TECHNICAL WORKING PAPER SERIES

REGRESSION DISCONTINUITY INFERENCE WITH SPECIFICATION ERROR

David S. Lee
David Card

Technical Working Paper 322
http://www.nber.org/papers/T0322

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2006

We are grateful to Guido Imbens and Thomas Lemieux for helpful suggestions, and to Michael Jansson, James Powell, Keisuke Hirano, and participants in the 2003 Banff International Research Station Regression Discontinuity Conference for helpful discussions. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Regression Discontinuity Inference with Specification Error
David S. Lee and David Card
NBER Technical Working Paper No. 322
March 2006
JEL No. C1, C5, J0

ABSTRACT

A regression discontinuity (RD) research design is appropriate for program evaluation problems in which treatment status (or the probability of treatment) depends on whether an observed covariate exceeds a fixed threshold. In many applications the treatment-determining covariate is discrete. This makes it impossible to compare outcomes for observations "just above" and "just below" the treatment threshold, and requires the researcher to choose a functional form for the relationship between the treatment variable and the outcomes of interest. We propose a simple econometric procedure to account for uncertainty in the choice of functional form for RD designs with discrete support. In particular, we model deviations of the true regression function from a given approximating function -- the specification errors -- as random. Conventional standard errors ignore the group structure induced by specification errors and tend to overstate the precision of the estimated program impacts. The proposed inference procedure that allows for specification error also has a natural interpretation within a Bayesian framework.

David S. Lee
Department of Economics
549 Evans Hall, #3880
University of California
Berkeley, CA 94720-3880
and NBER
dslee@econ.berkeley.edu

David Card
Department of Economics
549 Evans Hall, #3880
University of California
Berkeley, CA 94720-3880
and NBER
card@econ.berkeley.edu

1 Introduction

In the classic regression-discontinuity (RD) design (Thistlethwaite and Campbell, 1960) the treatment status of an observation is determined by whether an observed covariate is above or below a known threshold. If the covariate is predetermined it may be plausible to think of treatment status is "as good as randomly assigned" among the subsample of observations that fall just above and just below the threshold.[1] As in a true experiment, no functional form assumptions are necessary to estimate program impacts when the treatment-determining covariate is continuous: one simply compares average outcomes in small neighborhoods on either side of the threshold. The width of these neighborhoods can be made arbitrarily small as the sample size grows, ensuring that observed and unobserved characteristics of observations in the treatment and control groups are identical in the limit. This idea underlies the approach of Hahn et al. (2001) and Porter (2003), who describe non-parametric and semi-parametric estimators of regression-discontinuity gaps.

In many applications where the RD design seems compelling, however, the covariate that determines treatment is inherently discrete or is only reported in coarse intervals. For example, government programs like Medicare and Medicaid have sharp age-related eligibility rules that lend themselves to an RD framework, but in most data sets age is only recorded in months or years. In the discrete case it is no longer possible to compute averages within arbitrarily small neighborhoods of the cutoff point, even with an infinite amount of data. Instead, researchers have to choose a particular functional form for the model relating the outcomes of interest to the treatment-determining variable. Indeed, with an irreducible gap between the "control" observations just below the threshold and the "treatment" observations just above, the causal effect of the program is not even identified in the absence of a parametric assumption about this function.

In this paper we propose a simple procedure for inference in RD designs in which the treatment-determining covariate is discrete. The basic idea is to model the deviation between the expected value of the outcome and the predicted value from a given functional form as a random specification error. Modeling potential specification error in this way has a number of immediate implications.

[1] This assumption may or may not be plausible, depending upon the context. In particular, if the treatment is under perfect control of individuals, and there are incentives to "sort" around the threshold, the RD design may be invalid. On the other hand, even when individuals have partial control over the covariate, as long as there is a stochastic component that has continuous density, the treatment variable is as good as (locally) randomly assigned. See Lee (2006) for details.

Most importantly, it introduces a common component of variance for all the observations at any given value of the treatment-determining covariate. This creates a problem similar to the one analyzed by Moulton (1990) for multi-level models in which some of the covariates are only measured at a higher level of aggregation (e.g., micro models with state-level covariates). Random specification errors can be easily incorporated in inference by constructing sampling errors that include a grouped error component for different values of the treatment-determining covariate. The use of "clustered" standard errors will generally lead to wider confidence intervals that reflect the imperfect fit of the parametric function away from the discontinuity point.

More subtly, inference in an RD design involves extrapolation from observations below the threshold to construct a counterfactual for observations above the threshold. As in a classic out-of-sample forecasting problem, the sampling error of the counterfactual prediction for the point of support just beyond the threshold includes a term reflecting the expected contribution of the specification error at that point. Since the estimated (local) treatment effect is just the difference between the mean outcome for these observations and the counterfactual prediction, the precision of the estimated treatment effect depends on whether one assumes that *the same* specification error would prevail in the counterfactual world. If so, this error component vanishes. If not, the confidence interval for the local treatment effect has to be widened even further.

The paper is organized as follows. Section 2 describes the RD framework and why discreteness in the treatment-determining covariate implies that the treatment effect is not identified without assuming a parametric functional form. Section 3 describes the proposed inference procedure under a model where specification errors are considered random. Section 4 describes a modified procedure under less restrictive assumptions about the specification errors. Section 5 proposes an alternative, efficient estimator for the treatment effect, and Section 6 relates this estimator to a Bayesian approach. Section 7 concludes.

2 The Regression Discontinuity Design with Discrete Support

2.1 The Problem of Discreteness

To illustrate how discreteness causes problems for identification in an RD framework, consider the following potential outcomes formulation.[2] There is a binary indicator D of treatment status which is determined by whether an observed covariate X is above or below a known threshold x_0: $D = 1[X \geq x_0]$. Let Y_1 represent the potential outcome if an observation receives treatment and let Y_0 represent the potential outcome if not. The goal is to estimate $E[Y_1 - Y_0|X = x_0]$, the average treatment effect at the threshold. As usual, Y_1 and Y_0 are not simultaneously observed for any individual. Instead, we observe $Y = DY_1 + (1 - D)Y_0$.

When the support of X is continuous and certain smoothness assumptions are satisfied, $E[Y_1 - Y_0|X = x_0]$ is identified by the discontinuity in the regression function for the *observed outcome* Y at x_0. More specifically, if $E[Y_1|X = x]$ and $E[Y_0|X = x]$ are both continuous in x at x_0, then

$$
\begin{aligned}
& E[Y|X = x_0] - \lim_{e \to 0^+} E[Y|X = x_0 - e] \\
& \quad = E[Y_1|X = x_0] - \lim_{e \to 0^+} E[Y_0|X = x_0 - e] \\
& \quad = E[Y_1 - Y_0|X = x_0]
\end{aligned}
$$

This idea is illustrated in Figure 1. The data identifies $E[Y_1|X = x]$ when $x \geq x_0$, and $E[Y_0|X = x]$ when $x < x_0$, as indicated by the solid lines. Because of the discontinuous rule that determines treatment status, the data do not identify the dashed lines, or the counterfactual mean $E[Y_0|X = x_0]$ (the open circle). What the data do yield is $E[Y_0|X = x_0 - e]$, which can be an arbitrarily good approximation to $E[Y_0|X = x_0]$, with e sufficiently small. In this setting, non-parametric and semi-parametric procedures for estimation are appropriate (Hahn et al. (2001); Porter (2003)), particularly when the sample size is large, in which case one can precisely estimate local averages just above and below x_0.

This limiting argument, however, does not work when the support of X is discrete. Suppose X can take on J distinct values $(x_1, ..., x_J)$ and let $x_k = 0$ be the value of the covariate at the

[2]For an overview of the potential outcomes framework for program evaluation problems see, for example, Angrist and Krueger (1999).

discontinuity threshold. Figure 2 is a discrete analogue to Figure 1. As before, the counterfactual mean $E[Y_0|X=0]$ is unobservable. Here, the discrete analogue to $E[Y|X=x_0]-\lim_{e\to 0^+} E[Y|X=x_0-e]$ is $E\left[Y|X=0\right]-E\left[Y|X=x_{k-1}\right]$, which substantially over-estimates the true effect $E[Y_1-Y_0|X=0]$.

Unlike the continuous case, even if the population quantities $E\left[Y|X=x_j\right]$ $(j=1,\ldots,J)$ are known, $E[Y_1-Y_0|X=0]$ remains unidentified. Identification can be achieved by assuming that the regression function can be expressed as

$$E\left[Y|X=x_j\right]=D_j\beta_0+h\left(x_j\right) \tag{1}$$

where $h\left(\cdot\right)$ is a continuous function, $D_j=1\left[x_j\geq 0\right]$, and $h\left(0\right)=E\left[Y_0|X=0\right]$. With this specification β_0 (equal to $E[Y_1-Y_0|X=0]$) is the parameter of interest. Equation (1) is equivalently expressed as a model for the micro-data

$$Y_{ij}=D_j\beta_0+h\left(x_j\right)+\varepsilon_{ij} \tag{2}$$

where Y_{ij} is the outcome for the ith individual with the jth value of X, and $\varepsilon_{ij}\equiv Y_{ij}-E\left[Y_{ij}|X=x_j\right]$, with conditional variance $\sigma^2_{\varepsilon j}$.

It is important to note that β_0 is only identified when $h\left(\cdot\right)$ is determined by a limited number of parameters. With only J distinct values of X, if $h\left(\cdot\right)$ contains J or more parameters, there is no way for the data to distinguish between a discontinuity in the regression function, and a continuous function that connects $E\left[Y|X=x_{k-1}\right]$ and $E\left[Y|X=0\right]$.

In addition, the asymptotic arguments used to justify non-parametric estimation of β_0 (as in Hahn et al. (2001)) cannot be applied here. Even with an infinite amount of data, there are no data in a region in an “arbitrarily” small neighborhood below 0. For example, a one-sided kernel (or local linear) estimator will, in the limit, place no weight on observations for which $X\leq x_{k-1}$, and all of the weight on observations slightly below 0 (but above x_{k-1}). But because of the discrete support there are no data in this neighborhood.

2.2 Parametric Estimation and Inference

It is common practice for researchers to estimate RD designs by regressing Y on a low-order polynomial in x_j, and the treatment indicator D_j (e.g., Card and Shore-Sheppard (2004); Kane (2003); DiNardo and Lee (2004); Lee (2006)). If the polynomial function is the correct form for $h(\cdot)$, then conventional least-squares inference is appropriate.

When the covariate is discrete, a simple goodness-of-fit statistic for the polynomial functional form can be calculated as

$$G \equiv \frac{(ESS_R - ESS_{UR})/(J-K)}{ESS_{UR}/(N-J)} \tag{3}$$

where ESS_R is the (restricted) error sum of squares from estimating (2) with a polynomial in x_j for $h(x_j)$, and ESS_{UR} is the (unrestricted) sum of squares from regressing Y_{ij} on a full set of dummy variables for the J values of X. Under normality (and homoskedasticity) of ε_{ij}, this statistic is distributed as $F(J-K, N-K)$, where K is the number of parameters estimated in (2) and N is the number of observations.[3] If the statistic exceeds the critical value, it suggests that the polynomial function is too restrictive.

