JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 0, NO. 0, 1–11, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1328356

A Permutation Test for the Regression Kink Design

Peter Ganong[a,b] and Simon Jäger[a,c,d,e,f]

[a]NBER, National Bureau of Economic Research, Cambridge, MA; [b]University of Chicago Harris School of Public Policy, Chicago, IL; [c]Massachusetts Institute of Technology, Department of Economics, Cambridge, MA; [d]Institute on Behavior and Inequality (briq) Bonn, Germany; [e]IZA Institute of Labor Economics, Bonn, Germany; [f]CESifo, Munich, Germany

ABSTRACT

The regression kink (RK) design is an increasingly popular empirical method for estimating causal effects of policies, such as the effect of unemployment benefits on unemployment duration. Using simulation studies based on data from existing RK designs, we empirically document that the statistical significance of RK estimators based on conventional standard errors can be spurious. In the simulations, false positives arise as a consequence of nonlinearities in the underlying relationship between the outcome and the assignment variable, confirming concerns about the misspecification bias of discontinuity estimators pointed out by Calonico, Cattaneo, and Titiunik. As a complement to standard RK inference, we propose that researchers construct a distribution of placebo estimates in regions with and without a policy kink and use this distribution to gauge statistical significance. Under the assumption that the location of the kink point is random, this permutation test has exact size in finite samples for testing a sharp null hypothesis of no effect of the policy on the outcome. We implement simulation studies based on existing RK applications that estimate the effect of unemployment benefits on unemployment duration and show that our permutation test as well as inference procedures proposed by Calonico, Cattaneo, and Titiunik improve upon the size of standard approaches, while having sufficient power to detect an effect of unemployment benefits on unemployment duration. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2014
Accepted March 2017

KEYWORDS
Permutation test; Policy discontinuities; Randomization inference

1. Introduction

We develop a permutation test for Regression Kink (RK) designs that rely on an identification principle analogous to the one underlying the better-known Regression Discontinuity designs. Regression Discontinuity designs estimate the change in the *level* of an outcome Y at the threshold level of the assignment variable V at which the *level* of the policy changes discontinuously (see, e.g., Thistlethwaite and Campbell 1960, Imbens and Lemieux 2008). RK designs exploit discontinuous changes in the *slope* of a policy B at a specific level of the assignment variable and assess whether there is also a discontinuous change in the *slope* of the outcome variable. By comparing the ratio of the slope change in the outcome variable with the slope change in the policy variable at the kink point, the RK design recovers a causal effect of the policy on the outcome at the kink point. This is again analogous to RD designs that calculate the ratio of the change in the level of the outcome to the change in the level of treatment at the discontinuity. The slope change at the kink point identifies the average effect of increasing the policy conditional on the level of the assignment variable at the kink point. Key identification and inference results for the RK design were derived in Nielsen, Sørensen, and Taber (2010), Card, Lee, Pei and Weber (2015a, CLPW in the following), Calonico, Cattaneo, and Titiunik (2014b, CCT in the following), and Calonico, Cattaneo, and Farrell (2016).

This article discusses the proposed permutation test for the RK design, its underlying assumptions and implementation. For motivation, we begin by describing the RK design based on an example from a growing body of literature that implements the RK design to estimate the causal effect of unemployment benefits B on an outcome Y such as unemployment duration or employment (see, e.g., Britto 2015, CLPW; Card et al. 2015a; Kyyrä and Pesola 2015; Landais 2015; Sovago 2015; Kolsrud, Landais, Nilsson, and Spinnewijn 2015, KLNS in the following). Despite being crucial for policy design, it is difficult to address the question of whether unemployed individuals stay out of work for longer if they receive more generous benefits in the absence of a randomized experiment. RK studies of unemployment insurance aim to fill this gap by exploiting the fact that many unemployment insurance systems pay out benefits B that rise linearly with prior income V, up to a benefit cap $\bar{B}$ for individuals earning above a reference income $\bar{V}$, the kink point. In such a schedule, the slope of the policy variable B with respect to the assignment variable V decreases discontinuously at $\bar{V}$. To study the effect of benefits B on unemployment duration, researchers estimate the extent to which the slope of the outcome variable Y—unemployment duration—with respect to the assignment variable changes discontinuously at such kink points. Intuitively, if unemployment benefits have no impact on unemployment duration, one would not expect to see

CONTACT Simon Jäger *sjaeger@mit.edu* Massachusetts Institute of Technology, Cambridge, MA 02139-4307.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

discontinuous changes in the slope of unemployment duration Y and prior income V at a kink point $\bar{V}$. However, if unemployment benefits do deter individuals from finding employment, one would expect discontinuous changes in the slope of unemployment duration with respect to prior income at kink points that depend on the strength of the causal effect of benefits B on unemployment duration Y. Section 2 provides a more detailed review of the RK design and key identification results.

While the RK design has become increasingly popular, RK estimators as typically implemented may suffer from non-negligible misspecification bias and consequently incorrectly centered confidence intervals (CCT). In most applications of the RK design, researchers use local linear or quadratic estimators to estimate the slope change at the kink and choose bandwidths with the goal of minimizing mean squared error (Fan and Gijbels 1996). Table A.1 in the Appendix provides an overview of 44 RK studies, the vast majority of which use a linear or quadratic specification. In these specifications, misspecification bias can arise as a consequence of nonlinearity in the conditional expectation function. To observe how, consider Figure 1, which displays data-with a piecewise linear data-generating process (DGP) featuring a kink and a quadratic DGP with no kink. The top panel of Figure 1 shows that the estimated conditional means from these two DGPs are visually indistinguishable. However, applying local linear estimators that are common in the RK literature to both DGPs indicates statistically highly significant slope changes at the kink, even though the quadratic DGP does not feature a discontinuous slope change. CCT show that such misspecification bias is non-negligible with standard bandwidth selectors and leads to poor empirical coverage of the resulting confidence intervals. As a remedy, CCT develop an alternative estimation and inference approach for RD and RK designs based on a bias-correction of the estimators and a new standard error estimator that reflects the bias correction.

To provide a complement and a robustness check to existing RK inference, we propose a permutation test that treats the location of the kink point $\bar{V}$ as random and has exact size in finite samples. The key assumption underlying our test is that the location of the policy kink point is randomly drawn from a set of potential kink locations. This assumption is rooted in a thought experiment in which the data are taken as given and only the location of the kink point $\bar{V}$ is thought of as a random variable with a known distribution. In many RK contexts, this assumption is appealing because the kink point's location is typically not chosen based on features of the DGP and in some cases—such as kinks in many unemployment insurance schedules—is determined as the outcome of a stochastic process (see Section 3). Under the null hypothesis that the policy has no effect on the outcome and the assumption that the location of the policy kink is randomly drawn from a specified support, the distribution of placebo estimates provides an exact null distribution for the test statistic at the policy kink. We prove that the permutation test controls size exactly in finite samples.

We implement simulation studies to compare the size and power of our permutation test to that of standard RK inference as well as the RK estimators and confidence intervals developed in CCT. Specifically, we simulate data based on estimated DGPs from existing RK applications aimed at estimating the effect of unemployment benefits on unemployment duration

