CENTER FOR LABOR ECONOMICS
UNIVERSIY OF CALIFORNIA, BERKELEY
WORKING PAPER NO. 74

Regression Discontinuity Inference with Specification Error

David S. Lee
UC Berkeley and NBER

David Card
UC Berkeley and NBER

June 2004

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. Allowance for specification error in the RD estimation is equivalent to a parametric empirical Bayes procedure.

JEL: C12, C11

*We are grateful to Guido Imbens for many helpful discussions, and to Michael Jensen, James Powell, Kei Hirano, and participants in the 2003 Banff International Research Station Regression Discontinuity Conference for comments and suggestions.

I. 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, Todd, and van der Klauww [2001] and Porter [2003], who describe non-parametric and semi-parametric estimators of regression-discontinuity gaps.

In many applications where the RD idea 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"

[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

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. 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*

(locally) randomly assigned. See Lee [2003] for details.

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 II 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 III describes the proposed inference procedure under a model where specification errors are considered random. Section IV describes a modified procedure under less restrictive assumptions about the specification errors. Section V proposes an alternative, efficient estimator for the treatment effect, and Section VI relates the estimator to Bayes' and Empirical Bayes' approaches. Section VII concludes.

I. The Regression Discontinuity Design with Discrete Support

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 $E[Y_1 - Y_0 \mid X=x_0]$, the local treatment effect at the threshold. As usual in an evaluation problem, Y_1 andY_0 are not simultaneously observed for any individual. Instead, we observe $Y = DY_1 + (1-D) Y_0$.

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

When the support of X is continuous and certain smoothness assumptions are satisfied, $E[Y_1 - Y_0 \mid X=x_0]$ is identified as the discontinuity in the regression function for the *observed outcome* Y at x_0. In particular, if $E[Y_1 \mid X]$ and $E[Y_0 \mid X]$ are both continuous at x_0, then

$$\begin{aligned} E[Y_1 - Y_0 \mid X=x_0] &= E[Y_1 \mid X=x_0] - \lim_{\varepsilon \to 0+} E[Y_0 \mid X = x_0 - \varepsilon] \\ &= E[Y \mid X=x_0] - \lim_{\varepsilon \to 0+} E[Y \mid X = x_0 - \varepsilon] . \end{aligned}$$

This idea is illustrated in Figure 1. The data identifies $E[Y_1 \mid X=x]$ when $x \geq x_0$, and $E[Y_0 \mid X=x]$ when $x<x_0$, as indicated by the solid lines. Because of the discontinuous rule that determines treatment status, the data do not provide information on either the dashed lines, or the counterfactual mean $E[Y_0 \mid X=x_0]$ (the open circle). What the data do yield is $E[Y_0 \mid X = x_0 - \varepsilon]$, which can be an arbitrarily good approximation to $E[Y_0 \mid X = x_0]$, with ε sufficiently small.

This limiting argument, however, does not work when the support of X is discrete, as Figure 2 illustrates. Let the kth value of X, x_k, denote the value of the discontinuity threshold. As before, the counterfactual mean $E[Y_0 \mid X = x_k]$ is unobservable. But now there is a limit to how well it can be approximated. $E[Y \mid X = x_{k-1}]$ – the discrete analogue to $E[Y \mid X = x_0 - \varepsilon]$ – could be a poor approximation, resulting in misleading inferences. For example, in Figure 2, the difference between $E[Y \mid X = x_k]$ and $E[Y \mid X = x_{k-1}]$ substantially over-estimates the true effect $E[Y_1 - Y_0 \mid X=x_k]$. This approximation error is unaffected by sample size, so the asymptotic arguments employed by non-parametric and semi-parametric methods are inapplicable when the discreteness in X is an important feature of the data. The researcher is forced to extrapolate the quantity $E[Y_0 \mid X=x_k]$ using data “away” from the discontinuity threshold. Doing this without choosing a parametric form to approximate $E[Y_0 \mid X=x]$ is impossible.