A rejection of the polynomial, however, need not imply that the least squares estimate $\widehat{\beta}$ is inconsistent for β_0. Following White (1980) and Chamberlain (1994), $\widehat{\beta}$ is consistent for β^*, the discontinuity in the function that is the least squares approximation to the true function in Equation (1).[4] The difference between β^* and β_0 is unknown, but may be small (or could even be zero), even if the goodness-of-fit statistic leads one to reject the polynomial specification.

Despite this possibility, it seems natural for a researcher to be relatively more "skeptical" of $\widehat{\beta}$ as an estimate of β_0 when the goodness-of-fit statistic rejects the model, and relatively more "confident"

[3]Under non-normal (homoskedastic) ε_{ij}, $(J-K)\cdot G$ will be asymptotically distributed as $\chi^2(J-K)$.Letting W_j be the vector of regressors (the polynomial and dummy variable), under heteroskedastic ε_{ij}, one can compute the statistic as

$$\widetilde{G} \equiv \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{\widehat{\sigma}^2_{\varepsilon j}}\left(Y_{ij} - W_j\widehat{\theta}\right)^2 - \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{\widehat{\sigma}^2_{\varepsilon j}}\left(Y_{ij} - \overline{Y}_j\right)^2$$

which is a version of $ESS_R - ESS_{UR}$, weighted by the reciprocal of $\widehat{\sigma}^2_{\varepsilon j} = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2$. Equivalently, $\widetilde{G} = [\ \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{1}{\widehat{\sigma}^2_{\varepsilon j}}\left(Y_{ij} - W_j\widehat{\theta}\right)^2]\ -N$, or $\widetilde{G} = \sum_{j=1}^{J}\frac{n_j}{\widehat{\sigma}^2_{\varepsilon j}}\left(\overline{Y}_j - W_j\widehat{\theta}\right)^2$. It can be shown that $\widetilde{G}$ is distributed asymptotically as $\chi^2(J-K)$.

[4]When this interpretation of $\widehat{\beta}$ is adopted, the conventional heteroskedasticity-consistent standard errors are appropriate for inferences about β^*. Chamberlain (1994) derives the asymptotic distribution of minimum distance estimators under mis-specification, and shows the equivalence of the variance to the heteroskedasticity-consistent variance in a least-squares regression.

in $\widehat{\beta}$ when the F-statistic is relatively close to 1. The inference procedures proposed below formalizes this notion. We propose to inflate conventional standard errors to reflect "modeling uncertainty". As we show below, the degree of inflation is directly related to the goodness-of-fit statistic G.

3 Random Specification Error

Suppose a polynomial is chosen to approximate $h(\cdot)$. The regression in Equation (2) can be re-written as

$$Y_{ij} = \alpha_0 + D_j\beta_0 + X_j\gamma_0 + a_j + \varepsilon_{ij} \tag{4}$$

where X_j is a row vector of polynomial terms in x_j (with the normalization $x_k = 0$), and $a_j \equiv h(x_j) - X_j\gamma_0$ is specification error – the degree to which the true function $h(\cdot)$ deviates from the polynomial function.[5] Throughout the paper, we focus on the case of no other individual-level covariates, but it will be clear that the analysis can be extended to include such covariates. Moreover, if the RD design is valid, they can be excluded in the same way that baseline covariates can be excluded in an analysis of a randomized experiment (see, for example, the discussion in Lee (2006)). We also focus on the case of the "sharp" RD design – in which the treatment is a deterministic function of X. It will be clear, however, that these ideas also extend to "fuzzy" RD designs – in which there is imperfect compliance of the treatment.[6] The Appendix describes how to apply the inference procedures described below to the "fuzzy" design.

Our first proposed inference procedure stems from treating this modeling error as *random* and orthogonal to X (or, alternatively, $E[a_j|X = x_j] = 0,\ j = 1, \ldots, J$). This assumption implies that the least squares estimate $\widehat{\beta}$ will be consistent for β_0. More importantly, it implies that the conventional heteroskedasticity-consistent variance estimators will generally be inconsistent for the true variance of $\widehat{\beta}$. This is because the randomness in a_j has induced a within-group correlation (at the j level) in the error. Essentially, the specification error here is a random effect, and it is well known that standard error estimates that ignore this within-group correlation will under-state the true variability of the least squares estimates (Moulton, 1990).

Thus, our first observation is that *if the polynomial function is viewed as an approximation that*

[5] X_j may include interactions between the polynomial terms and the treatment indicator. This allows the regression function to have different derivatives (up to the order of the interaction terms) on either side of the threshold.

[6] Discussion of the distinction between the "sharp" and "fuzzy" designs can be found in Hahn et al. (2001).

nonetheless gives unbiased estimates of the discontinuity, and specification errors are considered to be random, then conventional standard error formulas understate the variability of the least-squares estimate of the discontinuity gap.

Letting $\theta_0 \equiv (\alpha_0, \beta_0, \gamma_0)$, and $\widehat{\theta}$ be the least squares estimator in the regression of Y_{ij} on $W_j \equiv (1, D_j, X_j)$, a consistent estimator for the asymptotic variance of $\sqrt{N}\left(\widehat{\theta} - \theta_0\right)$ is given by

$$\left(\tfrac{1}{N}\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} W_j' W_j\right)^{-1} \left(\left(\tfrac{J}{N}\right)\tfrac{1}{J}\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} W_j'\left(Y_{ij} - W_j\widehat{\theta}\right)\right)\left(\sum_{i=1}^{n_j} W_j\left(Y_{ij} - W_j\widehat{\theta}\right)\right)\right) \cdot \left(\tfrac{1}{N}\textstyle\sum_{j=1}^{J}\sum_{i=1}^{n_j} W_j' W_j\right)^{-1} \tag{5}$$

with n_j finite as $J \to \infty$. The computation of this variance is available as a standard option in today's typical statistical analysis software.[7]

The assumption that a_j is orthogonal to X may seem restrictive, but it should be noted that conventional inference using parametric functional forms (like polynomial functions) implicitly imposes the strictly more restrictive assumption of no specification error, $a_j = 0$.

3.1 Clustered Standard Errors and the Goodness-of-fit Statistic

There is a connection between the goodness-of-fit statistic given in (3), and the difference between the non-clustered and clustered variance estimators.

To see this, first note that (5) can be re-written as

$$\widehat{V}_C \equiv \left(\frac{1}{J}\sum_{j=1}^{J}\frac{n_j}{N/J}W_j'W_j\right)^{-1}\left(\frac{1}{J}\sum_{j=1}^{J}\left(\frac{n_j}{N/J}\right)^2 W_j'W_j\left(\overline{Y}_j - W_j\widehat{\theta}\right)^2\right)\left(\frac{1}{J}\sum_{j=1}^{J}\frac{n_j}{N/J}W_j'W_j\right)^{-1} \tag{6}$$

where $\overline{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$; note that this estimator has been re-normalized to be consistent for the asymptotic variance for $\sqrt{J}\left(\widehat{\theta} - \theta_0\right)$, rather than for $\sqrt{N}\left(\widehat{\theta} - \theta_0\right)$. This shows that the clustered standard error formula in the micro-level regression is equivalent to using the conventional heteroskedasticity-consistent standard error in a "cell-level" regression of $\overline{Y}_j$ on W_j, weighting each cell by the weight $\frac{n_j}{N/J}$.[8]

[7] For example, in STATA, this variance can be computed by regressing Y_{ij} on W_j, and using the "cluster" option, where the groups are defined by the discrete values of X.

[8] The sum of these weights across the J cells is equal to J.

Consider the simplified case where $n_j = n_0$ for all cells, so the weight becomes 1, and that a_j and ε_{ij} have constant variance σ_a^2 and σ_ε^2 across all J cells. In this case, we have

$$\widehat{V}_C \xrightarrow{p} E\left[W_j' W_j\right]^{-1}\left(\sigma_a^2 + \frac{\sigma_\varepsilon^2}{n_0}\right)$$

while the non-clustered variance estimator $\widehat{V}_{NC} \xrightarrow{p} E\left[W_j' W_j\right]^{-1}\left(\sigma_a^2 + \sigma_\varepsilon^2\right)(1/n_0)$.[9] It follows that the ratio of the clustered to the non-clustered estimated variance will converge in probability to

$$n_0 \frac{\sigma_a^2 + \frac{\sigma_\varepsilon^2}{n_0}}{\sigma_a^2 + \sigma_\varepsilon^2}. \tag{7}$$

This quantity represents the extent to which the non-clustered variance must be "inflated".

This ratio can be estimated by a Lagrange Multiplier version of the goodness-of-fit statistic in G in (3), which is given by

$$\begin{aligned} \frac{1}{J-K} LM &= \frac{\frac{1}{J-K}\left(ESS_R - ESS_{UR}\right)}{\frac{1}{N} ESS_R} \\ &= n_0 \frac{\frac{1}{J-K}\sum_{j=1}^{J}\left(\overline{Y}_j - W_j\widehat{\theta}\right)^2}{\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_0}\left(Y_{ij} - W_j\widehat{\theta}\right)^2} \end{aligned}$$

which, with n_0 fixed and $J \to \infty$, can be shown to converge in probability to the ratio in (7).

4 Mis-specification of Counterfactual Functions

In this section, we show that the special structure of an RD design implies that in some circumstances, the clustered standard errors may still understate the variability of $\widehat{\beta}$. If the specification error is random, then it is necessary to decide how the error in estimating $E\left[Y_1 | X = x_k\right]$ is related to the specification error in estimating $E\left[Y_0 | X = x_k\right]$. As shown below, if the errors are assumed to be identical, then the approach described above is appropriate. If the errors are independent, then the standard errors for $\widehat{\beta}$ must be inflated even further.

[9] $(1/n_0)$ is added because this is the estimator for the asymptotic variance for $\sqrt{J}\left(\widehat{\theta} - \theta_0\right)$, rather than for $\sqrt{N}\left(\widehat{\theta} - \theta_0\right)$.

Before describing these two cases in detail, we provide some intuition for the difference between the two cases. As we have argued above, in the case of discrete X, non-parametric identification of the RD design is impossible. Since it is necessary to impose some functional form, estimating the "discontinuity gap" amounts to using data away from the discontinuity threshold to estimate the average outcome at the threshold.

Consider Figure 3A, which abstracts from sampling error (i.e., suppose there is an infinite amount of data per value of X). The solid dots represent $E[Y|X = x_j]$ away from the discontinuity. Essentially, we are using data from the right, as well as an approximating function, to estimate the true $E[Y_1|X = x_k]$. In the figure, the approximating function (the solid line) is not perfect, and the true $E[Y_1|X = x_k]$ is larger than that predicted by the functional form. Similarly, the extrapolation of $E[Y_0|X = x_k]$ from data on the left also under-predicts the truth. Assuming "identical" specification errors means that we are assuming that the error in our "forecast" of $E[Y_1|X = x_k]$ is of the same sign and magnitude as our forecast error of $E[Y_0|X = x_k]$, *in repeated draws of the random effect error.* One realization of this process is illustrated in Figure 3A.

