.

Schmieder, J. F., T. von Wachter, and S. Bender (2012): "The Effects of Extended Unemployment Insurance Over the Business Cycle: Evidence from Regression Discontinuity Estimates over 20 Years," *Quarterly Journal of Economics*, 127(2), 701–752.

Shorack, G. R., and J. A. Wellner (2009): *Empirical Processes with Applications to Statistics*, vol. 59. Society for Industrial and Applied Mathematics (SIAM) Philadelphia.

Simonsen, M., L. Skipper, and N. Skipper (2016): "Price Sensitivity of Demand for Prescription Drugs: Exploiting a Regression Kink Design," *Journal of Applied Econometrics*, 31(2), 320–337.

Tibshirani, R., and L. A. Wasserman (1988): "Sensitive Parameters," *Canadian Journal of Statistics*, 6(2), 185–192.

Van der Vaart, A. W. (2000): *Asymptotic Statistics*. Cambridge University Press, Cambridge UK.

A Appendix

A.1 Proof of Corollary 1

We introduce some notation before embarking on the proof.

The density of $P \in \mathbb{P}$ wrt a σ-finite measure μ is p. The set of densities of all distributions in $\mathbb{P}$ is denoted $\mathbf{p}$. Similarly, the null and alternative sets of densities are $\mathbf{p}_0$ and $\mathbf{p}_1$, and their union equals $\mathbf{p}$. Define $co(\mathbf{p}')$ to be the convex hull of an arbitrary subset $\mathbf{p}' \subseteq \mathbf{p}$ in a similar fashion as in Equation (2.1).

The Total Variation (TV) metric between two distributions $P, Q \in \mathbb{P}$ with densities $p, q \in \mathbf{p}$ is defined as

$$d_{TV}(p,q) = \frac{1}{2}\int |p-q| \ d\mu. \tag{A.1}$$

The proof of the equivalence of (a) and (b) is shown in three parts.

Part 1: $(a) \Leftrightarrow (a')$ where

$$(a) : \forall q \in \mathbf{p}_1 \ \exists \{p_k\}_k \subseteq co(\mathbf{p}_0) \text{ such that } d_{TV}(p_k, q) \to 0$$
$$(a') : \forall q \in \mathbf{p}_1 \ \exists \{p_k\}_k \subseteq co(\mathbf{p}_0) \text{ and } \{\varepsilon_k\}_k \downarrow 0 \text{ such that } d_{TV}(p_k, q) < \varepsilon_k \ \forall k$$

Part 1, proof, $(a) \Rightarrow (a')$:

Fix q. For $\varepsilon_k = d_{TV}(p_k, q) \to 0$, there exists a monotone sub-sequence $\varepsilon_{k_j} = d_{TV}(p_{k_j}, q) \downarrow 0$. Create new sequences $\widetilde{p}_j = p_{k_j}$ and $\widetilde{\varepsilon}_j = \varepsilon_{k_j}/2$ so that $d_{TV}(\widetilde{p}_j, q) < \widetilde{\varepsilon}_j$.

Part 1, proof, $(a) \Leftarrow (a')$: straightforward.

Part 2: $(a') \Leftrightarrow (b')$ where

$$(a') : \forall q \in \mathbf{p}_1 \ \exists \{p_k\}_k \subseteq co(\mathbf{p}_0) \text{ and } \{\varepsilon_k\}_k \downarrow 0 \text{ such that } d_{TV}(p_k, q) < \varepsilon_k \ \forall k$$
$$(b') : \forall q \in \mathbf{p}_1 \ \exists \{\varepsilon_k\}_k \downarrow 0 \text{ such that } \forall \phi \ \int \phi q \ d\mu < \varepsilon_k + \sup_{p \in \mathbf{p}_0} \int \phi p \ d\mu \ \forall k$$

Part 2, proof, $(a') \Rightarrow (b')$:

Fix q, (a') implies there exists sequences $\{p_k\}_k \subseteq co(\mathbf{p}_0)$ and $\{\varepsilon_k\}_k \downarrow 0$ such that $d_{TV}(p_k, q) < \varepsilon_k \ \forall k$. Fix k. Use Theorem 1 with $\{\mathbf{p}_1\} = \{q\}$. (a') implies

$\forall \phi \ \int \phi q \ d\mu < \varepsilon_k + \sup_{p \in \mathbf{p}_0} \int \phi p \ d\mu$.

This is true for every k of a sequence ε_k that converges to zero, given an arbitrary q.