II. An Alternative Formulation

Many applications of the RD framework are implemented by regressing the outcome Y on a low-order polynomial in the treatment-determining covariate X and the binary treatment indicator D (e.g., Lee, 2003; Dinardo and Lee, 2004; Card and Shore Sheppard, 2002; Kane 2003). Sometimes other covariates are also included. Recognizing that X is discrete, let Y_{ij} represent the outcome for the ith observation with the jth value of X (hereafter, the jth cell) and let Z_{ij} represent a vector of individual-level covariates. The conventional set-up assumes that

(1) $E[\, Y_{ij} \mid X = x_j ,\; Z_{ij} \,] \;=\; Z_{ij}\, \varphi \;+\; h(x_j , \gamma) + D_j\, \beta$,

where φ is a vector of coefficients, $h(x_j, \gamma)$ is some function with coefficients γ, D_j is the treatment status indicator for subgroup j ($D_j=1[x_j \geq x_k]$), and β is the parameter of interest, measuring the discontinuity in the partial regression function for Y at $X=x_k$.[3] In this paper we ignore individual-level covariates in Equation (1). It is straightforward to extend our arguments to include them. Moreover, if the RD design is valid they can be excluded, since

$$E[\, Y_{ij} \mid X = x_j \,] \;=\; E[Z_{ij} \mid X = x_j]\, \varphi \;+\; h(x_j , \gamma) + D_j\, \beta$$

and in a valid design E[Z|X] should be a smooth function of X.[4] Thus E[Y|X] will have the same discontinuity at x_k as the partial regression function.[5]

Conventional inferences based on a model like Equation (1) may be misleading if the functional form of h is mis-specified. Researchers usually address this concern by plotting the mean values of Y against X, and super-imposing the parametric fit from their model. Ideally, the plot confirms that the parametric model provides a good fit for the mean of Y. Our approach

[3] h(.,.) 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.

[4] One way to test the validity of the design is to look for discontinuities in the regression function E[Z|X]. This is analogous to a test for random assignment based on comparisons of the characteristics of the treatment and control groups.

builds on this intuitive procedure by recognizing that any functional form is likely to be mis-specified and computing sampling errors for the estimated discontinuity that are valid not only if the functional form is correct, but also when the parametric model is incorrect (White, 1984; Chamberlain, 1994).

We assume that the cell means for Y, conditional on X, are generated as realizations from the process:

(2) $h(x_j, \gamma) + D_j \beta + a_j, \quad j=1,2,\ldots J,$

where a_j is an i.i.d. specification error with mean 0 and variance σ_a^2. The size of σ_a^2 reflects a researcher's ignorance about the "true" functional form for $E[Y_{ij} \mid X = x_j]$. When σ_a^2 is small, the functional form is approximately correct and conventional inference is appropriate. When σ_a^2 is larger, however, inference should take account of this uncertainly. Equation (2) implies that the data generating process for the observed outcomes is

(3) $Y_{ij} = h(x_j, \gamma) + D_j \beta + a_j + \varepsilon_{ij}, \quad i=1,2,\ldots n_j; \quad j=1,2\ldots J$

where

$$\varepsilon_{ij} = Y_{ij} - E[Y_{ij} \mid X = x_j].$$

Unlike Equation (1), this model has a grouped error structure, reflecting the fact that all the observations with $X = x_j$ share the same realization of the specification error. Assuming that the specification errors are i.i.d., this model can be estimated by standard feasible GLS procedures. For now, however, we focus on the commonly-used alternative of calculating valid "cluster-consistent" sampling errors for the OLS estimator, which is consistent under the i.i.d. specification error assumption.

[5] A more general functional form than Equation (1) would allow both h and β to depend on Z. In this case the discontinuity in $E[Y|X]$ at x_0 is an average of the discontinuities in the partial regression function, with weights given by the marginal distribution of Z at $X=x_0$.