Figure 3B, by contrast, depicts a single realization from a process that allows the prediction error for $E[Y_1|X = x_k]$ to be independent of the error for $E[Y_0|X = x_k]$. In the figure, the parametric functional form over-predicts $E[Y_0|X = x_k]$ and under-predicts $E[Y_1|X = x_k]$.

4.1 Identical Specification Errors

Suppose we approximate the following two counterfactual functions by the following polynomial functions

$$
\begin{aligned}
E[Y_1|X = x_j] &= \alpha_0 + X_j\gamma_0 + \beta_0 + a_{1j} \qquad (8) \\
E[Y_0|X = x_j] &= \alpha_0 + X_j\gamma_0 + a_{0j}
\end{aligned}
$$

where a_{1j} and a_{0j} are the random specification errors in the approximations for $E[Y_1|X = x_j]$ and $E[Y_0|X = x_j]$, respectively. The approximation for $E[Y_1|X = x_j]$ is parallel to the approximation for $E[Y_0|X = x_j]$, and different by exactly β_0 for each value of X.

If we assume that $a_{1j} = a_{0j}$, and we use the fact that $Y = DY_1 + (1 - D) Y_0$, then we obtain

$$E[Y|X = x_j] = \alpha_0 + X_j\gamma_0 + D_j\beta_0 + a_j$$

where $a_j \equiv D_j a_{1j} + (1 - D_j) a_{0j}$. This expression leads to the same regression specification given in (4). As before, β_0 (or, $E[Y_1 - Y_0|X = x_k]$) is the causal parameter of interest, and the clustered standard error formula is appropriate for inference.

The assumption of identical specification errors is equivalent to assuming that the same approximation error would arise whether the cell at the discontinuity point assigned to treatment or not. Equivalently, this assumption implies that that the treatment effect at the discontinuity is deterministic, that is, $E[Y_1 - Y_0|X = x_k] = \beta_0$.

One case where this assumption may be valid is when the researcher believes that the source of the approximation error is independent of treatment status. For example Card and Shore-Sheppard (2004) use a regression discontinuity design to examine the impact of the Medicaid expansions on health insurance. The family income eligibility limits for Medicaid were relaxed for children born after a certain date, and Card and Shore-Sheppard (2004) examine the relationship between Medicaid enrollment and quarter of birth. It is possible that there are small health differences by season of birth, implying that demand for Medicaid coverage varies by quarter of birth; here, a_j would reflect those seasonal differences. Arguably, the same seasonal differences would be present irrespective of treatment status.

Note that the specification errors a_{1j} and a_{0j} could be identical even when the counterfactual functions are not strictly parallel. To see this, consider the specification

$$E[Y_1|X = x_j] = \alpha_0 + X_j\gamma_1^* + \beta_0 + a_{1j}$$
$$E[Y_0|X = x_j] = \alpha_0 + X_j\gamma_0^* + a_{0j}$$

Here, the coefficients on the polynomial terms are allowed to be different. We now have

$$E[Y|X = x_j] = \alpha_0 + X_j\gamma_0^* + X_j D_j (\gamma_1^* - \gamma_0^*) + D_j\beta_0 + a_j$$

where again, $a_j \equiv D_j a_{1j} + (1 - D_j) a_{0j}$. This, too, leads to the "random-effects" regression equation given in (4), except that interactions between D_j and the polynomial terms are included. In this fully-interacted model the treatment effect function

$$E[Y_1 - Y_0|X = x_j] = X_j(\gamma_1^* - \gamma_0^*) + \beta_0 \tag{9}$$

is itself a polynomial in X. Therefore, in order to use this specification, it is necessary to assume that even if polynomials provide only an approximation to each counterfactual function separately, there is no approximation error in describing the *difference* in the counterfactual functions as a polynomial in X (at least at $X = x_k$).

4.2 Independent Specification Errors

Alternatively, one can allow $a_{1j} \neq a_{0j}$. When this is true, the treatment effect of interest is no longer equal to β_0. Instead, we have, using (8),

$$E[Y_1 - Y_0|X = x_k] = \beta_0 + a_{1k} - a_{0k}.$$

$\widehat{\beta}$ will be consistent for β_0, but not for the parameter of interest, $E[Y_1 - Y_0|X = x_k]$. Formally, with non-identical a_{1j}, a_{0j}, we have

$$\widehat{\beta} - E[Y_1 - Y_0|X = x_k] = (\widehat{\beta} - \beta_0) - (a_{1k} - a_{0k}) \tag{10}$$

where the first term converges in probability to 0 as $J \to \infty$, while the second term does not. No matter how much data are available, there is still uncertainty in the average treatment effect, induced by uncertainty about the realizations of a_{1k}, a_{0k}.

Inference about $E[Y_1 - Y_0|X = x_k]$ requires accounting for this uncertainty. In particular, it is necessary to assume that the specification errors are drawn from some parametric distribution. A natural choice is to assume that a_{1j} and a_{0j} are jointly and mutually independent, for each j. Independence implies that the forecast error for $E[Y_1|X = x_k]$ is independent of the forecast error for $E[Y_0|X = x_k]$.

In the Appendix, it is shown that, assuming that a_{1j} and a_{0j} have equal variance σ_a^2 across all

j values

$$\frac{\widehat{\beta} - E\left[Y_1 - Y_0 | X = x_k\right]}{\sqrt{\widehat{V\left(\widehat{\beta}\right)} + 2\widehat{\sigma}_a^2}} \xrightarrow{d} N\left(0,1\right) \tag{11}$$

where $\widehat{V\left(\widehat{\beta}\right)} \equiv \frac{\hat{V}_C}{J}$ is the standard cluster-consistent variance estimator.[10] $\widehat{\sigma}_a^2$ is a consistent estimator of σ_a^2, given by

$$\widehat{\sigma}_a^2 \equiv \frac{1}{N}\sum_{j=1}^{J} n_j \left(\overline{Y}_j - W_j\widehat{\theta}\right)^2 - \frac{1}{N}\sum_{j=1}^{J}\frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 \tag{12}$$

The first term is the weighted variance of the mean residual from the regression. With n_j fixed, and as $J \to \infty$, it converges in probability to $\sigma_a^2 + \lim_{J\to\infty} \frac{J}{N}\frac{1}{J}\sum_{j=1}^{J}\sigma_{\varepsilon j}^2$. It contains the variance in the specification error a_j, as well as sampling error in estimating the $\overline{Y}_j$s. The second term is an estimate of $\lim_{J\to\infty}\frac{J}{N}\frac{1}{J}\sum_{j=1}^{J}\sigma_{\varepsilon j}^2$, the average sampling variance.[11]

This implies that the interval

$$\left(\widehat{\beta} - 1.96\sqrt{\widehat{V\left(\widehat{\beta}\right)} + 2\widehat{\sigma}_a^2}, \widehat{\beta} + 1.96\sqrt{\widehat{V\left(\widehat{\beta}\right)} + 2\widehat{\sigma}_a^2}\right) \tag{13}$$

will contain $E\left[Y_1 - Y_0 | X = x_k\right]$ with approximately 0.95 probability. The interpretation of this confidence interval is similar to conventional confidence intervals, except that here, the parameter $E\left[Y_1 - Y_0 | X = x_k\right]$ is itself random, due to the randomness of the specification errors. Thus, the correct statement of inference is that the interval contains $E\left[Y_1 - Y_0 | X = x_k\right]$ about 95 percent of the time in repeated draws of both ε_{ij} and the (random) specification errors a_{1k} and a_{0k}.[12]

The interval in (13) strictly contains the usual confidence interval, and therefore leads to more conservative inferences. A wider interval is an intuitive result, since uncertainty regarding the extrapolation errors should yield less precise inferences. Another intuitive aspect of the interval in

[10] It may appear that the homoskedasticity and normality of a_{1j} and a_{0j} is restrictive, but it is important to remember that it is less restrictive than assuming that there is no specification error at all (i.e. $\sigma_a^2 = 0$).

[11] Under heteroskedasticity of ε_{ij} across the J groups, a consistent estimator is given by

$$\widehat{\sigma}_a^2 \equiv \frac{1}{\sum_{j=1}^{J}\frac{n_j}{\widehat{\sigma}_{\varepsilon j}^2}}\sum_{j=1}^{J}\frac{n_j}{\widehat{\sigma}_{\varepsilon j}^2}\left(\overline{Y}_j - W_j\widehat{\theta}\right)^2 - \frac{1}{\sum_{j=1}^{J}\frac{n_j}{\widehat{\sigma}_{\varepsilon j}^2}}\sum_{j=1}^{J}\frac{n_j}{\widehat{\sigma}_{\varepsilon j}^2}\left(\frac{\widehat{\sigma}_{\varepsilon j}^2}{n_j}\right)$$

where $\widehat{\sigma}_{\varepsilon j}^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$.

[12] (13) has been called an "Empirical Bayes"' Confidence Interval. See Morris (1983).

(13) is that it collapses to the conventional one when the chosen parametric form is exactly correct and σ_a^2 is known to be zero.

There is a close connection between $\widehat{\sigma}_a^2$ and the goodness-of-fit statistic G. Consider the case of a constant sampling error variance σ_ε^2 across all j cells. In this case, an alternative consistent estimator for σ_a^2 could be given by

$$\widetilde{\sigma}_a^2 \equiv \left(\frac{J}{J-K}\right)\frac{1}{N}\sum_{j=1}^{J} n_j\left(\overline{Y}_j - W_j\widehat{\theta}\right)^2 - \frac{J}{N}\frac{1}{N-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(Y_{ij}-\overline{Y}_j\right)^2$$

The probability limits of the first and second terms are $\sigma_a^2 + \lim_{J\to\infty}\frac{J}{N}\sigma_\varepsilon^2$ and $\lim_{J\to\infty}\frac{J}{N}\sigma_\varepsilon^2$, respectively. It is also true that $\widetilde{\sigma}_a^2 = (G-1)\frac{ESS_{UR}}{N-J}\frac{J}{N}$. Thus, the more that G exceeds 1 – evidence that the parametric approximation is too restrictive – the wider the confidence interval (13). Obtaining a negative value for $\widetilde{\sigma}_a^2$ simply implies that a goodness-of-fit statistic would be less than 1.

Finally, we draw attention to a technical point that leads to two complications. First, under conventional asymptotics, (11) only holds when $\sigma_a^2 > 0$. When $\sigma_a^2 = 0$, $\sqrt{J}\left(\widehat{\beta} - E\left[Y_1 - Y_0 | X = x_k\right]\right)$ converges in distribution to $N\left(0, V_C\right)$, (where $V_C = plim(\hat{V}_C)$). But $J\widehat{V\left(\widehat{\beta}\right)} + 2J\widehat{\sigma}_a^2$ does not converge to V_C: the first term converges to V_C, but the second term does not vanish. Secondly, under conventional asymptotics, even when $\sigma_a^2 > 0$, $\widehat{\beta} - E\left[Y_1 - Y_0 | X = x_k\right]$ converges in distribution to $N\left(0, 2\sigma_a^2\right)$, because the variance in the estimator of β_0 vanishes as the number of cells increases. Thus, with any fixed sample, the usual asymptotic approximation leads to an unintuitive result that the variance is $\frac{V_C}{J}$ when $\sigma_a^2 = 0$, but jumps to $2\sigma_a^2$ for $\sigma_a^2 > 0$ but arbitrarily small.