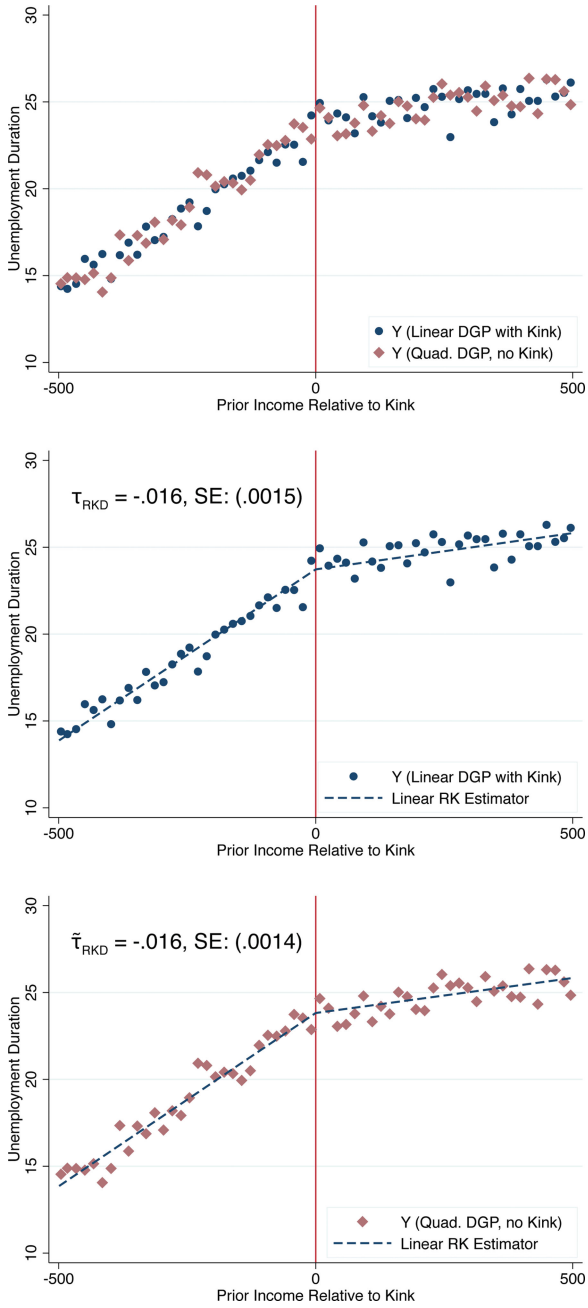

Figure 1. Piecewise linear and quadratic simulated DGPs. The data-generating process (DGP) is either linear with a kink (blue dots) or quadratic (red diamonds) without a kink. We generate 1000 observations with a variance of 12 and plot the data in bins based on the approach in Calonico, Cattaneo, and Titiunik (2014a, 2015). We estimate a linear regression kink model with heteroscedasticity-robust standard errors. The top panel shows the relationship between the outcome variable and the running variable for both the piecewise linear and the quadratic DGP. In the middle panel, we display the data for the piecewise linear DGP and add the estimates from a local linear model. In the bottom panel, we display the data for the quadratic (no kink) DGP and estimates from a local linear model.

in Austria and Sweden (Card et al. 2015b, Kolsrud et al. 2015). These Monte Carlo simulations document that asymptotic inference and standard bandwidth choice procedures (Fan and Gijbels 1996) lead to over-rejection of the null hypothesis; RK estimates are statistically significant using asymptotic methods even when the kink is fact zero. Further simulation studies show that asymptotic inference and standard bandwidth choice procedures are particularly problematic—as evidenced by high Type I error rates—when the relationship between outcome

and assignment variable is nonlinear. By contrast, CCT and the permutation test maintain actual size close to nominal size.

Following our analysis of size, we subsequently analyze power. The simulation studies further reveal that both the permutation test and the estimator based on CCT have sufficient power to reject the null hypothesis based on the effect sizes in KLNS, that is, they reject the hypothesis that unemployment benefits do not affect unemployment duration. In the simulation studies, we also document examples of settings in which the permutation test fails to detect nonzero kinks. This can occur when the relationship between the outcome and the assignment variable is sufficiently nonlinear relative to the magnitude of the kink. Overall, the simulation studies document that there is a spectrum of estimators' performance: standard asymptotic inference has much larger than nominal size, meaning that it substantially over-rejects the null hypothesis, CCT has closer to nominal size and lower power and the permutation test yields exact nominal size but the lowest power compared to its alternatives.

Our permutation test also serves as a complement to existing inference methods for Regression Discontinuity (RD) designs. Applying our procedure to simulated data based on two existing RD applications shows that the permutation test has similar size and power to existing RD methods. In particular, Cattaneo, Frandsen, and Titiunik (2015) developed a randomization inference approach for RD designs based on an interpretation of RD designs as local randomized experiments in a narrow window around the RD cutoff. While the procedure developed by Cattaneo, Frandsen, and Titiunik (2015) treats the locations of observations above and below the cutoff as random within a narrow window and also proposes a method for the selection of the window in which the randomization assumption holds, our test offers a complementary approach by treating the location of the kink point or cutoff itself as random.

By drawing on randomization inference, the permutation test that we propose builds on a long tradition in the statistics literature (Fisher 1935; Lehmann and Stein 1949; Welch and Gutierrez 1988; Welch 1990; Rosenbaum 2001; Ho and Imai 2006, see Rosenbaum 2002, for an introduction), which has recently seen renewed interest among econometricians (see, for instance, Bertrand, Duflo, and Mullainathan 2004; Imbens and Rosenbaum 2005; Chetty, Looney, and Kroft 2009; Abadie, Diamond, and Hainmueller 2010; Abadie, Athey, Imbens, and Wooldridge 2014; Cattaneo, Frandsen, and Titiunik 2015). Our approach generalizes a suggestion by Imbens and Lemieux (2008) for the RD design—"testing for a zero effect in settings where it is known that the effect should be 0"—to the RK design by estimating slope changes in regions where there is no change in the slope of the policy. Engström et al. (2015) and Gelman and Imbens (2014) pursued related placebo analyses. Our article builds on CCT in developing new methods to deal with misspecification bias in RK and RD designs. In related work, Landais (2015) proposed an alternative way of gauging the robustness in RK estimates by constructing difference-in-differences RK estimates based on the same kink point, albeit using data from time periods with and without the presence of an actual policy kink. Ando (forthcoming) uses Monte Carlo simulations to argue that linear RK estimates are biased in the presence of plausible amounts of curvature.

2. Notation and Review

2.1. Identification in the Regression Kink Design

The following section reviews key results and notation for the RK design based on CLPW and CCT and illustrates the method based on RK designs aimed at estimating the causal effect of unemployment benefits B on an outcome Y unemployment duration, by exploiting kinks in the unemployment insurance schedule (see Britto 2015, Card et al. 2015a, CLPW, KLNS, Kyyrä and Pesola 2015, Landais 2015, and Sovago 2015).

Formally, the outcome Y is modeled as

$$Y = y(B, V, U) \tag{1}$$

where B denotes the policy variable, V denotes a running variable, here thought of as prior income, which determines the assignment of B, and U denotes an error term. Analogous to treatment effects for binary treatments, defined as $y(1, V, U) - y(0, V, U)$ in a potential outcomes framework, the treatment parameter that RK designs intend to estimate is the marginal effect of increasing the level of the policy B on the outcome Y, that is,

$$\frac{dy(B, V, U)}{dB}. \tag{2}$$

Integrating this marginal effect over the distribution of U conditional on $B = b$ and $V = v$ leads to the "treatment on the treated" parameter in Florens, Heckman, Meghir, and Vytlacil (2008):

$$TT_{B=b,V=v} = \int \frac{\partial y(b, v, u)}{\partial b} dF_{U|B=b,V=v}(u) \tag{3}$$

where $F_{U|B=b,V=v}$ denotes the conditional c.d.f. of the error term U. In the context of unemployment benefits, this corresponds to the average effect of marginally increasing unemployment benefits on unemployment duration for individuals with unemployment benefits $B = b$ and prior income $V = v$.