Part 2, proof, $(a') \Leftarrow (b')$:

Fix q, get ε_k. Fix k. Use Theorem 1 with $\{\mathbf{p}_1\} = \{q\}$. (b') implies there exists $p_k \in co(\mathbf{p}_0)$ such that $d_{TV}(p_k, q) < \varepsilon_k$. Repeat this for every k to get a sequence $\{p_k\}_k \subseteq co(\mathbf{p}_0)$ such that $d_{TV}(p_k, q) < \varepsilon_k \ \forall k$.

Part 3: $(b') \Leftrightarrow (b)$ where

$$(b') : \forall q \in \mathbf{p}_1 \ \exists \{\varepsilon_k\}_k \downarrow 0 \text{ such that } \forall \phi \ \int \phi q \ d\mu < \varepsilon_k + \sup_{p \in \mathbf{p}_0} \int \phi p \ d\mu \ \forall k$$

$$(b) : \forall \phi \text{ and } q \in \mathbf{p}_1, \int \phi q \; d\mu \leq \sup_{p \in \mathbf{p}_0} \int \phi p \; d\mu$$

Part 3, proof $(b') \Rightarrow (b)$:
Fix q, get ε_k. Fix ϕ. It is true that
$\int \phi q \; d\mu < \varepsilon_k + \sup_{p \in \mathbf{p}_0} \int \phi p \; d\mu$.
Take limits on both sides,
$\int \phi q \; d\mu \leq \sup_{p \in \mathbf{p}_0} \int \phi p \; d\mu$.
This is true for every q and every ϕ.
Part 3, proof $(b') \Leftarrow (b)$:
Straightforward because for arbitrary ϕ, q, and $\{\varepsilon_k\}_k \downarrow 0$
$\int \phi q \; d\mu \leq \sup_{p \in \mathbf{p}_0} \int \phi p \; d\mu$
implies that
$\int \phi q \; d\mu < \varepsilon_k + \sup_{p \in \mathbf{p}_0} \int \phi p \; d\mu$.
□

A.2 Proof of Theorem 2

The proof of Theorem 2 follows the same lines as the proof of Theorem 1 by Romano (2004) except for the fact that our Assumption 1 is stated in terms of the LP metric and in terms of the convex hull of $\mathbb{P}_0$.

Pick an arbitrary $Q \in \mathbb{P}_1$. There exists a sequence of distributions $\{P_k\}_{k=1}^{\infty} \subseteq co(\mathbb{P}_0)$ such that $P_k \xrightarrow{d} Q$. Convergence in distribution is equivalent to $\mathbb{E}_{P_k}[g] \to \mathbb{E}_Q[g]$ for every bounded real-valued function g whose set of discontinuity points has probability zero under Q (Theorem 25.8, Billingsley (2008)). In particular, this is true for $g = \phi$ for an arbitrary ϕ that is a.s. continuous under Q.

Take an arbitrary sequence $\varepsilon_n \to 0$, and pick a sub-sequence $\{P_{k_n}\}_n$ from the sequence $\{P_k\}_k$ such that

$$-\varepsilon_n \leq \mathbb{E}_Q \phi - \mathbb{E}_{P_{k_n}} \phi \leq \varepsilon_n. \tag{A.2}$$

Therefore,

$$\mathbb{E}_Q \phi \leq \mathbb{E}_{P_{k_n}} \phi + \varepsilon_n \leq \sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi + \varepsilon_n. \tag{A.3}$$

Given $\varepsilon_n \to 0$, it follows that, for $\forall Q \in \mathbb{P}$,

$$\mathbb{E}_Q \phi \leq \sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi. \tag{A.4}$$

Consequently,

$$\sup_{Q \in \mathbb{P}_1} \mathbb{E}_Q \phi \leq \sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi. \tag{A.5}$$

It is clear that $\sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi \geq \sup_{P \in \mathbb{P}_0} \mathbb{E}_P \phi$. It remains to show that these are equal. Assume $\sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi > \sup_{P \in \mathbb{P}_0} \mathbb{E}_P \phi$. Select $\varepsilon > 0$ small enough such that $\sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi - \varepsilon > \sup_{P \in \mathbb{P}_0} \mathbb{E}_P \phi$. There exists $P_\varepsilon \in co(\mathbb{P}_0)$ such that

$$\sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi \geq \mathbb{E}_{P_\varepsilon} \phi > \sup_{P \in co(\mathbb{P}_0)} \mathbb{E}_P \phi - \varepsilon > \sup_{P \in \mathbb{P}_0} \mathbb{E}_P \phi. \tag{A.6}$$