The source of these problems is that the estimation error $\hat{\beta} - \beta_0$ is $O_p\left(\frac{1}{\sqrt{J}}\right)$, while the specification error $a_{1k} - a_{0k}$ is $O_p(1)$. In the Appendix, we propose a sequence for the data that allows the variance of X to shrinks as the number of cells J grows. Intuitively, although the increase in the number of cells tends to decrease the variability in the least squares estimator, the shrinking variance in the regressors offsets this tendency, leading to an estimation error that is of the same stochastic order as the specification error. The expression in equation (11) will then be valid whether or not $\sigma_a^2 = 0$, and the asymptotic variance in the overall error $\widehat{\beta} - E\left[Y_1 - Y_0 | X = x_k\right]$ will be continuous at $\sigma_a^2 = 0$.

5 Efficient Estimation

When the specification errors a_{1j} and a_{0j} are assumed to be different, there is an estimator for $E[Y_1 - Y_0|X = 0]$ that is more efficient than the OLS estimator $\widehat{\beta}$. This is because the least squares estimate of β_0 amounts to the difference between the prediction for $E[Y_1|X = 0]$ and the prediction for $E[Y_0|X = 0]$, using data away from the discontinuity threshold. While it is necessary to make such an extrapolation for $E[Y_0|X = 0]$ (since this quantitity is unobservable), information on $E[Y_1|X = 0]$ is available from the sample mean $\overline{Y}_k$. Use of this information can lead to a more efficient estimator of the treatment effect.

Figure 3B illustrates the point. In the figure, $\widehat{\beta}$ estimates the discontinuity in the function represented by the solid lines. In this particular realization of the data, the treatment effect at $X = 0$ is the difference between the solid circle, which is above the parametric function, and the open circle, which is below. The deviation of the open circle from the parametric line is unobservable, but the cell mean provides information on $E[Y_1|X = 0]$. Indeed, as the number of observations per cell tends to infinity, we can estimate $E[Y_1|X = 0]$ perfectly.

More formally, assume that equation (4) is valid, with the normalization that $x_k = 0$. Let $\widehat{\alpha} + \hat{\beta}$ be the least squares estimate of $E[Y_1|X = 0]$ that leaves out the kth cell in the estimation.[13] Now consider the combination estimator

$$\beta^* = \widehat{\beta} + \lambda\left(\overline{Y}_k - \left(\widehat{\alpha} + \hat{\beta}\right)\right) \tag{14}$$

which is the least squares estimator of the discontinuity adjusted by the kth cell mean's deviation from the least squares prediction. The error in the estimator is given by

$$\begin{aligned}\beta^* - E[Y_1 - Y_0|X = 0] &= (\widehat{\beta} - \beta_0) - (a_{1k} - a_{0k}) + \lambda\left(\alpha_0 + \beta_0 + a_{1k} + \overline{\varepsilon}_k - \left(\widehat{\alpha} + \hat{\beta}\right)\right) \\ &= (\widehat{\beta} - \beta_0) - (a_{1k} - a_{0k}) + \lambda\left(a_{1k} + \overline{\varepsilon}_k - \left(\widehat{\alpha} + \hat{\beta} - (\alpha_0 + \beta_0)\right)\right)\end{aligned}$$

which will be centered around zero.

[13] Note that this estimator is asymptotically equivalent to one that includes the kth cell.

The variance of this error is

$$V\left(\widehat{\beta}\right)+2\sigma_a^2+\lambda^2 V(a_{1k}+\overline{\varepsilon}_k-\left(\widehat{\alpha}+\hat{\beta}\right))+2\lambda C\left((\widehat{\beta}-(a_{1k}-a_{0k}),a_{1k}+\overline{\varepsilon}_k-\left(\widehat{\alpha}+\hat{\beta}\right)\right)$$
$$= V\left(\widehat{\beta}\right)+2\sigma_a^2+\lambda^2(\sigma_a^2+\tfrac{\sigma_{\epsilon k}^2}{n_k}+V\left(\widehat{\alpha}+\hat{\beta}\right))-2\lambda(C(\widehat{\beta},\left(\widehat{\alpha}+\hat{\beta}\right))+\sigma_a^2)$$

where the equality holds because the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ do not include data from the kth cell.

The optimal λ can be found by differentiating this variance with respect to λ and solving for the first order condition, yielding:

$$\lambda=\frac{\sigma_a^2+C(\widehat{\beta},\left(\widehat{\alpha}+\hat{\beta}\right))}{\sigma_a^2+V\left(\widehat{\alpha}+\hat{\beta}\right)+\frac{\sigma_{\epsilon k}^2}{n_k}} \tag{15}$$

The intuition behind this formula is illustrated by considering the case in which two separate parametric forms are used to model the function to the left and the right of the discontinuity threshold; that is, when the terms of the parametric function are completely interacted with the treatment dummy variable. Use the identity $C(\widehat{\beta},\hat{\alpha}+\widehat{\beta})=V\left(\hat{\alpha}+\widehat{\beta}\right)-C\left(\hat{\alpha},\hat{\alpha}+\widehat{\beta}\right)$, and note that $C\left(\hat{\alpha},\hat{\alpha}+\widehat{\beta}\right)=0$ here, because in a completely interacted model, only data to the left are used to estimate $\hat{\alpha}$ and only data to the right are used to estimate $\hat{\alpha}+\widehat{\beta}$. The optimal value of λ then becomes:

$$\lambda=\frac{\sigma_a^2+V\left(x_k\widehat{\gamma}+\widehat{\beta}\right)}{\sigma_a^2+V\left(x_k\widehat{\gamma}+\widehat{\beta}\right)+\frac{\sigma_{\epsilon k}^2}{n_k}} \tag{16}$$

When the parametric function is "good", σ_a^2 will be relatively small compared to the cell-level sampling error $\frac{\sigma_{\varepsilon k}^2}{n_k.}$ λ will thus tend to 0, and the linear combination estimator will be closer to the original parametric estimator $\widehat{\beta}$. On the other hand, if the parametric form is "bad", σ_a^2 will be relatively large. As a result, λ will tend towards 1, and the combination estimator will converge towards $\overline{Y}_k-\hat{\alpha}$, which is the difference between the cell mean and the prediction of $E\left[Y_0|X=x_k\right]$using data on the left side of the discontinuity threshold. The combination estimator thus provides a simple way to optimally combine two alternative estimators of $E\left[Y_1-Y_0|X=0\right]$ – $\widehat{\beta}$ and $\overline{Y}_k-\hat{\alpha}$. Note that the usual OLS estimator that *includes* the kth cell can also be written in the same form as (14), using the recursive residual formula of Brown et al. (1975). The implied weight by the OLS will in general not be equal to the weight given by (14).[14]

Whether or not the model is fully interacted, the optimal λ can be substituted into the expression above to yield the variance of this combination estimator:

$$V(\beta^*) = \left(V\left(\widehat{\beta}\right) + 2\sigma_a^2\right) - \lambda^2\left(\sigma_a^2 + V\left(\hat{\alpha} + \widehat{\beta}\right) + \frac{\sigma_{\varepsilon k}^2}{n_k}\right) \tag{17}$$

Note that the first set of parentheses is the error variance as discussed in the previous section, while the second term is non-negative. Thus, the variance of the combination estimator will be weakly smaller than the variance of the estimator $\widehat{\beta}$.

To make this estimator feasible, it is necessary to obtain sample analogues to the population variances and covariances in either (15) or (16). σ_a^2 can be estimated by $\widehat{\sigma}_a^2$ as defined in the previous section. The estimator for $V\left(\hat{\alpha} + \widehat{\beta}\right)$ is simply the "standard error of the prediction" at $X = 0$, which is a standard option in most statistical packages. $C(\widehat{\beta}, \hat{\alpha} + \widehat{\beta}) = V\left(\widehat{\beta}\right) + C\left(\hat{\alpha}, \widehat{\beta}\right)$ can be estimated using the estimated variance of $\widehat{\beta}$ and covariance between $\widehat{\beta}$ and (as long as the threshold is normalized to be zero) the estimated intercept $\hat{\alpha}$; these quantities are usually computed in most statistical packages. Finally, the usual variance estimator of $\overline{Y}_k$ can be used as the sample analogue to $\frac{\sigma_{\varepsilon k}^2}{n_k}$. Together, these quantities imply an estimator $\hat{\lambda}$, which can be used to construct $\hat{\beta}^*$, a feasible version of β^*.

In the Appendix, we provide conditions under which

$$\frac{\widehat{\beta^*} - E\left[Y_1 - Y_0 | X = x_k\right]}{\sqrt{\widehat{V\left(\widehat{\beta^*}\right)}}} \xrightarrow{d} N(0,1)$$

where $\widehat{V\left(\widehat{\beta^*}\right)}$ is defined by (17), with population quantities replaced by their sample analogues.

The usual asymptotic arguments lead to the same complications described in the previous section. Therefore, we continue to adopt the “shrinking variance” sequence in computing the asymptotic distribution, and providing a consistent variance estimator. In addition, as shown in the Appendix, in order to consistently estimate $\sigma_{\epsilon k}^2$, while maintaining that $\frac{\sigma_{\epsilon k}^2}{n_k}$ has the same order as $\hat{\beta}$, it is

[14] Using the recursive residual formula, the OLS coefficient using all observations can be written as

$$\hat{\theta} = \hat{\theta}_{-k} + \left(W'W\right)^{-1} W_k' \left(\overline{Y}_k - \hat{\theta}_{-k}\right)$$

where $-k$ denotes leaving out the k th cell, and W_k denotes the k th row of W.

necessary to assume that the number of observations and the variance of ϵ in the kth cell both grow as the number of cells increase. Without increasing the number of observations in the kth cell, one can neither consistently estimate $\hat{\lambda}$, nor the $V\left(\widehat{\beta^*}\right)$. Without further requiring that $\sigma^2_{\epsilon k}$ grows with the number of observations in the cell, the term $\frac{\sigma^2_{\epsilon k}}{n_k}$ will vanish in the expressions for λ and $V\left(\beta^*\right)$.

6 Relation to Bayesian Estimation

There is a close connection to the proposed estimator $\widehat{\beta^*}$ and a Bayesian approach to the problem. Specifically, the confidence intervals proposed above can be interpreted as Bayesian posterior intervals.

For example, note that the (14) can be re-written as

$$\beta^* = \left[\lambda \overline{Y}_k + (1-\lambda)\left(\hat{\alpha} + \widehat{\beta}\right)\right] - \hat{\alpha}$$

The expression in brackets can be viewed as an estimate of $E\left[Y_1|X=0\right]$ – a λ-weighted average of the kth cell mean and the predicted value from the regression – and the term $\hat{\alpha}$ as an estimate of $E\left[Y_0|X=0\right]$.