Assuming that h is a low order polynomial (or some other basis function), let $h(x_j,\gamma) = X_j\gamma$, where X_j is the row vector of polynomial terms in X (including an intercept). The model can then be rewritten as

(3') $$Y_{ij} = X_j \gamma + D_j \beta + a_j + \varepsilon_{ij} = W_j\theta + a_j + \varepsilon_{ij} ,$$

where $W_j = (X_j, D_j)$.

The consistent estimator for the variance of the OLS coefficients is

(4) $$\widehat{V(\widehat{\theta})} = \left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} W_j'W_j\right)^{-1}\left(\sum_{j=1}^{J}\left(\sum_{i=1}^{n_j} W_j'\widehat{u}_{ij}\right)\left(\sum_{i=1}^{n_j} W_j\widehat{u}_{ij}\right)\right)\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} W_j'W_j\right)^{-1}$$

where $\widehat{u}_{ij} = Y_{ij} - W_j\widehat{\theta}$. The computation of these standard errors is available as an option in today's statistical software packages.

Using this variance estimator instead of the conventional formula is intuitive. There are only J values of W that can be used to identify the parametric function. So the precision of the estimates should depend not on n_j , the number of observations for each cell value of W, but rather on J , the number of cells. To see this formally, define $\widehat{a}_j \equiv \overline{Y}_j - W\widehat{\theta}$ and $\widehat{\varepsilon}_{ij} = Y_{ij} - \overline{Y}_j$, where $\overline{Y}_j \equiv \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$. Consider the simple case where $n_j = n_0$ for all j . Equation (4) then becomes

(4’) $$\widehat{V(\widehat{\theta})} = \left(\frac{1}{J}\sum_{j=1}^{J} W_j'W_j\right)^{-1}\left(\frac{1}{J^2}\sum_{j=1}^{J} W_j'W_j\widehat{a}_j^2\right)\left(\frac{1}{J}\sum_{j=1}^{J} W_j'W_j\right)^{-1}$$

which depends on J . Thus, the formula for the clustered standard error is numerically equivalent to the heteroskedasticity-consistent standard error of the regression that uses the sample means (i.e. at the cell level) instead of the underlying micro-data. Note also, that infinite

n_0 would not shrink this variance estimator to zero.[6]

By contrast, the conventional standard error formula that ignores the group-error structure would yield

$$\widetilde{V\left(\hat{\theta}\right)} = \left(\frac{1}{J}\sum_{j=1}^{J} W_j' W_j\right)^{-1} \left(\frac{1}{J^2}\sum_{j=1}^{J} W_j' W_j \left(\frac{\hat{a}_j^2}{n_0} + \frac{1}{n_0^2}\sum_{i=1}^{n_0} \hat{\varepsilon}_{ij}^2\right)\right)\left(\frac{1}{J}\sum_{j=1}^{J} W_j' W_j\right)^{-1}$$

This latter variance estimator will tend to understate the true variability in $\hat{\theta}$.[7] Furthermore, it will tend to zero as n_0 increases, even with J fixed -- an unintuitive result, since the parameter θ is identified by variation across J cells.

Extension to Instrumental Variables Applications

Many interesting applications of the RD research design arise in situations where a program-induced discontinuity in Y_1 is used to identify the causal effect of Y_1 on some other outcome Y_2. Angrist and Lavy (1999), 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. A very simple version of this setup consists of two equations:

$$Y_{1ij} = h(x_j, \gamma) + D_j \beta + u_{ij}$$

[6] When cells have unequal numbers of observations, the formula becomes, $\widehat{V\left(\hat{\theta}\right)} = \left(\frac{1}{J}\sum_{j=1}^{J} n_j W_j' W_j\right)^{-1} \left(\frac{1}{J^2}\sum_{j=1}^{J} n_j^2 W_j' W_j \hat{a}_j^2\right)\left(\frac{1}{J}\sum_{j=1}^{J} n_j W_j' W_j\right)^{-1}$ which is the standard heteroskedasticity-consistent standard error from a cell-level regression using the number of observations per cell as weights. Note that this is numerically identical to the quantity given by the clustered standard error from the micro-level regression. Some statistical packages will give slightly different answers for the two methods, due to the finite sample correction adjustment, which may differ between the two methods.