The key feature that RK designs exploit is a discrete slope change in the assignment mechanism of the policy. Let $B = b(V)$ denote the continuous policy function. In many unemployment systems, benefits B rise linearly with the prior income V that an individual earned before becoming unemployed. A maximum level of benefits is typically offered for individuals earning above a higher reference income $\bar{V}$. A minimum floor on benefits is also typically offered for individuals with reference income below some minimum threshold. The existence of a benefit maximum implies that the slope between the policy variable B (unemployment benefits) and the assignment variable V (prior income) changes discontinuously at $\bar{V}$ when the prior income rises above the reference income. To illustrate an example of a kink at which the slope of the benefit schedule *increases* discretely around a minimum benefit floor, the top panel of Figure 2 shows plots of unemployment benefits plotted against earnings in the previous year based on Austrian UI data (CLPW). The benefit schedule or policy

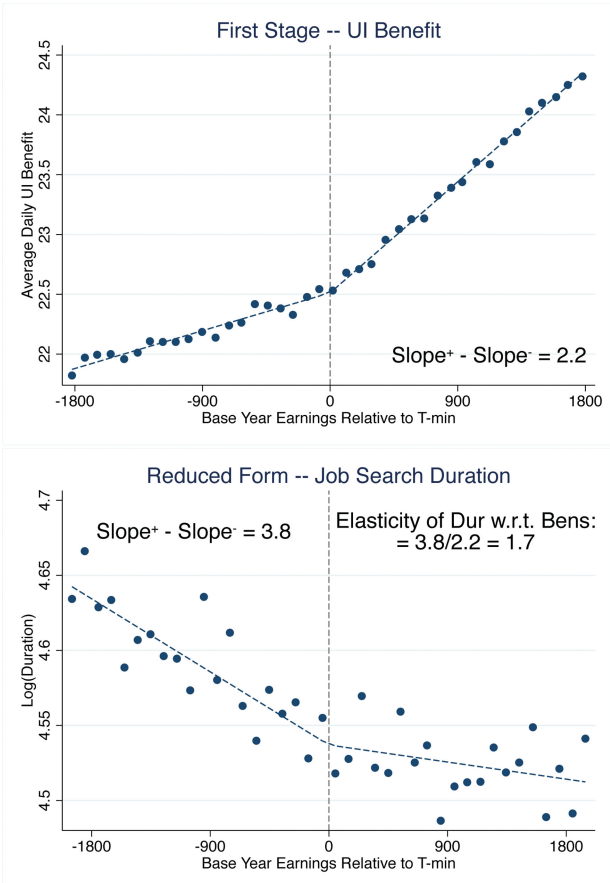

Figure 2. RK Example: UI Benefits in Austria. Notes: the figures are based on Card et al. (2015b). T-min refers to the earnings threshold at which benefits start to rise. Bin number chosen based on the approach by Calonico, Cattaneo, and Titiunik (2014a, 2015).

function $B = b(V)$ can be simply described as follows:

$$\frac{\partial b(V)}{\partial v} = \begin{cases} \alpha_1, & v < \bar{V} \\ \alpha_2, & \bar{V} < v \end{cases}, \tag{4}$$

where $\alpha_1 \neq \alpha_2$ and $\lim_{v \uparrow \bar{V}} db(v)/dv = \alpha_1$ and $\lim_{v \downarrow \bar{V}} db(v)/dv = \alpha_2$. In this example, the researchers analyze a location $\bar{V}$ at which the slope of unemployment benefits B with respect to prior income V rises discontinuously, that is, $\alpha_1 < \alpha_2$.

Researchers can exploit the discrete change in the policy function $b(V)$ at the kink point to identify the marginal effect of the policy. Intuitively, if unemployment benefits have no impact on unemployment duration, one would not expect to see discontinuous changes in the slope of unemployment duration and prior income at kink points. However, if unemployment benefits do deter individuals from going back to work, then one would expect discontinuous changes in the slope of unemployment duration with respect to prior income at kink points. At the kink, one would expect a positive slope change as individuals above the reference income receive more generous benefits and consequently stay unemployed for longer. To study the effect of benefits B on unemployment duration, researchers can then estimate the extent to which the slope of the outcome variable Y—unemployment duration—with respect to the assignment variable changes discontinuously at such kink points. To illustrate, the bottom panel in Figure 2 plots a measure of unemployment duration Y against the running variable V, earnings in the previous year, again based on Austrian UI data (CLPW) and documents an apparent slope change at the kink point.

2.2. Estimation and Identification

The RK estimand, τ_{RK}, is defined in the population as the change in the slope of the outcome variable at the kink point normalized by the slope change in the policy at the kink point:

$$\tau_{\text{RK}} \equiv \frac{\lim_{v \downarrow \bar{V}} dE(Y|V=v)/dv - \lim_{v \uparrow \bar{V}} dE(Y|V=v)/dv}{\lim_{v \downarrow \bar{V}} db(v)/dv - \lim_{v \uparrow \bar{V}} db(v)/dv}. \tag{5}$$

In the example of unemployment benefits, the denominator of this expression—that is, the slope change in the policy variable at the reference income—corresponds to $\lim_{v \downarrow \bar{V}} db(v)/dv - \lim_{v \uparrow \bar{V}} db(v)/dv = \alpha_2 - \alpha_1$. This is analogous to the denominator in fuzzy RD designs, which scales up the difference in the level of the outcome variable at the discontinuity by the difference in the level of the treatment at the discontinuity.

CLPW prove that the RK estimand in (5) identifies the "treatment on the treated" parameter in (3) (Florens et al. 2008) for individuals at the kink point under mild regularity conditions, in particular an assumption of smoothness of y, so that

$$\tau_{\text{RK}} = \int \frac{\partial y(b, v, u)}{\partial b} dF_{U|B=b,V=\underline{V}}(u). \tag{6}$$

Local polynomial regression techniques are used for estimation of τ_{RK} (Fan and Gijbels 1996). The data are split into two subsamples to the left and right of the kink point (denoted by + and −, respectively) and a local polynomial regression is estimated separately for each subsample. For the sharp RK design, in which the slope change in the policy at the kink point—normalized to $\bar{V} = 0$ in the following—is known, this amounts to solving the following least squares problem in the sample:

$$\begin{aligned} &\min_{\{\beta_j^-\}} \sum_{i=1}^{N^-} \left\{ Y_i^- - \sum_{j=0}^{p} \tilde{\beta}_j^- \left(V_i^-\right)^j \right\}^2 K\left(\tfrac{V_i^-}{h}\right) \\ &\min_{\{\beta_j^+\}} \sum_{i=1}^{N^+} \left\{ Y_i^+ - \sum_{j=0}^{p} \tilde{\beta}_j^+ \left(V_i^+\right)^j \right\}^2 K\left(\tfrac{V_i^+}{h}\right) \\ &\text{subject to } \tilde{\beta}_0^- = \tilde{\beta}_0^+ \\ &\hat{\tau}_{RK}^p \equiv \left(\tilde{\beta}_1^+ - \tilde{\beta}_1^-\right)/(\alpha_2 - \alpha_1). \end{aligned} \tag{7}$$

Here, p denotes the order of the polynomial, K the kernel function, and h the bandwidth used for estimation. In the literature, the bandwidth is typically chosen based on the formula in Fan and Gijbels or the procedure in CCT. The numerator of the left-hand side of equation (7) is identified as $\hat{\beta}_1^+ - \hat{\beta}_1^-$. The papers in the RK literature have primarily adopted a uniform kernel as the choice of K and overwhelmingly use local linear and quadratic specifications.

2.3. Asymptotic Bias

A potential problem of RK designs is that nonlinearities of the conditional expectation function $E[Y|V = v]$ can generate bias in the estimator $\hat{\tau}_{\text{RK}}^p$. Panel 3 of Figure 1 illustrates the intuition for this result as curvature of the conditional expectation function generates bias of linear RK estimators. A formal argument

supporting this intuition follows from CCT who derive a general formula for the asymptotic bias of RK and RD estimators. Based on the general formula, the asymptotic misspecification bias of local linear RK estimators is proportional to $(m_+^{(2)} + m_-^{(2)})h$, where h is the bandwidth and the terms $m_+^{(j)}$ and $m_-^{(j)}$ denote the limits of the jth derivative of $m(v) \equiv E[Y|V = v]$ from above and below at the kink. A similar expression can be derived for local quadratic estimators for which first-order bias is proportional to third-order terms of the conditional mean function. CCT prove that such misspecification bias is non-negligible with standard bandwidth selectors and thus leads to poor empirical coverage of the resulting confidence intervals.

3. A Permutation Test for the Regression Kink Design

3.1. The Thought Experiment

We propose a simple permutation test to assess the null hypothesis that treatment has no effect on the outcome of interest. At the core of our test is the assumption that the location of the policy kink can be considered as randomly drawn from a known set of placebo kink points. This assumption needs to be evaluated in the context of the specific research design under scrutiny. We describe a method for how researchers can estimate a distribution of placebo kink points in the context of unemployment insurance systems. In this interval, we can reassign the location of the kink and calculate RK estimates, $\hat{\tau}_{RK}^p$, at these placebo kinks. The permutation test assesses the extremeness of the estimated change in the slope at the kink point relative to estimated slope changes at nonkink points under the null hypothesis that the policy does not affect the outcome.