Consider a simple Bayesian approach to estimating $E\left[Y_1|X=x_k\right] - E\left[Y_0|X=x_k\right]$. A likelihood for the observed data would be specified; for example, $Y_{ik} \sim N\left(E\left[Y_1|X=0\right], \sigma^2\right)$; assume here that σ^2 is known. Now consider a prior distribution for $\left(E\left[Y_1|X=x_k\right], E\left[Y_0|X=x_k\right]\right)$ given by

$$N\left(\left(E_1, E_0\right), \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_0^2 \end{pmatrix}\right).$$

In this simple setup, given the observed data, the posterior distribution for the quantity $E[Y_1 \mid X = x_k]$ would be given by

$$N\left(\lambda \overline{Y}_k + (1-\lambda) E_1, (1-\lambda)\sigma_1^2\right)$$

with $\lambda = \sigma_1^2 / \left(\frac{\sigma^2}{n_k} + \sigma_1^2\right)$. Since at $X=0$, there are no data for the outcome in the untreated regime, the posterior for $E\left[Y_0|X=x_k\right]$ is the same the prior, $N\left(E_0, \sigma_0^2\right)$. With some re-arrangement, the resulting posterior distribution for $E\left[Y_1 - Y_0|X=0\right]$ is

$$N(\left[\lambda\overline{Y}_k + (1-\lambda)\,E_1\right] - E_0, \sigma_1^2 + \sigma_0^2 - \lambda^2(\frac{\sigma^2}{n_k} + \sigma_1^2))$$

Note that under an uninformative (diffuse) prior on $E\left[Y_0|X=0\right]$, the posterior for the treatment effect will also be uninformative. In the case where only data on the kth cell are provided, this is intuitive: without any outside information, one should not be able to provide an informative estimate of the treatment effect.

What are reasonable choices for the components of the prior distribution E_1 , E_0 , σ_1^2 , and σ_0^2 ? One possibility is to use the data away from the discontinuity threshold to generate values for these parameters. For example, $\hat{\alpha} + \widehat{\beta}$, the predicted value of $E\left[Y_1|X=0\right]$ using all data to the right of the kth cell in a parametric regression could be viewed as a reasonable value for E_1. The variance of that prediction, $\widehat{V\left(\hat{\alpha} + \widehat{\beta}\right)} + \hat{\sigma}_a^2$, is a reasonable value for σ_1^2. Similarly, a regression using all data to the left of the k th cell could generate $\hat{\alpha}$ and $\widehat{V\left(\hat{\alpha}\right)} + \hat{\sigma}_a^2$, which could be used as values for E_0 and σ_0^2, yielding the prior distribution for $E\left[Y_0 \mid X = x_k\right]$. Using these values – and substituting $\hat{\sigma}_\epsilon^2$ for σ^2– yields a posterior distribution for $E\left[Y_1 - Y_0|X=0\right]$ given by $N\left(\widehat{\beta^*}, \widehat{V\left(\widehat{\beta^*}\right)}\right)$.[15] It is important to note that a hierarchical Bayesian approach could be used for this problem. Rather than choosing values E_1, E_0, σ_1^2, and σ_0^2, a prior distribution could be specified for the hyperparameters of the model α_0, β_0, γ_0, σ_a^2, and $\sigma_{\epsilon j}^2$.

7 Summary

This paper draws attention to functional form issues in the estimation of regression discontinuity designs when the index variable determining treatment, X, has discrete support. In the discrete case, the conditions for non-parametric or semi-parametric methods are not satisfied; indeed, the treatment effect is not non-parametrically identified. Our goal is to formally incorporate uncertainty in the necessary parametric modeling of the underlying RD function.

[15] Here we are referring to the combination estimator for the model that completely interacts the treatment indicator with the polynomial. This notion of improving upon the estimate for the k th cell, by using information from other cells, is what underlies the parametric Empirical Bayes approach. Indeed, the estimator $\left[\hat{\lambda}\overline{Y}_k + \left(1-\hat{\lambda}\right)\left(\hat{\alpha} + \widehat{\beta}\right)\right]$ of $E\left[Y_1|X=x_k\right]$ is a type of "shrinkage"/Stein estimator (see Morris (1983)). Thus, the confidence intervals provided here could also be viewed as Empirical Bayes confidence intervals.

We have proposed a procedure for inference that explicitly acknowledges errors in whatever parametric functional form is chosen. Instead of assuming that the chosen functional form "correctly" describes the underlying regression function, we model the deviations of the true conditional means from the parametric function as random specification errors with an unknown variance. Viewing these deviations as random errors requires – at a minimum – the use of cluster-consistent standard errors (clustered on the distinct values of X), rather than conventional heteroskedasticity-consistent standard errors. An even more flexible model of the RD counterfactual functions requires further adjustment; the resulting confidence intervals can also be viewed as Bayesian posterior intervals, when the prior distribution is based on data away from the discontinuity threshold.

The inference procedure proposed in this paper can be summarized as follows.

1. Normalize the X variable so the threshold is at 0 , so the intercept in the regression can be interpreted as the estimate of $E\left[Y_0|X=0\right]$. Choose the parametric form for the approximation. Run the regression on the micro-data, computing both heteroskedasticity- and cluster-consistent (clustering on the individual values of X) standard errors.

2. Consider whether or not the counterfactual functions can be modeled so that specification errors in $E\left[Y_1|X=x_k\right]$ and $E\left[Y_0|X=x_k\right]$ are the same. If so, then the clustered standard errors can be used for inference.

3. If not, collapse the data to the cell level, retaining information on the means, variances, and number of observations in each cell. Run the (cell size-weighted) regression using the cell-data.[16] Use mean squared error from the regression and cell variances to compute $\widehat{\sigma}_a^2$ as in (12). Adjust the sampling variance by $2\widehat{\sigma}_a^2$ according to (13).

4. If a more efficient estimator is desired, use the estimated variances and covariances of the discontinuity coefficients and intercept, as well as the k th cell variance, compute $\hat{\lambda}$,and use this estimator for computing $\widehat{\beta^*}$ and $\widehat{V\left(\widehat{\beta^*}\right)}$.

Although our proposed procedure allows for specification error, there remains the issue of how to choose the functional form for the systematic part of the functional form (e.g., the order of the

[16] At this point, it is useful to verify that the point estimate is identical to that computed in step 1, and that the heteroskedasticity-consistent standard error is identical to the cluster-consistent standard error in step 1 (except for a possible finite-sample correction factor).

polynomial in X). Nevertheless, we believe our approach is better than than simply assuming the parametric form is correct. Moreover, our proposed procedures can be easily implemented using the variances and covariances provided by regression routines in standard statistical packages.

References

Angrist, Joshua and Alan Krueger, "Empirical Strategies in Labor Economics," in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3A, Amsterdam: North Holland, 1999, pp. 1278–1284.

_ **and Victor Lavy**, "Using Maimondies' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 1998, *114*, 533–575.

Brown, R.L., J. Durbin, and J.M. Evans, "Techniques for testing for the constancy of regression relationships over time (with discussion)," *Journal of the Royal Statistical Society B*, 1975, *37*, 149–192.

Card, David and Lara Shore-Sheppard, "Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low Income Children," *Review of Economics and Statistics*, 2004, *86*, 752–766.

Chamberlain, Gary, "Quantile Regression, Censoring, and the Structure of Wages," in C. A. Sims, ed., *Advances in Econometrics, Sixth World Congress*, Vol. 1, Cambridge: Cambridge University Press, 1994.

DiNardo, John and David S. Lee, "Economic Impacts of New Unionization on Private Sector Employers: 1984-2001," *Quarterly Journal of Economics*, 2004, *119*, 1383–1442.

Hahn, Jinyang, Wilbert van der Klaauw, and Petra Todd, "Identification and Estimation of Treatment Effects with a Regression–Discontinuity Design," *Econometrica*, January 2001, *69* (1), 201–209.

Kane, Thomas J., "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going," Working Paper, National Bureau of Economic Research May 2003.

Lee, David S., "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, Forthcoming 2006.

Morris, Carl N., "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 1983, *78*, 47–55.

Moulton, Brent, "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 1990, *57*, 334–338.

Porter, Jack, "Estimation in the Regression Discontinuity Model," Working Paper, Harvard University, Cambridge, MA May 2003.

Shore-Sheppard, Lara, "The Precision of Instrumental Variables Estimates with Grouped Data," Working Paper 374, Industrial Relations Section, Princeton University, Princeton 1996.

Thistlethwaite, D. and Donald Campbell, "Regression–Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 1960, *51*, 309–317.

White, Halbert, "Using Least Squares to Approximate Unknown Regression Functions," *International Economic Review*, 1980, *21*, 149–170.

Appendix

A Proofs

Notation

Consider the regression in matrix form

$$y = Db + X\gamma + e$$

where y $(J \times 1)$ is a vector of cell means for the outcome, the two columns of D $(J \times 2)$ are the intercept and treatment indicator variable, the columns of X $(J \times K)$ are the K polynomial terms in the treatment-determining covariate, and each element of e $(J \times 1)$ is the composite error term $a_j + \bar{\epsilon}_j$. b and γ are the corresponding coefficient vectors. The proofs below are for an unweighted least squares estimate, but they also hold for weighted (by the number of observations per cell) least squares estimates, by first pre-multiplying the regression equation by the square root of an appropriate weighting matrix. Let y_j, D_j, X_j, e_j be the jth row of the corresponding matrices (vectors).

Assumptions

The main assumption is that X has a shrinking variance – after partialling out the intercept and the teratment dummy – as the number of cells increases. That is, we assume that $X \equiv E\left[X^*|D\right] + \frac{1}{\sqrt{J}}\left(X^* - E\left[X^*|D\right]\right)$, where X^* is a $J \times K$ random matrix. For the proofs below, note that this definition is equivalent to $X \equiv DE\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right] + \frac{1}{\sqrt{J}}(X^* - DE\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right])$, where X_j^* is the jth row of X^*. By adopting this sequence, the estimated discontinuity – which amounts to the difference between two linear forecasts at the discontinuity threshold – will not become more precise as J increases. Instead the discontinuity estimator will converge to a normal distribution with finite variance.

Further assume that $E\left[\left(X_j^* - D_jE\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]\right)'\left(X_j^* - D_jE\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]\right)\right] = C$, a positive definite matrix, and that $E\left[D_j'D_je_j^2\right]$, $E\left[D_j'X_j^*e_j^2\right]$, and $E\left[X_j^{*\prime}X_j^*e_j^2\right]$ are finite matrices.

Asymptotic Distribution of $\hat{b}$ as $J \to \infty$

It can be shown that the least squares estimator for b can be written as

$$\hat{b} = \left(D'D\right)^{-1}D'y - \left(D'D\right)^{-1}D'X\left(X'MX\right)^{-1}X'My$$

where $M \equiv I - D\left(D'D\right)^{-1}D'$. It follows that

$$\hat{b} - b = \left(D'D\right)^{-1}D'e - \left(D'D\right)^{-1}D'X \cdot \left(X'MX\right)^{-1}X'Me \tag{18}$$

The first term is $o_p(1)$. $(D'D)^{-1} D'X \xrightarrow{p} E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right]$. $X'MX \xrightarrow{p} C$, because

$$\begin{aligned}
X'MX = &\frac{1}{J}(X^* - DE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])'(X^* - DE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right]) \\
&- ((D'D)^{-1} D'X^* - E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])'\frac{1}{J}(D'X^* - D'DE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right]) \\
&- \frac{1}{J}(X^{*\prime}D - E\left[X_j^{*\prime}D_j\right] E\left[D_j'D_j\right]^{-1} D'D)((D'D)^{-1} D'X^* - E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right]) \\
&((D'D)^{-1} D'X^* - E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])'\frac{1}{J}D'D((D'D)^{-1} D'X^* - E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])
\end{aligned}$$

where the first line converges to C, and the second, third, and fourth lines are $o_p(1)$.