By definition, $P_\varepsilon = \sum_{i=1}^N \alpha_i P_i$ for $N \in \mathbb{N}$, $P_i \in \mathbb{P}_0$ $\forall i$, $\alpha_i \in [0,1]$ $\forall i$, and $\sum_{i=1}^N \alpha_i = 1$. Then, $\mathbb{E}_{P_\varepsilon}\phi = \sum_{i=1}^N \alpha_i \mathbb{E}_{P_i}\phi \leq \sup_{P\in\mathbb{P}_0} \mathbb{E}_P\phi$, a contradiction. Therefore, $\sup_{P\in co(\mathbb{P}_0)} \mathbb{E}_P\phi = \sup_{P\in\mathbb{P}_0} \mathbb{E}_P\phi$, and

$$\sup_{Q\in\mathbb{P}_1} \mathbb{E}_Q\phi \leq \sup_{P\in\mathbb{P}_0} \mathbb{E}_P\phi. \tag{A.7}$$

□

A.3 Proof of Theorem 3

The proof is a combination of proofs by Dufour (1997) and Gleser and Hwang (1987).

Part (2.6):

Fix $m \in \mu(\mathbb{P})$. Define $\phi_m = \mathbb{I}\{m \notin C(Z)\}$, and note that $\sup_{P\in\mathbb{P}(m)} \mathbb{E}_P\phi_m = \sup_{P\in co(\mathbb{P}(m))} \mathbb{E}_P\phi_m$ (see proof of Theorem 2). It follows that $1-\alpha \leq \inf_{P\in\mathbb{P}(m)} P[m \in C(Z)] = \inf_{P\in co(\mathbb{P}(m))} P[m \in C(Z)]$. Therefore, $\forall P \in co(\mathbb{P}(m))$, $P[\mu(P) \in C(Z)] \geq 1-\alpha$.

By Assumption 2, there exists $\{P_k\}$ in $co(\mathbb{P}(m))$ such that $P_k \xrightarrow{d} P^*$. Then,

$$1-\alpha \leq P_k[\mu(P_k) \in C(Z)] = P_k[m \in C(Z)] \to P^*[m \in C(Z)] \tag{A.8}$$

where the convergence follows by Portmanteau's theorem because $P^*(\partial\{m \in C(Z)\}) = 0$ (Theorem 29.1 of Billingsley (2008)). This proves (2.6).

Part (2.7):

Pick a sequence $m_n \in \mu(\mathbb{P})$ such that m_n is unbounded. Without loss of generality, assume $m_n \uparrow \infty$. We have that

$$1-\alpha \leq P^*[m_n \in C(Z)] \leq P^*[m_n \leq U[C(Z)]]. \tag{A.9}$$

Taking the limit as $n \to \infty$,

$$1-\alpha \leq P^*[U[C(Z)] = \infty] \tag{A.10}$$
$$\leq P^*[U[C(Z)] - L[C(Z)] = \infty] = P^*[D[C(Z)] = \infty]. \tag{A.11}$$

Part (2.8):

Assumption 2 gives a sequence $\{P_k\}_k$ in $co(\mathbb{P})$ that converges in distribution to P^*. By assumption, $P^*[\partial\{D[C(Z)] = \infty]\}] = 0$, so Portmanteau's theorem gives $P_k[D[C(Z)] = \infty]] \to P^*[D[C(Z)] = \infty]] \geq 1-\alpha$. There exists a sequence $\delta_k \downarrow 0$ such that $P_k[D[C(Z)] = \infty]] \geq 1-\alpha-\delta_k$.

Fix $\varepsilon > 0$. The set $B_\varepsilon(P^*) \cap co(\mathbb{P})$ contains infinitely many P_ks from the sequence above. For these P_ks,

$$1-\alpha-\delta_k \leq P_k[D[C(Z)] = \infty]] \tag{A.12}$$
$$\leq \sup_{P\in B_\varepsilon(P^*)\cap co(\mathbb{P})} P[D[C(Z)] = \infty]] \tag{A.13}$$

$$= \sup_{P \in B_\varepsilon(P^*) \cap \mathbb{P}} P\left[D[C(Z)] = \infty\right]] \tag{A.14}$$

where the last equality follows by the same argument seen in the proof of (2.6) above. Taking the limit as $k \to \infty$ gives (2.8).

□

A.4 Proof of Lemma 1

Lemma 1. *Let $C(Z)$ be constructed as in Equation (2.9). Then,*

$$\inf_{P \in \mathbb{P}} P\left[\mu(P) \in C(Z)\right] = 1 - \sup_{m \in \mu(\mathbb{P})} \alpha(m). \tag{A.15}$$

Proof of Lemma 1. Suppose

$$\sup_{m \in \mu(\mathbb{P})} \sup_{P \in \mathbb{P}_{0,m}} P(\phi_m(Z) = 1) = \alpha. \tag{A.16}$$

Now, pick $\varepsilon > 0$. Then, there exists m_ε such that

$$\alpha - \varepsilon/2 \leq \sup_{P \in \mathbb{P}_{0,m_\varepsilon}} P(\phi_{m_\varepsilon}(Z) = 1) \leq \alpha. \tag{A.17}$$