The thought experiment underlying this permutation test and randomization inference more generally is different from the one underlying asymptotic inference. The idea underlying asymptotic inference is one of sampling observations from a large population. In contrast, the thought experiment in randomization inference is based on a fixed population that the researcher observes in the data, with the realizations of the running variable v and the outcome variable y, in which the assignment of treatment is sampled repeatedly. In the latter approach, treatment assignment is thought of as the random variable. Our test therefore does not treat the sample as being drawn from a (super) population for which we seek inference but rather takes the observed sample as given and tests hypotheses regarding this particular sample, treating the location of the policy kink as a random variable.

3.2. The Permutation Test Statistic

This subsection introduces the reduced form RK estimator as the test statistic for the permutation test, which can be easily adjusted to correspond to researchers' modeling choices in a given RK application. We let $\mathbf{y}$ denote the vector of y_i values, $\mathbf{v}$ denote the vector of v_i realizations and k denote a potential kink point, with a policy kink featuring a discontinuous slope change in the policy *or* a placebo kink not featuring such a discontinuous slope change. The data are a vector of n observations each with $(y_i, v_i, b(v_i))$ denoting outcome, running variable, and policy variable: in the context of using the RK design to estimate the effect of unemployment benefits on unemployment duration, these would correspond to unemployment duration, prior income, and unemployment benefits, respectively.

For notational tractability and expositional clarity, our exposition pertains to the linear RK model with a uniform kernel. This can be easily generalized to higher-order polynomials and other kernels. Define the matrix

$$\mathbf{v}^k \equiv \tilde{\mathbf{v}}(k) \equiv \begin{pmatrix} 1 & (v_1 - k) & (v_1 - k)\mathbf{1}(v_1 \geq k) \\ \vdots & \vdots & \vdots \\ 1 & (v_n - k) & (v_n - k)\mathbf{1}(v_n \geq k) \end{pmatrix}. \tag{8}$$

We define the test statistic for the slope change at the potential kink point k as

$$T(\mathbf{v}, \mathbf{y}, k) \equiv (0 \quad 0 \quad 1)' (\mathbf{v}^{k\prime}\mathbf{v}^k)^{-1}\mathbf{v}^{k\prime}\mathbf{y}, \quad |v_i - k| \leq h(\mathbf{v}, \mathbf{y}, k), \tag{9}$$

where $h(\mathbf{v}, \mathbf{y}, k)$ denotes the bandwidth used for estimation. This test statistic corresponds to the reduced form of a linear RK estimator. At the true kink point, which we label k^*, this estimator—scaled up by the slope change at the policy—identifies the causal effect of the policy on the "treated," $\int \frac{\partial y(b,v,u)}{\partial b} dF_{U|B=b(k^*),V=k^*}(u)$, under the assumptions laid out in CLPW. We can calculate the test statistic $T(\mathbf{v}, \mathbf{y}, k)$ at the true policy kink point k^*, $T(\mathbf{v}, \mathbf{y}, k^*)$, and at other points $k \in [v_{\min}, v_{\max}]$ in the range of v.

Modeling choices. The test statistic used for the permutation test should correspond to the RK estimator and preferred modeling choices, including bandwidth (or a bandwidth selection mechanism), polynomial order, and bias correction using the approach developed by CCT, implemented by the researcher for the RK estimator at the actual policy kink. The permutation test approach can be easily generalized to incorporate alternative RK estimators, polynomial orders, and bandwidth choices. It also can be easily applied to a Regression Discontinuity application (see Section 5). The Monte Carlo studies that we present in Section 4 provide some guidance for the modeling choices and suggest that estimators based on the procedure in CCT perform relatively well compared to local polynomial estimators with bandwidth choice based on Fan and Gijbels (1996).

3.3. The Randomization Assumption

The core assumption underlying our permutation test is that the location of the policy kink point k^* can be thought of as being randomly drawn:

Assumption: Random Kink Placement. k^* is a realization of a random variable K distributed according to a known distribution P.

The assumption that the policy kink location can be thought of as being randomly drawn is a strong but natural one in the context of many RK designs. It would be violated if, for instance, policy-makers had chosen a kink location explicitly or implicitly in response to the shape of the conditional expectation function $E[Y|V]$, for example, at a location where curvature is particularly high or low. For example, the KLNS study an example where the benefit cap was raised beginning in

2001. If policy-makers raised the cap specifically because they observed curvature in unemployment duration near the policy kink, this would contaminate this analysis. However, the institutional setup in KLNS makes this type of violation unlikely. We discuss four implementable strategies for researchers to identify P.

1. Estimation of stochastic process based on institutional features. In the example of estimating the causal effect of unemployment benefits on unemployment duration, researchers implementing the permutation test for the RK design can exploit features of many unemployment insurance systems to directly estimate the distribution P. In many unemployment insurance systems, the location of the kink point at which benefits are capped is determined as a consequence of past aggregate wage growth in the economy. For instance, in Austria—the setting of the study by CLPW—the earnings ceiling in the unemployment insurance system changes as a function of aggregate wage growth from the third to the second previous calendar year (§ 108 *Allgemeines Sozialversicherungsgesetz*). Therefore, data on past wage growth can be used to directly estimate the properties of the stochastic process that determines the realization of k^* in a given year or researchers can directly use the distribution of past changes in the location of the kink point as a set of potential kink locations for implementing the permutation test. In Section 4, we describe how we draw from the distribution of realized kink location changes in the context of the Austrian unemployment insurance system.

If directly estimating the stochastic process determining k^* is infeasible, then there is a class of alternative selection mechanisms involving a discretized set of kinks on range $[\underline{v}, \bar{v}]$.

2. Documentary evidence on rule-making. Researchers can still proxy P by drawing on information on the institutional environment of the relevant RK application. In the spirit of randomization inference, P can correspond to a grid of points spanning the range of proposals for kinks that could have been adopted. For example, if several policy proposals existed in a political debate regarding the choice of a reference income $\bar{V}$ in the example of unemployment insurance we discussed, researchers could use the discretized range $[\underline{v}, \bar{v}]$ that includes all of these proposals. For instance, prior to switching to a system of automatic updates of the earnings ceiling based on aggregate wage growth in 1969, the German *Bundestag* adjusted the earnings ceiling in a discretionary fashion so that the minutes of plenary proceedings can be used to gauge the range of discussed proposals.

3. Local randomization neighborhood (Cattaneo, Frandsen, and Titiunik 2015). In the context of developing a randomization inference approach for the RD design treating observations as randomly assigned, Cattaneo, Frandsen, and Titiunik (2015) designed a data-driven procedure to select a window around an RD cutoff based on balance tests of pre-treatment covariates in which treatment status is arguably as good as randomly assigned. A natural extension of their procedure is to treat the location of the cutoff or kink as randomly assigned within this window.

4. Range of available data. As a final benchmark, we suggest that researchers consider the whole range of available data $[v_{\min}, v_{\max}]$ and treat the empirical distribution of V as the distribution for P. This follows the approach in Section 4 of Gelman and Imbens (2014) for selecting pseudo-thresholds in the context of evaluating RD designs. This approach is natural in the context of using the RK design to estimate the causal effect of unemployment benefits on unemployment duration, because there are a wide range of policy kink locations in practice. For example, the maximum weekly UI benefit is a direct function of prior income varies across US states from $235 to $698.

3.4. Exact Size For Testing the Null Hypothesis of Policy Irrelevance

The goal of our permutation test is to assess whether the data reject the null hypothesis that the policy does not affect outcomes. We formalize this as a sharp null hypothesis where $\mathcal{B}$ and $\mathcal{V}$ denote the range of the policy and assignment variable, respectively:

Null Hypothesis: Policy Irrelevance. The policy does not affect outcomes at any v: $\frac{dy(b,v,U)}{db} = 0, \quad \forall b \in \mathcal{B}, \forall v \in \mathcal{V}$.