[7] It is numerically possible for the computed heteroskedasticty-consistent standard error to be larger than the cluster-consistent standard error. This is more likely when the specification error variance is very small or zero.

$$Y_{2ij} \;=\; g(x_j\,,\delta) \;+\; Y_{1ij}\,\alpha + \; v_{ij}\,,$$

where (Y_{1ij}, Y_{2ij}) is a pair of observed outcomes for the ith individual in the jth cell, h and g are smooth functions (e.g., low order polynomials), $D_j = 1[x_j \geq x_k]$ is an indicator of treatment status, β is the discontinuity in Y_1 at x_k induced by the treatment effect, α 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 α cannot be estimated consistently by a simple OLS procedure. When the functional forms of h and g are known, however, α can be estimated by the instrumental variables (IV) method using D_j as an instrument for Y_{1ij}. The maintained assumptions are that program status has no direct effect on Y_2, controlling for Y_1, and that the partial regression function g is smooth in a neighborhood of x_k.

As in the program evaluation setting, an important concern is that the functional forms of h and g are unknown. A natural extension of our framework is to assume that the data generating process for the observed outcomes is

$$Y_{1ij} \;=\; h(x_j\,,\gamma) \;+\; D_j\,\beta \;+\; a_{1j} \;+\; \varepsilon_{1ij}\,,$$

$$Y_{2ij} \;=\; g(x_j\,,\delta) \;+\; Y_{1ij}\,\alpha + \; a_{2j} \;+\; \varepsilon_{2ij}\,, \qquad i=1,2,...n_j\,; \quad j=1,2...J\,,$$

where (a_{1j}, a_{2j}) represents an i.i.d. vector of specification errors with mean 0 and variance Σ. Assuming that these errors are random, the model can be estimated consistently by standard IV, using D_j as an instrument for Y_{1ij}. The conventional IV sampling errors, however, ignore the group structure of the residuals and may overstate the precision of the IV estimator, especially if the number of observations per cell is large relative to the number of points of support of X. (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 (White, 1984).

To summarize, we propose that in RD research designs where the treatment-determining covariate (X) is discrete, researchers report standard errors that are "clustered" by the values of X. Sampling errors estimated in this way use information about the fit of the parametric model throughout the range of X to infer the precision of the estimated discontinuity at the treatment threshold x_k. It is important to emphasize, however, that clustered standard errors are only appropriate for **random** misspecification error.

III. Incorporating Extrapolation Errors

Under the assumption of random specification errors, cluster-consistent standard errors are adequate for the model given by Equation (3). There is, however, a hidden assumption that is required to justify these standard errors: the "specification error" that arises in estimating $E[Y_0 \mid X=x_k]$ needs to be equal in direction and magnitude to the error in estimating $E[Y_1 \mid X=x_k]$.

This circumstance in illustrated in Figure 3A, which abstracts from sampling error. The solid circles represent realized conditional means from the data generating process, the open circle represents the unobserved, counterfactual mean under the control regime, and the solid lines represent the underlying parametric function. By assuming that ß is equal to the parameter of interest $E[Y_1 - Y_0 \mid X=x_k]$, it is necessary to also assume that the parametric form understates (or overstates) $E[Y_0 \mid X=x_k]$ by the same magnitude as it understates (or overstates) $E[Y_1 \mid X=x_k]$; this needs to be true in repeated draws of the specification error.