There also exists $P_\varepsilon \in \mathbb{P}_{0,m_\varepsilon}$ such that

$$\alpha - \varepsilon \leq P_\varepsilon(\phi_{m_\varepsilon}(Z) = 1) \leq \alpha. \tag{A.18}$$

Rearranging the expression above, we obtain

$$1 - \alpha + \varepsilon \geq P_\varepsilon(\mu(P_\varepsilon) \in C(Z)) \geq 1 - \alpha. \tag{A.19}$$

Therefore, we find that

$$\inf_{P \in \mathbb{P}} P\left[\mu(P) \in C(Z)\right] = 1 - \alpha, \tag{A.20}$$

as we wanted to prove.

□

A.5 Proof of Corollary 3

Fix $m \in \mathbb{R}$. Pick an arbitrary $Q \in \mathbb{P}_{1,m}$, and let $m' = \mu(Q) \neq m$. Define $g(x) = \mathbb{E}_Q[Y_i | X_i = x]$. Construct a sequence of functions $g_k : \mathbb{R} \to \mathbb{R}$, $k = 1, 2, \ldots$ as follows:

$$g_k(x) = g(x) + (m' - m)\left[\Lambda\left(k^2(x - c)\right) - \mathbb{I}\{x \geq c\}\right] \tag{A.21}$$

where $\Lambda(\cdot)$ is the cumulative distribution function (CDF) of the logistic distribution.

The function g_k is infinitely continuously differentiable on $\mathbb{X} \setminus \{c\}$, so $g_k \in \mathcal{G}$ $\forall k$, and $\lim_{x \downarrow c} g_k(x) - \lim_{x \uparrow c} g_k(x) = m$. Moreover, as $k \to \infty$, $g_k(x) \to g(x)$ for every $x \neq c$. Define

P_k to be the distribution of $(X_i\ ,\ Y_i - g(X_i) + g_k(X_i))$ when $(X_i, Y_i) \sim Q$. It follows that $\mu(P_k) = m$ and $P_k \in \mathbb{P}_{0,m}\ \ \forall k$.

It remains to show that $P_k \xrightarrow{d} Q$, or equivalently, to show that

$$(X_i\ ,\ Y_i - g(X_i) + g_k(X_i)) \xrightarrow{d} (X_i, Y_i) \tag{A.22}$$

as $k \to \infty$ where $(X_i, Y_i) \sim Q$. Note that $(X_i\ ,\ Y_i - g(X_i) + g_k(X_i)) = (X_i, Y_i) + (0, g_k(X_i) - g(X_i))$, so it suffices to show that $g_k(X_i) - g(X_i) \xrightarrow{p} 0$ as $k \to \infty$.

Define $A_k = \{c - k^{-1} < X_i < c + k^{-1}\}$, and let A_k^c be the complement of A_k. Fix $\varepsilon > 0$.

$$Q\left[|g_k(X_i) - g(X_i)| > \varepsilon\right] \tag{A.23}$$

$$= Q\left[|g_k(X_i) - g(X_i)| > \varepsilon \mid\ A_k\right]\ Q\left[A_k\right] \tag{A.24}$$

$$+ Q\left[|g_k(X_i) - g(X_i)| > \varepsilon \mid\ A_k^c\right]\ Q\left[A_k^c\right]. \tag{A.25}$$

Part (A.24) vanishes as $k \to \infty$ by the continuity property of probability measures, because $A_k \downarrow \{c\}$ and $Q\left[\{c\}\right] = 0$ by assumption.

For part (A.25), note that $|g_k(x) - g(x)| \leq |m' - m|\Lambda\left(-k\right)$ for any $x \in A_k^c$ because $\Lambda\left(k^2(x - c)\right)$ is strictly increasing in x and symmetric around $x = c$, so $|g_k(x) - g(x)|$ attains its maximum at $x = c - k^{-1}$ and $x = c + k^{-1}$. Therefore,

$$(A.25) \leq \mathbb{I}\left\{|m' - m|\Lambda\left(-k\right) > \varepsilon\right\}\ Q\left[A_k^c\right] \to 0 \tag{A.26}$$

because $\Lambda\left(-k\right) \to 0$ as $k \to \infty$.

Therefore, Assumption 1 is satisfied for every $m \in \mathbb{R}$. Theorem 2 applies, and Corollary 2 applies with $\mu(\mathbb{P}) = \mathbb{R}$.