Note that this hypothesis implies that the policy is irrelevant, that is, $y(b_1, \tilde{v}, U) = y(b_2, \tilde{v}, U),\ \forall b_1,\ b_2 \in \mathcal{B}, \forall \tilde{v} \in \mathcal{V}$. Under the *Policy Irrelevance Hypothesis* and the *Assumption of Random Kink Placement*, the distribution of kink estimates over P corresponds to the exact distribution of possible estimates that could have arisen had the policy kink been at a different location in the same dataset. Under these assumptions, we can construct an exact test following the logic of Fisher (1935) and Pitman (1937). Note that the null hypothesis of policy irrelevance across the distribution of potential kink points $\mathcal{V}$ is stronger than the null hypothesis in Cattaneo, Frandsen, and Titiunik (2015) and Cattaneo, Titiunik, and Vazquez-Bare (forthcoming), who assume policy irrelevance in a local region around the discontinuity.

Proposition 1. Under the Null Hypothesis of Policy Irrelevance and the Random Kink Location assumption, there exists a test function $\phi(\mathbf{v}, \mathbf{y}, k)$ for significance level α that has an exact finite sample level of α.

In Appendix B, we follow the structure of a simple proof by Romano (1990) documenting that under the Null Hypothesis of Policy Irrelevance and the Random Kink Location assumption there exists a test function $\phi(\mathbf{v}, \mathbf{y}, k)$ for significance level α that has an exact finite sample level of α.

Under the assumption of random kink placement, the null hypothesis thus leads to a testable implication that can be assessed by measuring how unusual a given realization of the test statistic is at the policy kink. Analogous to the test outlined above, researchers can also calculate p-values for assessing the likelihood that the null hypothesis is true given the RK estimate at the policy kink k^* and the distribution of placebo kink estimates. Suppose a researcher had calculated 1000 placebo kink estimates and the estimate at the policy kink k^* was the 20th lowest of these estimates. Subsequently, the two-sided p-value would be calculated to be 4% corresponding to twice the one-sided p-value of 2%. More generally, the two-sided p-value can be calculated as twice the minimum of the two one-sided p-values, that is, the minimum of the fraction of placebo estimates—including the one at the actual policy kink k^*—that are no greater than or no smaller than the test statistic at the policy kink k^*.

3.5. Confidence Intervals

We also construct confidence intervals by inverting the permutation test following Rosenbaum (2002, chap. 2.6.2) and Imbens and Rubin (2015, chap. 5.7). The confidence interval is defined as the region of potential constant treatments effects for which the permutation test does not reject the null hypothesis of policy irrelevance when applied to the transformed data. We transform the data by subtracting out a set of potential treatment effects and then apply the permutation test to the transformed data. Appendix C provides details on the algorithm that we implement for identifying confidence intervals.

4. Applications of the Permutation Test

4.1. The Effect of Unemployment Benefits on Duration: Simulation Studies Comparing Asymptotic and Randomization Inference

We simulate data based on two existing RK applications estimating the effect of unemployment insurance on unemployment duration, namely CLPW and KLNS. The simulation studies compare the size and power of asymptotic inference and the permutation test for linear, quadratic, and cubic estimators based on FG and CCT bandwidth choice. The simulations reveal a tradeoff between size and power among inference procedures: FG bandwidth choice has high power but size well above nominal levels, while CCT and the permutation test have much improved size but have the power to reject the null hypothesis only in the KLNS setting and not in the CLPW setting.

4.1.1. Simulation Procedure and Methods

Our data simulation procedure has two steps. In a first step, we estimate cubic spline models on binned means of data given to us by the authors of the respective papers. The running variable in both applications corresponds to measures of prior income and the outcome variable to a measure of unemployment duration. The top panel of Figure 3 shows data that CLPW provided to us, outlining the relationship between log unemployment duration and base year earnings at the bottom kink for 100 nonoverlapping bins. Similarly, the bottom panel of Figure 3 shows the relationship between unemployment duration and the running variable for the application in KLNS. We use a cubic spline model as approximation to the DGP because it offers a very flexible approximation of the conditional mean function with only one additional parameter per knot. We estimate two cubic spline models on each dataset: one with a cubic spline that assumes there is no discontinuous slope change to evaluate size and again with a modified cubic spline that allows for a discontinuous slope change at the policy kink to evaluate power. In all cases, we use a spline with 100 equally spaced knots covering the full support of the running variable. To illustrate, the solid maroon lines in both panels of Figure 3 show the estimated conditional mean function in both applications using a cubic spline without allowing for a slope change at the kink point, while the dashed maroon lines show a cubic spline fit allowing for discontinuous slope changes at the kink point. In a second step, we simulate KLNS datasets and CLPW datasets, each with 10,000 unemployment durations where $y = E(y|x) + \varepsilon$ and $\varepsilon \sim N(0, 0.125)$. We report results from 200 simulated datasets for FG and 50 simulated datasets for CCT in Table 1. We implement a higher number of simulations for FG due to the increased dispersion of estimates based on FG bandwidth choice relative to CCT and because of the higher computational burden of CCT's methodology.

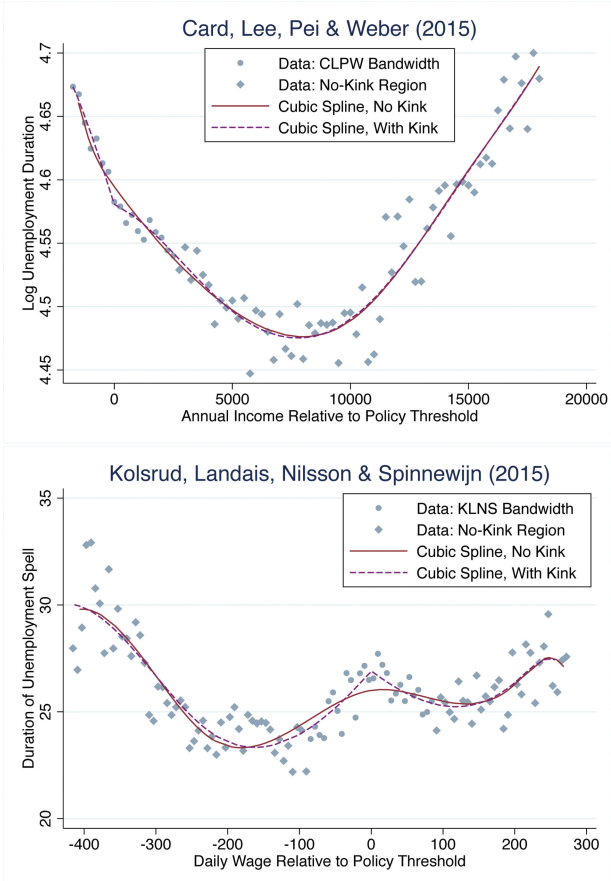

Figure 3. Conditional mean functions for RK applications in Card et al. (2015b) and Kolsrud et al. (2015). The data showing the global relationship between the outcome variable and the running variable were shared with the authors by Andrea Weber and Camille Landais. The solid maroon line denotes the fit for a natural cubic spline estimated on the data without allowing for a slope change at the kink point; the dashed maroon line denotes a natural cubic spline allowing for a slope change at the kink point.

To assess size, we implement the permutation test described in Section 3 and asymptotic inference in datasets that were generated under the assumption that the null hypothesis is true. The kink locations span the feasible range of the running variable and we consider a grid of 100 equally spaced placebo kinks. The proof in Appendix B discusses how to handle ties. In every simulated dataset, we treat every kink as a policy kink, meaning that we implement the permutation test and asymptotic inference at each of the 100 placebo kinks. We set the nominal level of the test to 5% for both inference methodologies and reject the null hypothesis if the 95% asymptotic confidence interval excludes zero. We compute the asymptotic Type I error rate as the fraction of times that the null hypothesis is rejected.

4.1.2. Results

When the null hypothesis is true, asymptotic inference with FG bandwidth choice has size far exceeding nominal levels, while

Table 1. Empirical study: regression kink estimators.