Finally, we have

$$\begin{aligned}
X'Me &= \frac{1}{\sqrt{J}}(X^* - DE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])'e - ((D'D)^{-1} D'X^* - E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])'\frac{1}{\sqrt{J}}D'e \\
&= \frac{1}{\sqrt{J}}(X^* - DE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])'e + o_p(1)
\end{aligned}$$

Thus, we have

$$\hat{b} - b \xrightarrow{d} N\left(0, E\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right] C\Omega C E\left[D_j'X_j^*\right]' E\left[D_j'D_j\right]^{-1}\right)$$

where $\Omega \equiv E[(X_j^* - D_jE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])' (X_j^* - D_jE\left[D_j'D_j\right]^{-1} E\left[D_j'X_j^*\right])e_j^2]$.

Proof of Consistency of $\widetilde{V\left(\hat{b}\right)}$ (Variance Estimator using True b)

The expression in (18) can be used to construct a natural consistent variance estimator assuming a known b. Using (18), consider

$$\begin{aligned}
\widetilde{V\left(\hat{b}\right)} \equiv\; & (D'D)^{-1} (\sum_{j=1}^{J} D_j'D_je_j^2) (D'D)^{-1} \qquad (19) \\
+\; & B(\sum_{j=1}^{J} \left(X_j - D_j(D'D)^{-1}D'X\right)' D_je_j^2) (D'D)^{-1} \\
+\; & (D'D)^{-1} (\sum_{j=1}^{J} D_j' \left(X_j - D_j(D'D)^{-1}D'X\right) e_j^2)B' \\
+\; & B(\sum_{j=1}^{J}(X_j - D_j (D'D)^{-1} D'X)'(X_j - D_j (D'D)^{-1} D'X)e_j^2)B'
\end{aligned}$$

where $B \equiv -(D'D)^{-1} D'X (X'MX)^{-1}$. We first show that this is a consistent estimator for the variance given above, and then show that it is numerically identical to the conventional least-squares clustered variance estimator (with known b).

The first three terms in (19) will be shown to be $o_p(1)$, and the final term will converge to the desired asymptotic variance. The first term is $o_p(1)$. The second term in (19) can be equivalently written as

$$
\begin{aligned}
& B(\sum_{j=1}^{J}\left(X_j - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right]\right)' D_j e_j^2)\left(D'D\right)^{-1} \\
+ \quad & B\left(E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right] - (D'D)^{-1} D'X\right)' (\sum_{j=1}^{J} D_j' D_j e_j^2)\left(D'D\right)^{-1} \\
= \quad & B(\sum_{j=1}^{J}\left(X_j - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right]\right)' D_j e_j^2)\left(D'D\right)^{-1} + o_p(1) \\
= \quad & B(\frac{1}{\sqrt{J}}\sum_{j=1}^{J}\left(X_j^* - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right]\right)' D_j e_j^2)\left(D'D\right)^{-1} + o_p(1) \\
= \quad & o_p(1) + o_p(1)
\end{aligned}
$$

where the first equality follows because $\left(D'D\right)^{-1} D'X$ is consistent for $E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right]$ and $X'MX \xrightarrow{p} C$, which implies that B is $O_p(1)$, the second equality follows by the definition of X_j, and the third equality follows because $(D'D)^{-1}$is $O_p\left(\frac{1}{J}\right)$. The third term in (19) is similarly $o_p(1)$.

The fourth term in (19) can be re-written as

$$
\begin{aligned}
& B(\textstyle\sum_{j=1}^{J}(X_j - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right])'(X_j - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right])e_j^2)B' \\
+ \quad & B\left(E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right] - \left(D'D\right)^{-1} D'X\right)' (\textstyle\sum_{j=1}^{J} D_j' D_j e_j^2) \cdot \\
& \left(E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right] - \left(D'D\right)^{-1} D'X\right) B' \\
+ \quad & B\left(E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right] - \left(D'D\right)^{-1} D'X\right)' \cdot \\
& (\textstyle\sum_{j=1}^{J} D_j'(X_j - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right])e_j^2)B' \\
+ \quad & B(\textstyle\sum_{j=1}^{J}(X_j - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right])' D_j e_j^2) \cdot \\
& \left(E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right] - \left(D'D\right)^{-1} D'X\right) B'
\end{aligned}
$$

which is equal to

$$
\begin{aligned}
& B(\tfrac{1}{J}\textstyle\sum_{j=1}^{J}(X_j^* - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right])'(X_j^* - D_j E\left[D_j' D_j\right]^{-1} E\left[D_j' X_j^*\right])e_j^2)B' \\
+ \quad & O_p(1) \cdot O_p(\tfrac{1}{J}) \cdot O_p(J) \cdot O_p(\tfrac{1}{J}) \cdot O_p(1) \\
+ \quad & O_p(1) \cdot O_p(\tfrac{1}{J}) \cdot O_p(\sqrt{J}) \cdot O_p(1) \\
+ \quad & O_p(1) \cdot O_p(\sqrt{J}) \cdot O_p(\tfrac{1}{J}) \cdot O_p(1)
\end{aligned}
$$

because $E\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]-(D'D)^{-1}D'X$ is $O_p\left(\frac{1}{J}\right)$, and $(\sum_{j=1}^{J}D_j'(X_j-D_jE\left[D_j'D_j\right]^{-1}$ $E\left[D_j'X_j^*\right])e_j^2)$ is $O_p\left(\sqrt{J}\right)$, which can be seen by noting that X is, by definition, shrinking towards the predicted means. The first line also follows by the definition of X. Thus the fourth term in (19) converges in probability to $E\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]C\Omega CE\left[D_j'X_j^*\right]'E\left[D_j'D_j\right]^{-1}$.

Next, (19) can be shown to be numerically identical to the conventional least squares clustered variance estimator (with β known), after some re-arrangement of terms. Specifically, after expanding the middle two terms, (19) becomes

$$\begin{aligned}
&(D'D)^{-1}(\sum_{j=1}^{J}D_j'D_je_j^2)\left(D'D\right)^{-1} \quad -BX'D\left(D'D\right)^{-1}(\sum_{j=1}^{J}D_j'D_je_j^2)\left(D'D\right)^{-1}\\
&-\left(D'D\right)^{-1}(\sum_{j=1}^{J}D_j'D_je_j^2)\left(D'D\right)^{-1}D'XB'\\
&+B(\sum_{j=1}^{J}X_j'D_je_j^2)\left(D'D\right)^{-1} \;+\; \left(D'D\right)^{-1}(\sum_{j=1}^{J}D_j'X_je_j^2)B'\\
&+B(\sum_{j=1}^{J}(X_j-D_j\left(D'D\right)^{-1}D'X)'(X_j-D_j\left(D'D\right)^{-1}D'X)e_j^2)B'
\end{aligned}$$

After expanding the last term and collecting terms with $\sum_{j=1}^{J}D_j'D_je_j^2$, $\sum_{j=1}^{J}X_j'D_je_j^2$, $\sum_{j=1}^{J}D_j'X_je_j^2$, and $\sum_{j=1}^{J}X_j'X_je_j^2$, we obtain

$$A(\sum_{j=1}^{J}D_j'D_je_j^2)A \;+B(\sum_{j=1}^{J}X_j'D_je_j^2)A+A(\sum_{j=1}^{J}D_j'X_je_j^2)B'+B(\sum_{j=1}^{J}X_j'X_je_j^2)B'$$

where $A\equiv(D'D)^{-1}-BX'D\left(D'D\right)^{-1}=(D'D)^{-1}+(D'D)^{-1}D'X\left(X'MX\right)^{-1}X'D\left(D'D\right)^{-1}$, and $B\equiv-\left(D'D\right)^{-1}D'X\left(X'MX\right)^{-1}$. This is exactly the expression that would be obtained by using the partitioned inverse formula for the conventional least squares clustered variance estimator (with β known) for $\hat{\beta}$.

Proof that $\widehat{V\left(\hat{b}\right)}-\widetilde{V\left(\hat{b}\right)}$ is $o_p(1)$

Let $\widehat{V\left(\hat{b}\right)}$ be the conventional clustered variance estimator (with unknown β); it is defined as

$\widetilde{V\left(\hat{b}\right)}$ except after replacing e_j with $\hat{e}_j \equiv Y_j - D_j\hat{b} - X_j\hat{\gamma}$. It follows that

$$
\begin{aligned}
\hat{e}_j &= e_j - D_j\left(\hat{b}-\beta\right) - X_j\left(\hat{\gamma}-\gamma\right) \\
&= e_j - D_j\left(D'D\right)^{-1}D'e + D_j\left(D'D\right)^{-1}D'X\left(\hat{\gamma}-\gamma\right) - X_j\left(\hat{\gamma}-\gamma\right) \\
&= e_j - D_j\left(D'D\right)^{-1}D'e + D_j\left(\left(D'D\right)^{-1}D'X - E\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]\right)\left(\hat{\gamma}-\gamma\right) \\
&\quad -\left(X_j - D_jE\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]\right)\left(\hat{\gamma}-\gamma\right) \\
&= e_j - D_j\cdot O_p\left(\frac{1}{\sqrt{J}}\right) + D_j\cdot O_p\left(\frac{1}{J}\right)\cdot O_p\left(1\right) \\
&\quad -\frac{1}{\sqrt{J}}\left(X_j^* - D_jE\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]\right)\cdot O_p\left(1\right)
\end{aligned}
$$

The second and third equalities follow from re-arranging terms. The final equality follows from noting that $\left(D'D\right)^{-1}D'X - E\left[D_j'D_j\right]^{-1}E\left[D_j'X_j^*\right]$is $O_p\left(\frac{1}{J}\right)$ and $\left(\hat{\gamma}-\gamma\right)$ is $O_p\left(1\right)$, as shown in the proof of asymptotic normality.