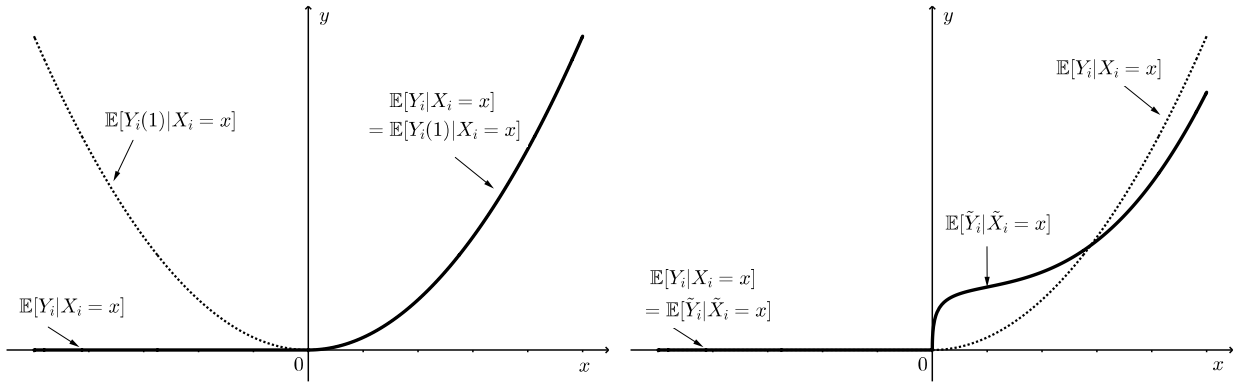
□

A.6 Example of RDD Model with Manipulation

In this section, we provide an example of DGP with manipulation that gives rise to a model similar to Equation (5.2) in our Monte Carlo experiment. The potential outcome of losing an election is normalized to zero ($Y_i(0) = 0$), and the potential outcome of winning an election is $Y_i(1)$. Suppose the expected potential gain of winning an election is small for tight elections but large otherwise; that is, let $\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x] = x^2$, where X_i is the margin of victory of a given political party in district i. Assume the distribution of X_i is Uniform$[-1, 1]$. Each district is an iid draw $(Y_i(1), X_i, \varepsilon_i)$ from a given distribution, where ε_i denotes district i's potential to influence the election outcome in a world where manipulation is possible. In a world *without* manipulation, the researcher observes (Y_i, D_i, X_i), where $D_i = \mathbb{I}\{X_i \geq 0\}$ is the victory indicator, and $Y_i = D_iY_i(1) + (1 - D_i)Y_i(0) = D_iY_i(1)$ is the outcome. It follows that $\mathbb{E}[Y_i|X_i = x] = x^2\mathbb{I}\{x \geq 0\}$. There is no discontinuity at the cutoff, the causal effect is zero, and the DGP is under the null hypothesis of zero effect at the cutoff (Figure 4(a)).

Figure 4: Example of RDD with Manipulation

(a) Conditional Mean without Manipulation (b) Conditional Mean with Manipulation



Notes: In figure (a) there is no manipulation, and the researcher observes a sample of (Y_i, X_i). The solid line denotes the conditional mean of the observed outcome given the margin of victory $\mathbb{E}[Y_i|X_i]$. The conditional mean of the potential outcome in case of victory $\mathbb{E}[Y_i(1)|X_i]$ is the the dotted line, and the potential outcome in case of loss is normalized to zero $Y_i(0) = 0$. In figure (b) there is manipulation, and the researcher observes a sample of $(\tilde{Y}_i, \tilde{X}_i)$. The solid line is $\mathbb{E}[\tilde{Y}_i|\tilde{X}_i]$ while the dotted line depicts $\mathbb{E}[Y_i|X_i]$. Manipulation increases the slope of the conditional mean function at the cutoff.

In a world *with* manipulation, the given party in district i decides to influence the election if the expected potential gain of doing so is positive. In other words, if the margin of victory without manipulation X_i leads to the loss of the election, and the expected potential gain of winning the election $\mathbb{E}[Y_i(1) - Y_i(0)|X_i]$ is strictly positive, then the party decides to manipulate. Thus, manipulation occurs if $X_i < 0$, and the margin of victory changes from X_i to $\varepsilon_i > 0$. Although the party manipulates as little as possible to win the election, it does not have perfect control over its vote share. Assume $\varepsilon_i = 5\chi^2_{(3)}$, that is, five times a Chi-square distribution with three degrees of freedom. The pdf f_ε evaluated at zero equals zero, but it is highly sloped to the right of zero. Let $\tilde{X}_i$ be the manipulated margin of victory defined as $\tilde{X}_i = \mathbb{I}\{X_i < 0\}\varepsilon_i + \mathbb{I}\{X_i \geq 0\}X_i$. The researcher observes $(\tilde{Y}_i, \tilde{D}_i, \tilde{X}_i)$, where $\tilde{Y}_i = \tilde{D}_i Y_i(1) + (1 - \tilde{D}_i)Y_i(0) = \tilde{D}_i Y_i(1)$, and $\tilde{D}_i = \mathbb{I}\{\tilde{X}_i \geq 0\}$ is the victory indicator. The conditional mean function under manipulation is given by