			Estimate		Interval Length			Error Rate	
Data	Kink?	Method	Mean	SD	Asymp	Permute	Error Type	Asymp	Permute
CLPW	No	FG	1.09	1.80	3.82	651.80	Type I (Size)	0.38	0.05
CLPW	No	CCT	−0.17	8.17	22.02	40.11	Type I (Size)	0.05	0.05
KLNS	No	FG	0.15	0.49	0.93	5.77	Type I (Size)	0.37	0.05
KLNS	No	CCT	0.00	0.39	1.49	1.67	Type I (Size)	0.05	0.05
CLPW	Yes	FG	4.08	1.65	5.21	653.07	Type II (1 - Power)	0.12	0.73
CLPW	Yes	CCT	7.80	13.05	48.33	40.38	Type II (1 - Power)	0.92	0.73
KLNS	Yes	FG	− 6.01	0.14	0.52	5.69	Type II (1 - Power)	0.00	0.00
KLNS	Yes	CCT	− 6.44	0.24	1.07	1.52	Type II (1 - Power)	0.00	0.00

NOTES: To compare the false rejection rate (size) and false acceptance rate (power) of asymptotic and permutation-based methods, we analyze data from two empirical applications which compute the elasticity of unemployment duration with respect to benefits: CLPW (2015) and KLNS (2015). We fit a natural cubic spline to each dataset. To estimate the Type I error rate, we assume no kink (rows 1–4), and require that the first derivative is continuous at all of the knots. To estimate the Type II error rate (rows 5–8), we allow the first derivative to change discontinuously at the policy kink. We randomly generate 10,000 unemployment durations $y = E(y|x) + \varepsilon$ with $E(y|x)$ from the cubic spline and $\varepsilon \sim N(0, 0.125)$. For ease of exposition, we have scaled up the outcome variable by 10^5 for CLPW and by 10^2 for KLNS. FG uses a local linear estimator with Fan and Gijbels's (1996) bandwidth, while CCT uses Calonico, Cattaneo, and Titiunik's (2014b) local linear estimator with quadratic bias-correction. We set the nominal level of the test to 5%. We reject the null hypothesis when the estimate at the kink location is outside the 95% confidence interval, where the interval is constructed either using standard asymptotic methods or the permutation method described in Section 3 for a set of placebo kinks on [−1400, 17800] for CLPW and [−400, 300] for KLNS. We use one draw of the dataset per simulation draw and 100 placebo kinks to estimate the Type I error rate. Reported results are based on 200 (FG) and 50 (CCT) draws for each specification.

CCT's procedure and the permutation test are reliable, as documented in Table 1. To evaluate size, we report the Type I error rate, which is the fraction of placebo kinks where each methodology rejects the null hypothesis at the 5% level. Linear estimators with FG bandwidth choice have substantially higher than nominal coverage rates, close to 40%, demonstrating a failure of asymptotic inference for these modeling choices. This failure arises because the procedure cannot distinguish between the global "U-shape" in the CLPW data-generating process and a discrete change in slope at the policy kink. In contrast, estimators following CCT's procedure perform substantially better and lead to Type I error rates much closer to the nominal level. The Type I error rate of the permutation test corresponds to the nominal level of 5% as proven in Proposition 1.

In the context of these two empirical applications, the permutation test produces longer confidence intervals precisely when asymptotic methods over-reject the null hypothesis. We construct asymptotic interval length using standard methods and permutation interval length using the methodology described in Section 3.5 and Appendix C following Rosenbaum (2002, chap. 2.6.2) and Imbens and Rubin (2015, chap. 5.7). The interval lengths based on the permutation test are substantially larger than the asymptotic ones in the case of FG bandwidth choice, in particular in the case of CLPW, where intervals based on the permutation test are two orders of magnitude larger than the asymptotic ones. In contrast, permutation-test based confidence intervals following CCT's procedure tend to be larger, in three of four specifications, but of a similar order of magnitude as asymptotic confidence intervals based on CCT.

To assess power, we implement the permutation test and asymptotic inference in datasets that were generated under the assumption that the null hypothesis is false, specifically by assuming that there is a slope change in the DGP at the policy kink. We use the same grid as in our analysis of size and again set the nominal level of the test to 5%. We analyze the Type II error rate, which is defined as the fraction of times, where the null hypothesis is rejected when there is a true policy kink.

When the null hypothesis is false, Table 1 shows that asymptotic inference with FG bandwidth choice consistently rejects the null hypothesis, while CCT's procedure and the permutation test rejects the null hypothesis in only one of two empirical applications. For CLPW, the error rate is 12% when the FG bandwidth procedure is used; in contrast, CCT and the permutation test fail to reject 92% and 73% of the time, respectively. For KLNS, the error rate is zero across all procedures, that is, the null hypothesis is rejected each time. This shows that asymptotic inference with FG bandwidth choice has high power in both settings, while CCT and the permutation test only have sufficient power in one of the settings. The permutation test delivers different conclusions in these two settings due to the differing shapes of the conditional mean functions. The estimated slope changes in CLPW away from the policy kink tend to be positive, just like the change at the true policy kink. In contrast, in KLNS, the slope change at the policy kink is negative, while the estimated slope changes away from the policy kink tend to be zero or positive.

We also assess the performance of quadratic and cubic specifications and find that standard inference achieves lower type I and higher Type II error rates as the order of the local polynomial model increases. Table A.2 in the Appendix demonstrates qualitatively similar results for quadratic specifications, while Table A.3 demonstrates that cubic specifications with standard inference also have lower Type I error rates. In addition, we also implement the permutation test by drawing on information on the institutional environment and using the past realization of changes in the earnings ceiling as described in Table A.4 in the Appendix and proposed in Section 3.3. Similar to the previous specifications, the results in Table A.4 illustrate that the permutation test achieves exact size, albeit with higher Type II error rates than standard procedures.

Taken together, the results show that linear and quadratic estimators with FG bandwidth choice over-reject the null hypothesis in data simulated based on actual RK applications. Our results show a trade-off between size and power across inference methods. Unlike asymptotic FG methods, CCT's estimator as well as permutation test-based inference lead to Type I error rates closer to the nominal level. However, CCT and the permutation test have sufficient power to detect an effect of unemployment benefits on unemployment duration in only one of the two empirical applications.

4.2. Comparing Standard and Randomization Inference: Additional Simulation Studies

To further understand where our approach has power, we extend the analysis to three artificial data-generating processes (DGPs). The conditional mean functions for the linear and nonlinear DGPs that we analyze are displayed in Figure A.1 in the Appendix. For each of these conditional mean functions, we simulate data with and without a kink at zero analogous to the procedure in the previous section and report the same set of statistics as in Table 1 to study size and power. DGP 1 is a linear function and piecewise linear in the specification with a kink. DGP 2 is based on a combination of trigonometric polynomial and exponential functions with and without kinks. DGP 3 follows a sine function. The simulations show that there is a spectrum where asymptotic inference using FG bandwidth substantially over-rejects and CCT has closer to nominal size but lower power and the permutation test has size close to nominal levels but lowest power. In general, the power of RK estimators is lowest when the DGP is highly nonlinear relative to the effect size studied.

We implement simulation procedures analogous to those described in Section 4.1. Specifically, we study a grid of 100 equally spaced placebo kinks on the interval $[-1,1]$ and analyze 200 simulated datasets using the FG methodology and 50 simulated datasets using the CCT methodology for computational reasons and due to the higher dispersion of FG estimates. The only modification is that while in the real datasets there was exactly one location for the policy kink—that is, the actual policy kink in the applications in CLPW and KLNS—in our simulated datasets we consider a grid of equally spaced locations for the policy kink in our analysis of the Type II error rate.

Table 2 reports the Type I error rate of linear RK estimators under asymptotic and permutation test-based inference. In line with the previous section, the results document that asymptotic inference with FG bandwidth choice leads to substantial overrejection of the null hypothesis—Type I error rates greater than 50%—when the DGP features nonlinearity (DGPs 2 and 3), while it performs well when the DGP is linear—in the case of DGP 1—with a Type I error rate of 5%. Asymptotic inference based on CCT's procedure outperforms the estimator with FG bandwidth, but leads to overrejection of the null hypothesis in some settings with nonlinearity (DGPs 2 and 3). Inference based on the permutation test achieves Type I error rates at the nominal level of the test by construction. In Appendix Tables A.5 and A.3, we repeat the exercise for the case of quadratic and cubic RK estimators. The relative performance of these estimators with asymptotic inference does not improve relative to the case of linear RK estimators. Again, the permutation test leads to empirical coverage at the nominal level.