There is an alternative formulation of the problem that relaxes this restriction. Once specification error is considered to be "random," one could assume that the error for $E[Y_1 \mid X=x_k]$ is independent of the error in extrapolating from the data to "forecast" $E[Y_0 \mid X=x_k]$. This possibility is illustrated in Figure 3B, which plots an example of one realization of the data generating process. Here, the parametric form understates $E[Y_1 \mid X=x_k]$ but overstates $E[Y_0 \mid X=x_k]$. As might be expected, since this alternative model is less restrictive, it ought to lead to more conservative inferences. Indeed, the cluster-consistent standard errors need to be further adjusted to account for this extra degree of uncertainty, as outlined below.

Inference Under Independent Counterfactual Specification Errors

To derive the necessary adjustment for the case of independent errors, consider the following potential outcomes version of the RD design with specification errors:

(5a) $E[Y_0 \mid X=x_j] = X_j \gamma + a_{j0}$,

(5b) $E[Y_1 \mid X=x_j] = X_j \gamma + \beta + a_{j1}$,

where (as in Section II) Y_0 and Y_1 represent outcomes in the absence and presence of treatment, and a_{j0} and a_{j1} represent the specification errors for the jth cell in the presence and absence of treatment, respectively. Assume that (a_{j0}, a_{j1}) are jointly normal, i.i.d. across cells, each with mean 0 and variance σ_a^2. As in the case where X is continuous, the object of interest in an RD framework is

$$E[Y_1 - Y_0 \mid X=x_k] = \beta + a_{k1} - a_{k0} .$$

Given an estimate $\hat{\beta}$ of the discontinuity in the systematic part of the mean outcome at $X=x_k$, the error in the forecast of the actual discontinuity in the potential outcomes is

$$\left(\hat{\beta} - \beta\right) - (a_{k1} - a_{k0}) .$$

Note that this reduces to the sampling error of $\hat{\beta}$ if and only if $a_{k1} = a_{k0}$. In this case, the discontinuity gap in the potential outcome at $X=x_k$ is just the discontinuity in the systematic part of the mean prediction for Y. Otherwise, when program status changes from 0 to 1 there is a new “draw” on the specification error, and the forecast error is larger.

Arguably a more natural assumption than $a_{j1} = a_{j0}$ is that the specification errors are independent and normally distributed. It follows that the error has variance $Var\left(\widehat{\beta}\right) + 2\sigma_a^2$, and hence the interval

$$\left(\widehat{\beta} - 1.96\sqrt{V\left(\widehat{\beta}\right) + 2\sigma_a^2}, \widehat{\beta} + 1.96\sqrt{V\left(\widehat{\beta}\right) + 2\sigma_a^2}\right) \tag{6}$$

contains $E[Y_1 - Y_0|X = x_k]$ with 0.95 probability.[8] The interpretation of this confidence interval is similar to conventional confidence intervals, except that here, the parameter $E[Y_1 - Y_0|X = x_k]$ is itself random. Thus, the correct inference statement is that the interval contains $E[Y_1 - Y_0|X = x_k]$ about 95 percent of the time in repeated draws of not only the sampling error -- but also the specification errors (and hence $E[Y_1 - Y_0|X = x_k]$).[9]

The interval in (6) strictly contains the usual confidence interval, and therefore leads to more conservative inferences. A wider interval is an intuitive result, since uncertainty regarding the true functional form ought to lead to more tentative inferences. Another intuitive aspect of the interval in (6) is that it collapses to the conventional one when the chosen parametric form is exactly correct and σ_a^2 equals zero.

Only one additional quantity is needed to construct the interval -- an estimate of σ_a^2 -- which can be consistently estimated by

$$\widehat{\sigma}_a^2 \equiv \frac{1}{N}\sum_{j=1}^{J} n_j \widehat{a}_j^2 - \frac{1}{N}\sum_{j=1}^{J} \widehat{\sigma}_{\varepsilon j}^2 \tag{7}$$

where *N* is the total number of (micro-level) observations and $\widehat{\sigma}_{\varepsilon j}^2$ is the unbiased estimate of the