$$\begin{aligned}
&\mathbb{E}[\tilde{Y}_i|\tilde{X}_i = x] = \mathbb{I}\{x \geq 0\}\mathbb{E}[Y_i(1)|\tilde{X}_i = x] \\
&= \mathbb{I}\{x \geq 0\}\mathbb{E}\Big[Y_i(1) \;\Big|\; \{\varepsilon_i = x, X_i < 0\} \text{ or } \{X_i = x, X_i \geq 0\}\Big] \\
&= \mathbb{I}\{x \geq 0\}\Big\{\theta(x)\mathbb{E}\left[Y_i(1) \mid \varepsilon_i = x, X_i < 0\right] + (1 - \theta(x))\mathbb{E}\left[Y_i(1) \mid X_i = x, X_i \geq 0\right]\Big\} \\
&= \mathbb{I}\{x \geq 0\}\Big\{\theta(x)\mathbb{E}\left[Y_i(1) \mid X_i < 0\right] + (1 - \theta(x))\mathbb{E}\left[Y_i(1) \mid X_i = x\right]\Big\} \\
&= \mathbb{I}\{x \geq 0\}\Big\{\theta(x)(1/3) + (1 - \theta(x))x^2\Big\},
\end{aligned}$$

where the weight

$$\theta(x) = \frac{f_\varepsilon(x)\mathbb{P}(X_i < 0)}{f_\varepsilon(x)\mathbb{P}(X_i < 0) + f_X(x)\mathbb{I}\{x \geq 0\}}$$

$$= \frac{f_\varepsilon(x)0.5}{f_\varepsilon(x)0.5 + 0.5\mathbb{I}\{x \geq 0\}}$$

is such that $\theta(x) \in [0,1]$, $\theta(0) = 0$, $\theta(x)$ is continuous in x, and it is positively and highly sloped near $x = 0$. The conditional mean function $\mathbb{E}[\tilde{Y}_i|\tilde{X}_i = x]$ increases sharply at the cutoff (Figure 4(b)) because districts with low X_i and high causal effects manipulate their X_i to $\tilde{X}_i = \varepsilon_i$ to the right of the cutoff. There is no discontinuity at the cutoff, and the DGP is still under the null hypothesis of zero effect. However, manipulation makes it harder to distinguish a zero effect from a positive effect at the cutoff.

A.7 Proof of Corollary 5

Fix $Q \in \mathbb{P}_1$ with CDF $F_Q(x)$. The CDF $F_Q(x)$ has a jump discontinuity of size $\delta > 0$ at $x = x_0$. Call f_Q the derivative of F_Q at $x \neq x_0$, which is a continuous function of x for every $x \neq x_0$. The integral of f_Q over $\mathbb{R}$ equals $1 - \delta$. The side limits of f_Q at x_0, $f_Q(x_0^+)$ and $f_Q(x_0^-)$, may be different from each other. Pick a sequence $\varepsilon_k \downarrow 0$. Construct a continuous "hat-shaped" function $g_k(x) : [x_0 - \varepsilon_k; x_0 + \varepsilon_k] \to \mathbb{R}$ such that: (i) $g_k(x_0 - \varepsilon_k) = f_Q(x_0 - \varepsilon_k)$; (ii) $g_k(x_0 + \varepsilon_k) = f_Q(x_0 + \varepsilon_k)$; (iii) $g_k(x)$ has constant and positive slope for $x \leq x_0$, and constant and negative slope for $x \geq x_0$; (iv) $g_k(x) \geq f_Q(x)$; and (v) $\int (g_k(x) - f_Q(x))\, dx = \delta$. It is always possible to construct such a function for a small enough ε_k. Define $f_{P_k}(x) = f_Q(x) + \mathbb{I}\{x_0 - \varepsilon_k \leq x \leq x_0 + \varepsilon_k\}\,(g_k(x) - f_Q(x))$. This is a continuous PDF function, and let it define the distribution P_k. Then the CDF F_{P_k} converges to F_Q as $k \to \infty$ at every continuity point of F_Q, so that $P_k \xrightarrow{d} Q$.

□

A.8 Proof of Corollary 6

Fix $m \in \mathbb{R}$. Choose an arbitrary $Q \in \mathbb{P}_{1,m}$, and let $m' = \mu(Q) \neq m$. Define $g(x,w) = \mathbb{E}_Q[Y_i|X_i = x, W_i = w]$, and $\tau_Q(w) = g(x,w) - \lim_{x \downarrow 0} g(x,w)$.