Table 3 reports the Type II error rate of linear RK estimators under asymptotic and permutation test-based inference for slope changes of 5 and 20 at the kink point and shows that the permutation test suffers from lower power for highly nonlinear DGPs. Asymptotic inference with RK estimators relying on the bandwidth choice procedure in FG leads to the lowest Type II error rates and has small interval lengths compared to asymptotic inference based on CCT as well as the permutation test. Nonetheless, comparing the mean of the estimates across specifications reveals that CCT's estimator is much less biased, in particular in the case of DGP 3, which is highly nonlinear. The results also suggest that the power of the permutation test is particularly low when the DGP is very nonlinear. Curvature makes the implementation of the RK design particularly problematic due to misspecification bias, as discussed in Section 2.3 and pointed out by CCT. In the case of highly nonlinear DGPs, point estimates are also much more dispersed as indicated by the standard deviation and longer intervals for the permutation test. In Appendix Tables A.6 and A.3 (last six rows), we also report results for the case of quadratic and cubic estimators.

5. Applying the Permutation Test to Existing RD Applications

In this setting, we briefly document results of applying the permutation test to RD designs based on two well-known studies by Lee (2008) and Ludwig et al. (2007) and illustrate that standard asymptotic inference and the permutation test deliver similar conclusions in the RD setting. We display the conditional mean functions in Figure A.2 in the Appendix. To apply our analysis to an RD setting, we modify equations (8) and (9) to allow for an intercept shift at the discontinuity:

$$\mathbf{v}^k \equiv \begin{pmatrix} 1 & \mathbf{1}(v_1 \geq k) & (v_1 - k) & (v_1 - k)\mathbf{1}(v_1 \geq k) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{1}(v_1 \geq k) & (v_n - k) & (v_n - k)\mathbf{1}(v_n \geq k) \end{pmatrix} \tag{10}$$

$$T(\mathbf{v}, \mathbf{y}, k) \equiv (0 \quad 1 \quad 0 \quad 0)' (\mathbf{v}^{k\prime}\mathbf{v}^k)^{-1}\mathbf{v}^{k\prime}\mathbf{y}. \tag{11}$$

Table 2. Simulation study: size of regression kink estimators.

		Estimate		Interval Length		Type I Error Rate	
DGP	Method	Mean	SD	Asymp	Permute	Asymp	Permute
1	FG	−0.00	0.15	0.44	9.18	0.05	0.05
1	CCT	0.01	0.55	1.73	4.14	0.05	0.05
2	FG	−0.89	1.81	2.67	10.92	0.56	0.05
2	CCT	0.07	2.01	5.21	8.76	0.24	0.05
3	FG	−1.86	20.42	3.10	117.07	0.94	0.05
3	CCT	−0.22	6.03	18.87	23.16	0.15	0.05

NOTES: To compare the false rejection rate (size) of asymptotic and permutation-based methods, we analyze the data-generating processes displayed in Figure A.1. For every DGP, we randomly generate 10,000 observations with x distributed uniformly on $[-2,2]$ and $y = E(y|x) + \varepsilon$ with $\varepsilon \sim N(0, 0.25)$. FG uses a local linear estimator with Fan and Gijbels's (1996) bandwidth, while CCT uses Calonico, Cattaneo, and Titiunik's (2014b) estimator. We set the nominal level of the test to 5%. We reject the null hypothesis when the estimate at the kink location is outside the 95% confidence interval, where the interval is constructed either using standard asymptotic methods or the permutation method described in Section 3 for a set of 100 placebo kinks on $[-1,1]$. In this setting, the null hypothesis is true by assumption and the Type I error rate for an accurate estimation method should be 5%. Reported results are based on 200 (FG) and 50 (CCT) draws for each specification.

Table 3. Simulation study: power of regression kink estimators.

DGP	True Kink Size	Method	Estimate		Interval Length		Type II Error Rate	
			Mean	SD	Asymp	Permute	Asymp	Permute
1	20	FG	19.99	0.14	0.46	120.82	0.00	0.04
1	20	CCT	20.06	0.47	1.69	3.53	0.00	0.00
2	20	FG	19.02	1.80	2.67	10.26	0.00	0.05
2	20	CCT	20.34	1.63	5.23	8.92	0.00	0.00
3	20	FG	17.81	20.22	3.13	662.21	0.06	0.89
3	20	CCT	20.07	6.25	19.43	23.13	0.02	0.15
1	5	FG	5.01	0.15	0.43	61.67	0.00	0.25
1	5	CCT	4.93	0.56	1.79	3.57	0.00	0.08
2	5	FG	4.12	1.88	2.65	10.72	0.09	0.57
2	5	CCT	5.12	1.84	5.25	9.35	0.12	0.39
3	5	FG	2.91	20.52	3.10	138.10	0.04	0.94
3	5	CCT	4.57	5.35	19.26	23.02	0.82	0.87

NOTES: To compare the false acceptance rate (power) of asymptotic and permutation-based methods, we analyze the data-generating processes displayed in Figure A.1. For every DGP, we randomly generate 10,000 observations with x distributed uniformly on $[-2,2]$ and $y = E(y|x) + \varepsilon$ with $\varepsilon \sim N(0, 0.25)$. FG uses a local linear estimator with Fan and Gijbels's (1996) bandwidth, while CCT uses Calonico, Cattaneo, and Titiunik's (2014b) estimator. In our baseline specification, we randomly choose a kink location on $[-1,1]$ with a kink size specified in column 2. We set the nominal level of the test to 5%. We accept the null hypothesis when the 95% confidence interval includes zero, where the interval is constructed either using standard asymptotic methods or the permutation method described in Section 3 for a set of 100 placebo kinks on $[-1,1]$. Reported results are based on 200 (FG) and 50 (CCT) draws for each specification.

Table 4. Empirical study: regression discontinuity estimator.

Data	Discontinuity?	Method	Estimate		Interval Length		Error	Error Rate	
			Mean	SD	Asymp	Permute		Asymp	Permute
Lee	No	IK	0.00	0.10	0.35	0.79	Type I (Size)	0.06	0.05
Lee	No	CCT	− 0.00	0.04	0.15	0.70	Type I (Size)	0.05	0.05
LM	No	IK	0.00	1.25	4.54	16.87	Type I (Size)	0.06	0.05
LM	No	CCT	0.01	1.56	5.35	8.17	Type I (Size)	0.06	0.05
Lee	Yes	IK	0.45	0.09	0.36	0.80	Type II (1 - Power)	0.00	0.00
Lee	Yes	CCT	0.44	0.04	0.14	0.70	Type II (1 - Power)	0.00	0.00
LM	Yes	IK	−1.65	1.23	5.15	18.89	Type II (1 - Power)	0.77	0.77
LM	Yes	CCT	−1.64	1.62	6.04	8.42	Type II (1 - Power)	0.80	0.83

NOTES: To compare the false rejection rate (size) and false acceptance rate (power) of asymptotic and permutation-based methods, we analyze data from Lee (2008) and Ludwig-Miller (LM, 2007). For the "Yes" Discontinuity rows, we estimate a natural cubic spline, allowing for a jump in the intercept in the knot at the policy discontinuity. For the "No" Discontinuity rows, we take the true DGP and subtract the estimated jump at the policy discontinuity so that the conditional mean function is continuous at zero. For Lee, we randomly generate results from 10,195 elections with a victory probability equal to the predicted mean of the cubic spline. For LM, we randomly generate a mortality rate in 2,810 counties using a predicted mean of the cubic spline and a standard deviation of 5.7. IK uses a local linear estimator with Imbens Kalyanaraman's (2012) bandwidth, while CCT uses Calonico, Cattaneo, and Titiunik's (2014b) local linear estimator with quadratic bias-correction. We compute two-sided asymptotic p-values and permutation-based p-values from placebo kinks on $[-48,47]$ for Lee and $[-40,9.5]$ for LM using the method described in Section 3 and set the nominal level of the test to 5%. The permutation test is based on 100 placebo kinks. Reported results are based on 200 (FG) and 50 (CCT) draws of each dataset.

The analysis of size and power of asymptotic and permutation test-based inference follows the same structure as in Sections 4.1 and 4.2 and we report the results of our simulations in Table 4, which also features randomization inference for the RD design based on the approach in Cattaneo, Frandsen, and Titiunik (2015). The simulation studies reveal that for both RD applications the permutation test and asymptotic inference lead to actual size close to nominal size and have comparable Type II error rates.