[8] This is approximately true if ε_{ij} are non-normal, since $\widehat{\beta}$ would be asymptotically normal.
[9] Equation (6) has been called an ``Empirical Bayes" confidence interval.

jth cell variance of Y_{ij}.[10] It is simple to compute this quantity, in two steps. First, the micro-data are "collapsed" to the cell level, while computing the mean, number of observations, and the unbiased estimate of the variance of Y for each cell. Second, the cell-level means of Y are regressed on the parametric function of X (along with the dummy variable D), using the number of observations per cell as weights. The first term in the above expression is simply the mean squared error from this regression. The second term is just the average of the $\widehat{\sigma}_{\varepsilon j}^2$ across the J cells, multiplied by (J/N).

In the appendix, we specify the conditions under which

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

as J tends to infinity. This justifies the use of the adjusted confidence interval in (6).

IV. Efficiency

Given the model in Equations (5a) and (5b), the OLS estimator $\widehat{\beta}$ of $E[Y_1 - Y_0|X = x_k]$ is not asymptotically efficient in the class of linear estimators (and neither is the corresponding GLS estimator). This is because the parametric regression is essentially the difference between the prediction for $E[Y_1|X = x_k]$ and the prediction for $E[Y_0|X = x_k]$, as extrapolated from data away from the discontinuity threshold. While it is necessary to make such an extrapolation for $E[Y_0|X = x_k]$ (since this counterfactual is unobservable), it is unnecessary for $E[Y_1|X = x_k]$; it can be estimated by the sample mean $\overline{Y}_k$. This cell mean can be used to construct a more efficient estimator of the treatment effect.

[10] In some cases, the difference will be negative, in which case the estimate of σ_a^2 is zero.

Figure 3B illustrate this intuition. 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=x_k is the difference between the solid circle, which is above the parametric fit, and the open circle which is below the parametric fit. We cannot know how much the open circle deviates from the parametric form, but the cell mean provides information on how much the solid circle deviates from the linear approximation.

More formally, let $\widehat{\beta}$ and $\widehat{\gamma}$ be the same estimators as before, except after leaving out the data for the kth cell in the regression. Now consider an estimator of the treatment effect of the following form:

(8) $$\beta^* = \widehat{\beta} + \lambda\left(\overline{Y}_k - \left(x_k\widehat{\gamma} + \widehat{\beta}\right)\right)$$

Essentially, this is the same estimator, but with an adjustment according to the size of the cell mean's deviation from the parametric function. The error in this estimator is

$$(\widehat{\beta} - \beta) - (a_{k1} - a_{k0}) + \lambda\left(x_k\gamma + \beta + a_{k1} + \overline{\varepsilon}_k - \left(x_k\widehat{\gamma} + \widehat{\beta}\right)\right)$$

which has a zero mean. The variance of this estimator is

$$\begin{aligned}&V\left(\widehat{\beta}\right) + 2\sigma_a^2 + \lambda^2 V(a_{k1} + \overline{\varepsilon}_k - \left(x_k\widehat{\gamma} + \widehat{\beta}\right))\\&+ 2\lambda C\left((\widehat{\beta} - (a_{k1} - a_{k0}), a_{k1} + \overline{\varepsilon}_k - \left(x_k\widehat{\gamma} + \widehat{\beta}\right)\right)\end{aligned}$$

Neither $\widehat{\gamma}$ nor $\widehat{\beta}$ contain data from the kth cell, so this reduces to

$$\begin{aligned}&V\left(\widehat{\beta}\right) + 2\sigma_a^2 + \lambda^2(\sigma_a^2 + \frac{\sigma_{\varepsilon k}^2}{n_k} + V\left(x_k\widehat{\gamma} + \widehat{\beta}\right))\\&- 2\lambda(C(\widehat{\beta}, x_k\widehat{\gamma} + \widehat{\beta}) + \sigma_a^2)\end{aligned}$$