Construct a sequence of functions $g_k : \mathbb{X} \times \mathbb{W} \to \mathbb{R}$, $k = 1, 2, \ldots$ as follows:

$$g_k(x,w) = g(x,w) + (\tau_Q(w) - m)\left[\mathbb{I}\{x > 0\} - \Lambda\left(k^2 x\right)\right] \tag{A.27}$$

where $\Lambda(\cdot)$ is the CDF of the logistic distribution.

The function g_k is infinitely many times continuously differentiable wrt x on $\{\mathbb{X}\backslash\{c\}\} \times \mathbb{W}$, so $g_k \in \mathcal{G}$ $\forall k$. Also, $g_k(0,w) - \lim_{x \downarrow 0} g_k(x,w) = m$. Moreover, as $k \to \infty$, $g_k(x,w) \to g(x,w)$ pointwise. Define P_k to be the distribution of $(X_i\ ,\ W_i\ ,\ Y_i - g(X_i,W_i) + g_k(X_i,W_i))$ when $(X_i, W_i, Y_i) \sim Q$. It follows that $\mu(P_k) = m$ and $P_k \in \mathbb{P}_{0,m}$ $\forall k$.

It remains to show that $P_k \xrightarrow{d} Q$, or equivalently, to show that

$$(X_i\ ,\ W_i\ ,\ Y_i - g(X_i,W_i) + g_k(X_i,W_i)) \xrightarrow{d} (X_i, W_i, Y_i) \tag{A.28}$$

as $k \to \infty$ where $(X_i, Y_i) \sim Q$. Note that $(X_i\ ,\ W_i\ ,\ Y_i - g(X_i,W_i) + g_k(X_i,W_i)) = (X_i, W_i, Y_i) + (0, g_k(X_i,W_i) - g(X_i,W_i))$, so it suffices to show that $g_k(X_i,W_i) - g(X_i,W_i) \xrightarrow{p} 0$ as $k \to \infty$.

Define $A_k = \{0 < X_i < k^{-1}\}$, and let A_k^c be the complement of A_k. Fix $\varepsilon > 0$.

$$Q\left[|g_k(X_i, W_i) - g(X_i, W_i)| > \varepsilon\right] \tag{A.29}$$

$$= Q\left[|g_k(X_i, W_i) - g(X_i, W_i)| > \varepsilon \mid A_k\right] Q\left[A_k\right] \tag{A.30}$$

$$+ Q\left[|g_k(X_i, W_i) - g(X_i, W_i)| > \varepsilon \mid A_k^c\right] Q\left[A_k^c\right]. \tag{A.31}$$

Part (A.30) vanishes as $k \to \infty$ by the continuity property of probability measures because $A_k \downarrow \{\emptyset\}$ where $\emptyset$ denotes the empty set and has zero probability.

For part (A.31), $|g_k(x, w) - g(x, w)| \leq |\tau_Q(w) - m||1 - \Lambda(k)|$ for any w and any $x \in A_k^c$ because $1 - \Lambda(k^2 x)$ is strictly decreasing in x. For fixed w, $|g_k(x, w) - g(x, w)|$ attains its maximum at $x = k^{-1}$. Therefore,

$$(A.31) \leq \mathbb{P}\left\{|\tau_Q(W_i) - m||1 - \Lambda(k)| > \varepsilon\right\} Q\left[A_k^c\right] \to 0 \tag{A.32}$$

because $\Lambda(k) \to 1$ as $k \to \infty$ and $|\tau_Q(W_i) - m|$ is bounded.

□

A.9 Simulations - RDD

This section contains additional results of the RDD simulation in the main text. The size and power analyses in the main text use the 5% nominal level. This section has the same analyses using 1% and 10% nominal levels. It also has the simulated critical values under the various choices of null (τ, M)-models.

Table 3: Simulated Rejection Rates and Critical Values

Panel 1: *Rejection Rate under the Null Model* (τ, M) *Using Critical Values Simulated under Model* $(\tau, 0)$

(a) Nominal Size 1%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	0.0100	0.0097	0.0108	0.0142	0.0108	0.0123
.02	0.0100	0.0130	0.0147	0.0120	0.0139	0.0132
.03	0.0100	0.0195	0.0192	0.0212	0.0210	0.0224
.04	0.0100	0.0231	0.0257	0.0232	0.0243	0.0276
.05	0.0100	0.0274	0.0323	0.0335	0.0373	0.0345
.06	0.0100	0.0339	0.0467	0.0523	0.0591	0.0577
.07	0.0100	0.0402	0.0546	0.0612	0.0702	0.0737
.08	0.0100	0.0451	0.0674	0.0827	0.0917	0.0968