We also find broadly similar results using a different permutation test proposed by Cattaneo, Frandsen and Titiunik (2015, CFT in the following), which holds the location of the discontinuity fixed and randomly varies which observations are assigned to the treatment and control. Specifically, we implemented CFT for 50 simulated datasets based on Lee (window of 1) and LM (window of 1.1, following Cattaneo, Titiunik, and Vazquez-Bare forthcoming). With nominal size of 5%, the Type I error rates were 0.28 (Lee) and 0.04 (LM), while the Type II error rates were 0.00 (Lee) and 0.78 (LM). CFT's performance is similar to the asymptotic and permutation method's in Table 4 except that CFT over-rejects the null in the Lee empirical setting.

6. Conclusion

We develop a permutation test for the regression kink design and document its performance compared to standard asymptotic inference. The thought experiment underlying our test differs from the one of standard asymptotic inference, which is based on the thought experiment of drawing observations from a large population so that standard errors reflect sampling uncertainty. Our test follows the randomization inference approach in taking the sample as given and takes the assignment of treatment—here the location of the kink point—as a random variable and thus the source of uncertainty.

Nonlinearity is ubiquitous in many of the settings in which RK designs are applied. In the presence of such nonlinearity, the test can offer a complement to existing methods and is more robust than asymptotic inference with standard bandwidth choice. Based on the results of our simulation studies, we recommend that practitioners: (1) avoid using linear and quadratic RK estimators with FG bandwidth choice, (2) use CCT's robust procedure as preferred procedure for estimating slope changes, (3) use the distribution of placebo estimates to assess whether they will have power to detect economically

meaningful results in their context, (4) report p-values constructed by comparing their point estimate to the distribution of placebo estimates, and (5) report the robustness of the permutation test to different assumptions on the placebo kink distribution.

Supplementary Materials

The online supplementary materials contain the appendices for the article.

Acknowledgments

We thank Alberto Abadie, Josh Angrist, David Card, Matias Cattaneo, Avi Feller, Edward Glaeser, Paul Goldsmith-Pinkham, Guido Imbens, Maximilian Kasy, Zhuan Pei, Mikkel Plagborg-Møller, and Guillaume Pouliot as well as participants at Harvard University's Research in Statistics Seminar and the Research in Econometrics Seminar for helpful comments and discussions. We are especially thankful to Gary Chamberlain, Raj Chetty and Larry Katz for guidance and suggestions. We thank Andrea Weber and Camille Landais for sharing supplemental figures based on administrative UI data. We thank Harvard's Lab for Economic Applications and Policy for financial support and Carolin Baum, David Poth, Michael Schöner, Shawn Storm, Cody Tuttle, and Thorben Wölk for excellent research assistance.

References

Abadie, A., Athey, S., Imbens, G., and Wooldridge, J. (2014), "Finite Population Causal Standard Errors," Working Paper. [3]

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [3]

Ando, M. (forthcoming), "How Much Should We Trust Regression-Kink-Design Estimates?" *Empirical Economics*, 53, 1287–1322. [3]

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249–75. [3]

Britto, D. G. C. d. (2015), "Unemployment Insurance and the Duration of Employment: Evidence From a Regression Kink Design," *SSRN Working Paper 2648166*. [1,3]

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2016), "Coverage Error Optimal Confidence Intervals for Regression Discontinuity Designs," *Working Paper*. [1]

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014a), "Robust Data-Driven Inference in the Regression-Discontinuity Design," *Stata Journal*, 14, 909–946. [2,4]

——— (2014b), "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326. [1,10]

——— (2015), "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110, 1753–1769. [2]

Card, D., Johnston, A., Leung, P., Mas, A., and Pei, Z. (2015a), "The Effect of Unemployment Benefits on the Duration of Unemployment Insurance Receipt: New Evidence from a Regression Kink Design in Missouri, 2003–2013," *American Economic Review*, 105, 126–30. [1,3]

Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015b), "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83, 2463–2483. [2]

Cattaneo, M., Frandsen, B., and Titiunik, R. (2015), "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," *Journal of Causal Inference*, 3, 1–24. [3,6,10]

Cattaneo, M., Titiunik, R., and Vazquez-Bare, G. (forthcoming), "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management*, 36, 643–681. [6,10]

Chetty, R., Looney, A., and Kroft, K. (2009), "Salience and Taxation: Theory and Evidence," *The American Economic Review*, 99, 1145–1177. [3]

Engström, P., Nordblom, K., Ohlsson, H., and Persson, A. (2015), "Tax Compliance and Loss Aversion," *American Economic Journal: Economic Policy*, 7, 132–164. [3]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and its Applications* (vol. 66), Boca Raton, FL: Chapman and Hall. [2,4,5]

Fisher, R. (1935), *The Design of Experiments*, Oxford, England: Oliver and Boyd. [3,6]

Florens, J.-P., Heckman, J. J., Meghir, C., and Vytlacil, E. (2008), "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. [3,4]

Gelman, A., and Imbens, G. (2014), "Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs," *National Bureau of Economic Research Working Paper*. [3,6]

Ho, D. E., and Imai, K. (2006), "Randomization Inference With Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election," *Journal of the American Statistical Association*, 101, 888–900. [3]

Imbens, G., and Kalyanaraman, K. (2012), "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79, 933–959. [10]

Imbens, G. W., and Lemieux, T. (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635. [1,3]

Imbens, G. W., and Rosenbaum, P. R. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Series A, 168, 109–126. [3]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, and in the Social and Biomedical Sciences*, Cambridge, UK: Cambridge University Press. [7,8]

Kolsrud, J., Landais, C., Nilsson, P., and Spinnewijn, J. (2015), "The Optimal Timing of Unemployment Benefits: Theory and Evidence From Sweden," *IZA Discussion Paper 9185*. [1,2]

Kyyrä, T., and Pesola, H. (2015), "The Effects of Unemployment Insurance Benefits on Subsequent Labor Market Outcomes: Evidence From an RKD Approach," VATT Working Paper. [1,3]

Landais, C. (2015), "Assessing the Welfare Effects of Unemployment Benefits Using the Regression Kink Design," *American Economic Journal: Economic Policy*, 7, 243–278. [1,3]

Lee, D. S. (2008), "Randomized Experiments From Non-Random Selection in US House Elections," *Journal of Econometrics*, 142, 675–697. [10]

Lehmann, E., and Stein, C. (1949), "On the Theory of Some Non-Parametric Hypotheses," *The Annals of Mathematical Statistics*, 20, 28–45. [3]

Ludwig, J., Miller, D. L., (2007), "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *The Quarterly Journal of Economics*, 122, 159–208. [9]

Nielsen, H. S., Sørensen, T., and Taber, C. R. (2010), "Estimating the Effect of Student Aid on College Enrollment: Evidence From a Government Grant Policy Reform," *American Economic Journal: Economic Policy*, 2, 185–215. [1]

Pitman, E. (1937), "Significance Tests Which May be Applied to Samples From any Populations," *Supplement to the Journal of the Royal Statistical Society*, 4, 119–130. [6]

Romano, J. P. (1990), "On the Behavior of Randomization Tests Without a Group Invariance Assumption," *Journal of the American Statistical Association*, 85, 686–692. [6]

Rosenbaum, P. R. (2001), "Stability in the Absence of Treatment," *Journal of the American Statistical Association*, 96, 210–219. [3]

——— (2002), *Observational Studies* (2nd ed.), New York: Springer. [3]

Sovago, S. (2015), "The effect of the UI benefit on labor market outcomes-regression kink evidence from the Netherlands," Working Paper. [1,3]

Thistlethwaite, D. L., and Campbell, D. T. (1960), "Regression-Discontinuity Analysis: An Alternative to the ex Post Facto Experiment," *Journal of Educational Psychology*, 51, 309. [1]

Welch, W. J. (1990), "Construction of Permutation Tests," *Journal of the American Statistical Association*, 85, 693–698. [3]

Welch, W. J., and Gutierrez, L. G. (1988), "Robust Permutation Tests for Matched-Pairs Designs," *Journal of the American Statistical Association*, 83, 450–455. [3]