Differentiating with respect to λ and solving for the first order condition yields the optimal λ, given by

$$\lambda = \frac{\sigma_a^2 + C(\widehat{\beta}, x_k\widehat{\gamma} + \widehat{\beta})}{\sigma_a^2 + V\left(x_k\widehat{\gamma} + \widehat{\beta}\right) + \frac{\sigma_{\varepsilon k}^2}{n_k}} \tag{9}$$

The intuition behind this combination is seen 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 -- or, in other words, the terms of the parametric function are completely interacted with the treatment dummy variable. Use the equality $C(\widehat{\beta}, x_k\widehat{\gamma} + \widehat{\beta})$ $=$ $V\left(x_k\widehat{\gamma} + \widehat{\beta}\right)$ $-$ $C\left(x_k\widehat{\gamma}, x_k\widehat{\gamma} + \widehat{\beta}\right)$, and note that $C\left(x_k\widehat{\gamma}, x_k\widehat{\gamma} + \widehat{\beta}\right) = 0$ here, because only data to the left are used to estimate $x_k\widehat{\gamma}$ and only data to the right are used to estimate $x_k\widehat{\gamma} + \widehat{\beta}$. The optimal value of λ then becomes:

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

When the parametric function is "good", then σ_a^2 is relatively small compared to the cell-level sampling error $\frac{\sigma_{\varepsilon k}^2}{n_k}$ -- due to small specification errors. λ will thus tend to 0, and the linear combination estimator collapses to the original parametric estimator $\widehat{\beta}$. On the other hand, if the parametric form is a "bad", then σ_a^2 will be relatively large. As a result, λ will tend towards 1, and the combination estimator will converge towards $\overline{Y}_k - x_k\widehat{\gamma}$ -- the difference between the cell mean and the "prediction" of $E[Y_0|X = x_k]$ 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 $E[Y_1 - Y_0|X = x_k]$ -- $\widehat{\beta}$ and $\overline{Y}_k - x_k\widehat{\gamma}$. Note that the usual OLS estimator that *includes* the kth cell can also be written in the same form as Equation (8), using the recursive residual formula of Brown, Durbin, and Evans (1975). The implied weight by the OLS will in general not be equal to the weight given by Equation (9).[11]

The optimal λ can be substituted into the expression above to yield the variance of this combination estimator:

$$V(\beta^*) = \left(V\left(\widehat{\beta}\right) + 2\sigma_a^2\right) - \lambda^2\left(\sigma_a^2 + V\left(x_k\widehat{\gamma} + \widehat{\beta}\right) + \frac{\sigma_{\varepsilon k}^2}{n_k}\right) \tag{10}$$

Note that the first set of parentheses is the error variance as discussed in the previous section. Thus, the combination estimator of the treatment effect will always have a smaller error variance than the parametric estimator $\widehat{\beta}$. Note that in adjusting the confidence intervals from the previous section, not only must the width of the interval be shortened using the expression above, it must be re-centered around the new point estimate β^* .

To make this estimator feasible, it is only necessary to obtain sample analogues to the population variances and covariances in the above two expressions. σ_a^2 can be estimated by $\widehat{\sigma}_a^2$ as defined in the previous section. The estimator for $V\left(x_k\widehat{\gamma} + \widehat{\beta}\right)$ is simply the "standard error of the prediction" at $X = x_k$, which is a standard option in most statistical packages. $C(\widehat{\beta}, x_k\widehat{\gamma} + \widehat{\beta}) = V\left(\widehat{\beta}\right) + C\left(x_k\widehat{\gamma}, \widehat{\beta}\right)$ 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

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

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

estimated intercept $x_k\widehat{\gamma}$; again, these quantities are automatically computed in most statistical packages. Finally, $\frac{\sigma^2_{\varepsilon k}}{n_k}$ can be estimated by the estimated variance of the mean $\overline{Y}_k$. Together, these quantities imply an estimator $\widehat{\lambda}$, which can be used to construct a feasible version of β^* .