Squaring the above residual yields

$$
\begin{aligned}
\hat{e}_j^2 - e_j^2 &= e_jD_j\left(O_p\left(\frac{1}{J}\right)+O_p\left(\frac{1}{\sqrt{J}}\right)-O_p\left(\frac{1}{\sqrt{J}}\right)\right) - e_jX_j^*O_p\left(\frac{1}{\sqrt{J}}\right) \\
&+ \left(O_p\left(\frac{1}{J}\right)+O_p\left(\frac{1}{\sqrt{J}}\right)-O_p\left(\frac{1}{\sqrt{J}}\right)\right)'D_j'D_j\left(O_p\left(\frac{1}{J}\right)+O_p\left(\frac{1}{\sqrt{J}}\right)-O_p\left(\frac{1}{\sqrt{J}}\right)\right) \\
&- \left(O_p\left(\frac{1}{J}\right)+O_p\left(\frac{1}{\sqrt{J}}\right)-O_p\left(\frac{1}{\sqrt{J}}\right)\right)'D_j'X_j^*O_p\left(\frac{1}{\sqrt{J}}\right) \\
&+ O_p\left(\frac{1}{\sqrt{J}}\right)'X_j^{*\prime}X_j^*O_p\left(\frac{1}{\sqrt{J}}\right)
\end{aligned} \tag{20}
$$

Note that each of the above terms is a summation of scalars. To see that $\widehat{V\left(\hat{b}\right)} - \widetilde{V\left(\hat{b}\right)}$ is $o_p\left(1\right)$, substitute each of these scalars for "e_j^2" in (19). The first three terms will be $o_p(1)$ as argued in the proof for the consistency of $\widetilde{V\left(\hat{b}\right)}$. In addition, the fourth term will also be $o_p(1)$ because each of these scalars is a product that includes a $O_p(\cdot)$ term in (20).

Proof that $\hat{\sigma}_a^2 \xrightarrow{p} \sigma_a^2$.

First, note the definition$\hat{\sigma}_a^2 \equiv \frac{1}{J}\sum_j \hat{e}_j^2 - \frac{1}{Jn}\sum_j \frac{1}{n-1}\sum_i \left(Y_{ij}-y_j\right)^2$. Next, summing over (20), it follows that $\frac{1}{J}\sum_j \hat{e}_j^2 \xrightarrow{p} \frac{1}{J}\sum_j e_j^2$, which converges to $\sigma_a^2 + \frac{\sigma_\epsilon^2}{n}$. Finally, the second term is a consistent estimator for $\frac{\sigma_\epsilon^2}{n}$, as $J \to \infty$.

Proof of Asymptotic Distribution of Shrinkage Estimator

In addition to the assumptions above, normalize so that x_k, the point of the threshold, is zero, and let $b = (\alpha, \beta)$, so that α is the intercept and β is the discontinuity gap. Also assume that $n_k = Jn_k^*$, n_k^* a finite constant, and $\varepsilon_{ik} = \sqrt{n_k}\varepsilon_{ik}^*$, so that $\sigma_{\varepsilon k}^2 = n_k\sigma_{\varepsilon k}^{*2}$, $\sigma_{\varepsilon k}^{*2}$ a finite constant.

We will show that

$$\frac{\widehat{\beta^{*}}-E\left[Y_{1}-Y_{0}\mid X=0\right]}{\sqrt{\widehat{V\left(\widehat{\beta^{*}}\right)}}} \xrightarrow{d} N(0,1)$$

by first showing that $\widehat{\beta^{*}}-E\left[Y_{1}-Y_{0}|X=0\right] \xrightarrow{d} N\left(0,V\left(\beta^{*}\right)\right)$, and then showing that $\widehat{V\left(\widehat{\beta^{*}}\right)}$ is consistent for $V\left(\beta^{*}\right)$.

First, re-write $\frac{1}{\sqrt{J}}\left(\widehat{\beta^{*}}-E\left[Y_{1}-Y_{0}|X=0\right]\right)$ as $\frac{1}{\sqrt{J}}(\widehat{\beta}-E\left[Y_{1}-Y_{0}|X=0\right]+\hat{\lambda}\left(\overline{Y}_{k}-\left(\widehat{\alpha}+\widehat{\beta}\right)\right))$. Define c_J as the vector $\left(\frac{1}{\sqrt{J}}\left(\hat{\beta}-E\left[Y_{1}-Y_{0}|X=0\right]\right),\frac{1}{\sqrt{J}}\left(\overline{Y}_{k}-\left(\hat{\alpha}+\hat{\beta}\right)\right),\hat{\lambda}\right)'$, so that $\frac{1}{\sqrt{J}}(\widehat{\beta^{*}}-E\left[Y_{1}-Y_{0}|X=0\right])=f\left(c_{J}\right)$, noting that $f\left(\cdot\right)$is a continuous function.

We need to show c_J has probability limit $c=(0,0,\lambda)$, and that $\sqrt{J}\left(c_{J}-c\right)$converges in distribution to $N\left(0,V^{*}\right)$. If true, then $\sqrt{J}\left(f\left(c_{J}\right)-f\left(c\right)\right)$ will converge in distribution to $N(0,(1,\lambda,0)'\,V^{*}\cdot(1,\lambda,0))$, by the delta method. The zero in the last element of the gradient vector implies that the resulting asymptotic variance does not include the variance of $\widehat{\lambda}$, or its covariance with any other element of b_J. As a result, it will be true that $\widehat{\beta^{*}}-E\left[Y_{1}-Y_{0}\mid X=0\right]\xrightarrow{d} N\left(0,V\left(\beta^{*}\right)\right)$.

To show $c_{J}\xrightarrow{p} c\equiv(0,0,\lambda)$, recall from above that $\widehat{\beta}-E\left[Y_{1}-Y_{0}|X=0\right]$is $O_{p}\left(1\right)$; multiplying by $\frac{1}{\sqrt{J}}$ yields $o_{p}\left(1\right)$. Similarly, $\overline{Y}_{k}-\left(\widehat{\alpha}+\widehat{\beta}\right)=\left(\widehat{\alpha}-\alpha\right)+\left(\beta-\widehat{\beta}\right)+a_{k}+\overline{\varepsilon}_{k}$is also $O_{p}\left(1\right)$, because $\frac{1}{n_{k}}\sum_{i=1}^{n_{k}}\varepsilon_{ik}=\frac{1}{\sqrt{J}n_{k}^{*}}\sum_{i=1}^{n_{k}}\varepsilon_{ik}^{*}$; multiplying by $\frac{1}{\sqrt{J}}$ yields $o_{p}\left(1\right)$. $\widehat{\lambda}$ is consistent for λ , because the sample analogs to each of its parts are consistent. For example, as shown above, the standard estimators for $C(\widehat{\beta},\widehat{\alpha}+\widehat{\beta})$ and $V\left(\widehat{\beta}\right)$ are consistent, as is $\widehat{\sigma}_{a}^{2}$. Also,

$$\begin{aligned}
\frac{1}{n_{k}^{2}}\sum\nolimits_{i=1}^{n_{k}}\left(Y_{ik}-\overline{Y}_{k}\right)^{2} &= \frac{1}{n_{k}^{2}}\sum\nolimits_{i=1}^{n_{k}}\epsilon_{ik}^{2}+\frac{2}{n_{k}^{2}}\sum\nolimits_{i=1}^{n_{k}}\epsilon_{ik}\left(E\left[Y_{ik}\right]-\overline{Y}_{k}\right)+\frac{1}{n_{k}^{2}}\sum\nolimits_{i=1}^{n_{k}}\left(E\left[Y_{ik}\right]-\overline{Y}_{k}\right)^{2}\\
&= \frac{1}{n_{k}}\sum\nolimits_{i=1}^{n_{k}}\varepsilon_{ik}^{*2}+\frac{\sqrt{n_{k}}}{n_{k}^{2}}\left(E\left[Y_{ik}\right]-\overline{Y}_{k}\right)\sum\nolimits_{i=1}^{n_{k}}\varepsilon_{ik}^{*}+\frac{1}{n_{k}}\left(E\left[Y_{ik}\right]-\overline{Y}_{k}\right)^{2}\\
&= \frac{1}{n_{k}}\sum\nolimits_{i=1}^{n_{k}}\varepsilon_{ik}^{*2}+O\left(\frac{\sqrt{J}}{J^{2}}\right)O_{p}\left(1\right)o_{p}\left(J\right)+O\left(\frac{1}{J}\right)O_{p}\left(1\right)
\end{aligned}$$

where the first and second equalities hold after some re-arrangement, and the third equality holds because $E\left[Y_{ik}\right]-\overline{Y}_{k}$is $O_{p}\left(1\right)$.

To show $\sqrt{J}\left(c_{J}-c\right)\xrightarrow{d} N\left(0,V^{*}\right)$, we decompose the vector as

$$\sqrt{J}\left(c_{J}-c\right)=\left(\begin{array}{c}\widehat{\beta}-\beta\\ \left(\alpha-\widehat{\alpha}\right)+\left(\beta-\widehat{\beta}\right)\\ \sqrt{J}\left(\widehat{\lambda}-\lambda\right)\end{array}\right)+\left(\begin{array}{c}0\\ \overline{\varepsilon}_{k}\\ 0\end{array}\right)+\left(\begin{array}{c}a_{k1}-a_{k0}\\ a_{k1}\\ 0\end{array}\right)$$

The element in the second vector is $\frac{1}{n_{k}}\sum_{i=1}^{n_{k}}\varepsilon_{ik}=\frac{1}{\sqrt{n_{k}^{*}}}\sum_{i=1}^{n_{k}}\varepsilon_{ik}^{*}$, which converges to a normal. The third vector is normal, by assumption. The first two elements converge to a normal as in the proof of the asymptotic normality of $\hat{b}$, as shown above. Finally, $\sqrt{J}\left(\widehat{\lambda}-\lambda\right)$ can also be expressed as a summation in the form of $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} z_j$. $\sqrt{J}\left(\frac{\widehat{\sigma}_a^2+\widehat{C(\widehat{\beta},\widehat{\alpha}+\widehat{\beta})}}{\widehat{\sigma}_a^2+\widehat{V\left(\widehat{\alpha}+\widehat{\beta}\right)}+\frac{\widehat{\sigma}_{\varepsilon k}^2}{n_k}} - \frac{\sigma_a^2+C(\widehat{\beta},\widehat{\alpha}+\widehat{\beta})}{\sigma_a^2+V\left(\widehat{\alpha}+\widehat{\beta}\right)+\frac{\sigma_{\varepsilon k}^2}{n_k}}\right)$ converges in probability to $\sqrt{J}\left(\frac{\widehat{\sigma}_a^2-\sigma_a^2+\widehat{C(\widehat{\beta},\widehat{\alpha}+\widehat{\beta})}-C(\widehat{\beta},\widehat{\alpha}+\widehat{\beta})}{\sigma_a^2+V\left(\widehat{\alpha}+\widehat{\beta}\right)+\frac{\sigma_{\epsilon k}^{*2}}{n_k^*}}\right)$. The numerator can be shown to be a summation in the form of $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} z_j + o_p(1)$. The central limit theorem applies.

We have shown that each of the parts that make up $\widehat{\lambda}$ is consistent. Those same terms are used to construct $\widehat{V\left(\widehat{\beta^*}\right)}$, which is therefore consistent for $V(\beta^*)$.