(b) Nominal Size 10%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	0.1000	0.1096	0.1111	0.1170	0.1159	0.1127
.02	0.1000	0.1253	0.1309	0.1270	0.1337	0.1254
.03	0.1000	0.1587	0.1618	0.1651	0.1629	0.1707
.04	0.1000	0.1733	0.1803	0.1818	0.1827	0.1901
.05	0.1000	0.1801	0.2049	0.2097	0.2244	0.2177
.06	0.1000	0.2126	0.2561	0.2635	0.2810	0.2805
.07	0.1000	0.2172	0.2596	0.2867	0.3022	0.3072
.08	0.1000	0.2244	0.2957	0.3238	0.3427	0.3594

Panel 2: *Rejection Rate under the Alternative Model* (τ, ∞) *Using Critical Values Simulated under Model* (τ, M)

(a) Nominal Size 1%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	0.0113	0.0104	0.0099	0.0100	0.0100	0.0101
.02	0.0159	0.0111	0.0105	0.0102	0.0101	0.0103
.03	0.0223	0.0111	0.0112	0.0109	0.0111	0.0105
.04	0.0286	0.0133	0.0114	0.0109	0.0114	0.0108
.05	0.0390	0.0188	0.0137	0.0125	0.0114	0.0111
.06	0.0668	0.0211	0.0149	0.0132	0.0124	0.0114
.07	0.0869	0.0237	0.0150	0.0139	0.0128	0.0132
.08	0.1165	0.0295	0.0184	0.0164	0.0132	0.0130

(b) Nominal Size 10%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	0.1150	0.1013	0.1004	0.1003	0.1001	0.1001
.02	0.1411	0.1058	0.1026	0.1025	0.1013	0.1013
.03	0.1695	0.1130	0.1060	0.1045	0.1031	0.1030
.04	0.2003	0.1239	0.1129	0.1078	0.1064	0.1038
.05	0.2389	0.1361	0.1182	0.1128	0.1116	0.1079
.06	0.3093	0.1543	0.1319	0.1203	0.1154	0.1109
.07	0.3335	0.1792	0.1438	0.1286	0.1192	0.1161
.08	0.4073	0.2085	0.1567	0.1342	0.1269	0.1200

Panel 3: *Critical Values Simulated under Null Model* (τ, M)

(a) Nominal Size 1%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	3.0222	3.0074	3.0631	3.1741	3.0641	3.0927
.02	3.0669	3.1787	3.2466	3.1459	3.1743	3.2177
.03	3.0506	3.3975	3.4213	3.4238	3.3659	3.4641
.04	3.0810	3.5389	3.5381	3.4846	3.4706	3.5916
.05	3.0786	3.5443	3.6329	3.6819	3.7531	3.7461
.06	2.9715	3.5573	3.7476	3.8374	3.9043	3.8987
.07	2.9829	3.7526	3.9103	3.9917	3.9896	4.0030
.08	2.9536	3.7593	4.0283	4.1033	4.2614	4.2798

(b) Nominal Size 10%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	1.8700	1.9312	1.9323	1.9577	1.9602	1.9437
.02	1.8801	2.0155	2.0295	2.0221	2.0558	2.0240
.03	1.8466	2.1393	2.1621	2.1594	2.1638	2.2066
.04	1.8822	2.2410	2.2767	2.2879	2.2917	2.3270
.05	1.9056	2.3180	2.4002	2.4403	2.4903	2.4722
.06	1.8316	2.3812	2.5359	2.5594	2.6196	2.6234
.07	1.8772	2.4403	2.5859	2.7054	2.7526	2.7885
.08	1.8651	2.4662	2.7184	2.8357	2.8925	2.9366

(c) Nominal Size 5%

τ	$M=0$	$M=2$	$M=4$	$M=6$	$M=8$	$M=10$
.01	2.2543	2.2888	2.3158	2.3511	2.3279	2.3525
.02	2.2491	2.4041	2.4038	2.3951	2.4190	2.4311
.03	2.1983	2.5607	2.5687	2.5776	2.5385	2.6150
.04	2.2350	2.6669	2.6713	2.6819	2.7055	2.7515
.05	2.2390	2.7010	2.8138	2.8575	2.9340	2.8821
.06	2.2030	2.7688	2.9399	2.9920	3.0553	3.0453
.07	2.2713	2.8591	3.0370	3.0936	3.1734	3.2180
.08	2.2556	2.9023	3.1370	3.2685	3.3442	3.3817

Notes: Model (τ, M) refers to Equation (5.2) in the main text. The estimates for the Wald test are obtained by the robust bias-corrected method of Calonico, Cattaneo, and Titiunik (2014) and implemented using the STATA package 'rdrobust'.