B Extension to "Fuzzy" Regression Discontinuity Designs

Many interesting applications of the RD research design involve "imperfect compliance": the relation between the treatment of interest is not a deterministic function of X. Instead the conditional expectation of the treatment is a discontinuous function of X. Angrist and Lavy (1998), for example, use discontinuities in the mapping from the number of students in a grade to average class size to identify the effect of class size on test scores. The rule, while not perfectly followed, nevertheless generates a discontinuity in the *expected* class size. A very simple version of this setup consists of two equations:

$$Y_{1ij} = D_j\delta_0 + X_j\gamma_1 + u_{ij}$$

$$Y_{2ij} = Y_{1ij}\beta_0 + X_j\gamma_2 + v_{ij}$$

where (Y_{1ij}, Y_{2ij}) is a pair of observed outcomes for the ith individual in the jth cell, $X_j$and D_j are as previously defined, δ_0 is the discontinuity in Y_1 at $X = 0$, β_0 is the causal effect of Y_1 on Y_2, and (u_{ij}, v_{ij}) is a pair of potentially correlated errors. Correlation between u_{ij} and v_{ij} implies that β_0 cannot be estimated consistently by a simple OLS procedure. β_0 can be estimated, however, by instrumental variables method using D_j as an instrument for Y_{1ij}. The maintained assumptions are that program status D_j has no direct effect on Y_2, controlling for Y_1. Note that the resulting IV estimator is equivalent to estimating two regression discontinuities – for the two outcomes $Y_1$and Y_2– and computing the ratio of the discontinuity gaps.

A natural extension of our framework is to assume that the data generating process for the observed outcomes is

$$Y_{1ij} = D_j\delta_0 + X_j\gamma_1 + a_{1j} + u_{ij}$$

$$Y_{2ij} = Y_{1ij}\beta_0 + X_j\gamma_2 + a_{2j} + v_{ij}$$

where (a_{1j}, a_{2j}) represents an i.i.d. vector of mean zero random specification errors. IV will still yield an asymptotically unbiased estimate of β_0, but the conventional IV sampling errors, as in the "sharp" design, ignore the group structure of the residuals and may overstate the precision of the IV estimator (See Shore-Sheppard (1996) for a discussion of grouped error structures in an IV setting similar to Moulton (1990)). The use of clustered standard errors is again a simple remedy in this situation.

Note that the above specification implicitly assumes the structure of "identical" specification errors in the counterfactual functions, as described in sub-section (4.1). If it is more desirable to assume "independent" errors, as in sub-section (4.2), then it is necessary to account for the variance in the forecast errors $a_{1j}$and a_{2j}. One way to proceed would be to apply the procedure in (4.2), separately for both "outcomes" Y_1 and Y_2. This would give us, for example, least squares estimate $\hat{\beta}$and $\hat{\pi}$ for the parameters $E\left[Y_1^1 - Y_1^0|X=0\right]$ and $E\left[Y_2^1 - Y_2^0|X=0\right]$ (where the superscripts denote potential outcomes). The error in these estimators include both estimation error and the forecast error, as in sub-section (4.2). It is possible to analogously compute the covariance in the estimation error in $\hat{\beta}$and $\hat{\pi}$ as well as the covariance between the specification errors for each outcome. Following an analogous argument as in (4.2), it would then follow that

$$\left(\hat{\beta} - E\left[Y_1^1 - Y_1^0|X=0\right], \hat{\pi} - E\left[Y_2^1 - Y_2^0|X=0\right]\right)' \hat{\Sigma}^{-1} \cdot \left(\hat{\beta} - E\left[Y_1^1 - Y_1^0|X=0\right], \hat{\pi} - E\left[Y_2^1 - Y_2^0|X=0\right]\right)$$

(where $\hat{\Sigma}$ is the corresponding consistent estimator of the variance-covariance matrix for the error vector) converges in distribution to $\chi^2(2)$. One can invert this test statistic to generate, for example, a 95 percent joint confidence set for $E\left[Y_1^1 - Y_1^0|X=0\right]$ and $E\left[Y_2^1 - Y_2^0|X=0\right]$, and from this generate the confidence set for the ratio $E\left[Y_2^1 - Y_2^0|X=0\right] / E\left[Y_1^1 - Y_1^0|X=0\right]$.

Figure 1: Regression Discontinuity, Continuous Covariate

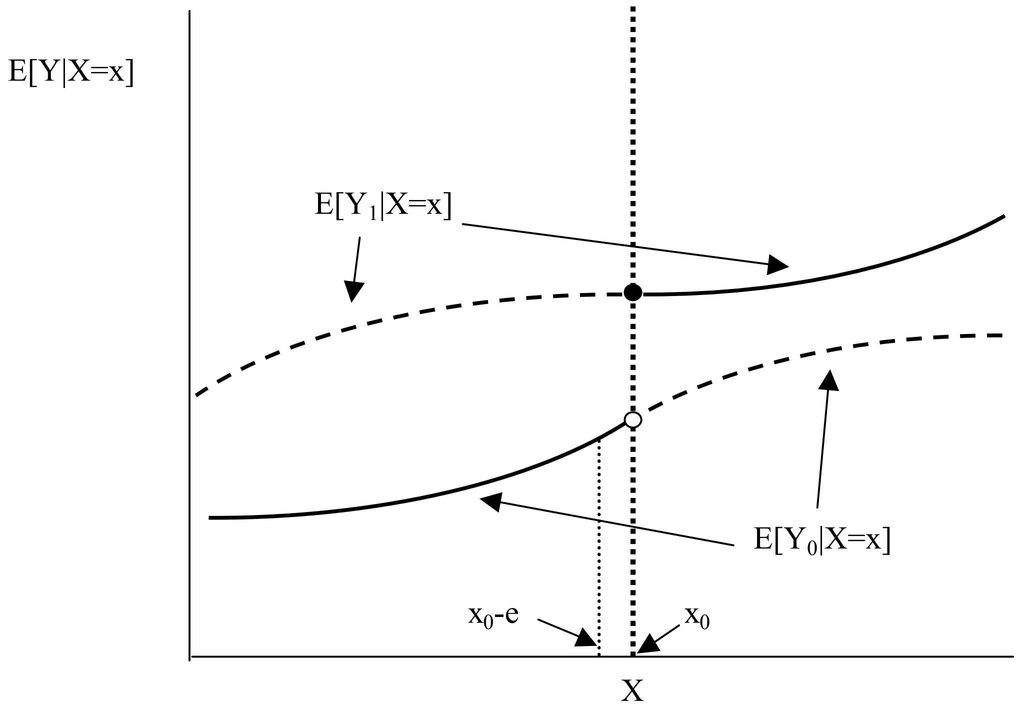

Figure 2: Regression Discontinuity, Discrete Covariate

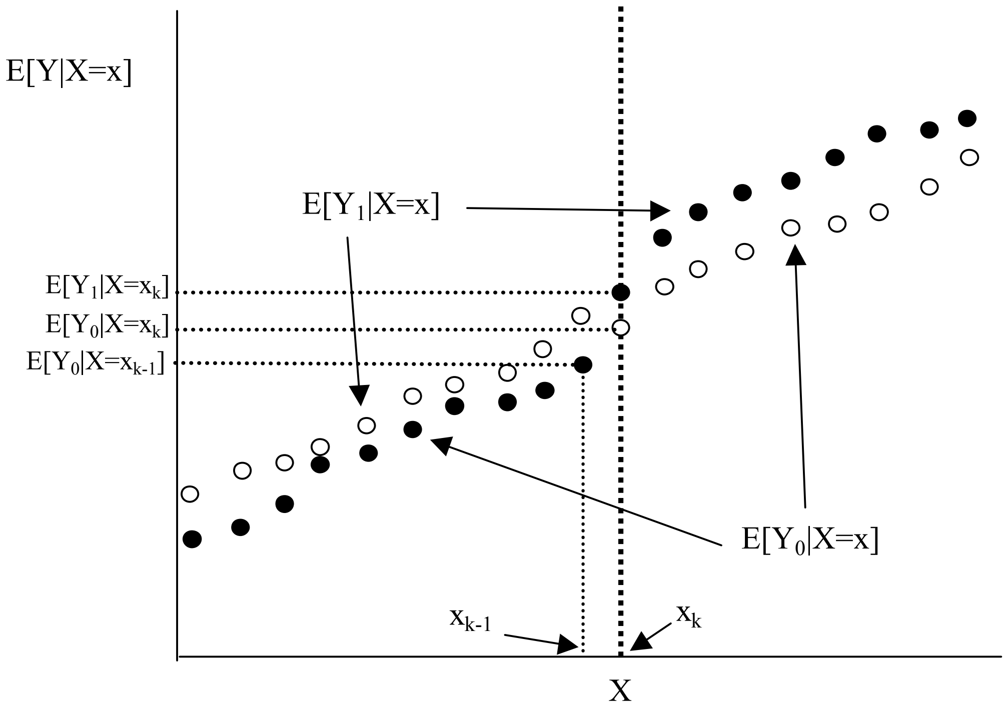

Figure 3A: Counterfactual Specification, Identical Errors

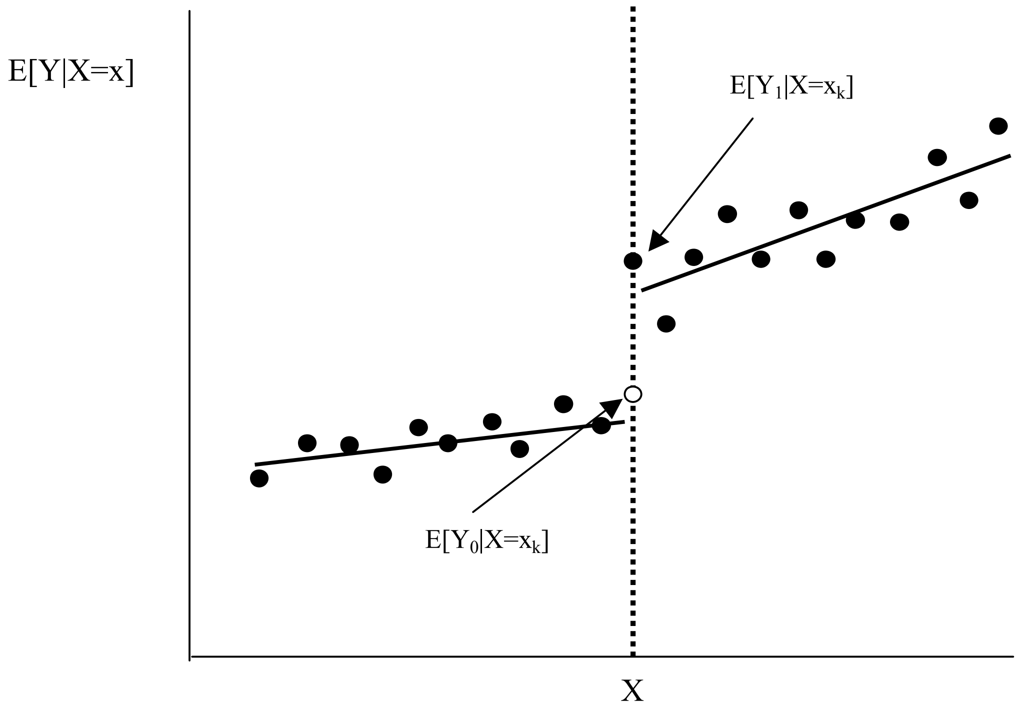

Figure 3B: Counterfactual Specification, Independent Errors