In the appendix, we provide conditions under which

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

where $\widehat{\beta^*}$ and $\widehat{V\left(\widehat{\beta^*}\right)}$ are defined by the above expressions, with population quantities replaced by their sample analogues.

Summary of Implementation

By way of summarizing all of the methods proposed in this paper, we recommend the following procedure:

1) Normalize the X variable so the threshold is at 0 , so the intercept in the regression can be interpreted as the estimate of $E[Y_0|X = x_k]$. 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. If the cluster-consistent standard errors are significantly larger, it suggests non-trivial specification error. If relaxing the "identical" counterfactual specification error assumption is desired, proceed to the next step.

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

2) Collapse the data keeping the means, variances, and number of observations at the cell level.
3) Run the (cell size-weighted) regression using the cell-data. Verify that the point estimate is identical to that computed in step 1. Also verify that the heteroskedasticity-consistent standard error is identical to the cluster-consistent standard error in step 1 (except for the finite-sample correction factor). Use mean squared error from the regression and cell variances to compute $\widehat{\sigma}_a^2$ as in Equation (8). Adjust standard errors by $2\widehat{\sigma}_a^2$. If a more efficient estimator is desired, proceed to the next step.
4) Perform step 3, except exclude data from the k th cell (the "first" cell in which $D = 1$). Using the estimated variances and covariances of the discontinuity coefficients and intercept, as well as the k th cell variance, compute $\widehat{\lambda}$ as in Equation (9). Using the k th cell mean, compute $\widehat{\beta^*}$; then compute $\widehat{V\left(\widehat{\beta^*}\right)}$ as in Equation (10).

V. Relation to Bayes' and Empirical Bayes' Procedures

There is a close connection to the proposed estimator $\widehat{\beta^*}$ and Bayes' and Empirical Bayes' approaches. As we show below, the confidence intervals provided in Sections IV and V can equivalently be viewed as either Bayesian "confidence intervals" or (parametric) Empirical Bayes confidence intervals.

First note that the Equation (7) can be rewritten as

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

The expression in brackets can be viewed as an estimate of $E[Y_1|X = x_k]$ -- an average of the k th cell mean and the predicted value from the regression -- and the term $x_k\widehat{\gamma}$ as an estimate of

$E[Y_0|X = x_k]$.

Consider the simplest Bayesian approach to estimating $E[Y_1|X = x_k]$ – $E[Y_0|X = x_k]$. A likelihood for the observed data would be specified; for example, $Y_{ik} \sim N(E[Y_1|X = x_k], \sigma^2)$; assume for sake of exposition that σ^2 is known. Suppose one assumes a prior distribution for $\begin{pmatrix} E[Y_1|X = x_k] & E[Y_0|X = x_k] \end{pmatrix}'$ given by $N\left(\begin{pmatrix} B_1 & B_0 \end{pmatrix}', \begin{matrix} \sigma_1^2 & 0 \\ 0 & \sigma_0^2 \end{matrix}\right).$ For $E[Y_1|X = x_k]$, the posterior distribution would be $N(\lambda \overline{Y}_k + (1-\lambda)B_1, (1-\lambda)\sigma_1^2)$ with $\lambda = \sigma_1^2 / \left(\frac{\sigma^2}{n_k} + \sigma_1^2\right)$; for $E[Y_0|X = x_k]$, since there is no data on the "control" regime at $X = x_k$, the posterior is equal to the prior: $N(B_0, \sigma_0^2)$. With some re-arrangement, the resulting posterior distribution for $E[Y_1 - Y_0|X = x_k]$ is $N([\lambda \overline{Y}_k + (1-\lambda)B_1] - B_0$, $\sigma_1^2 + \sigma_0^2 - \lambda^2(\frac{\sigma^2}{n_k} + \sigma_1^2))$. Note that under an uninformative (diffuse) prior on $E[Y_0|X = x_k]$, the posterior for the treatment effect will also be uninformative. In the case where only data on the k th cell is provided, this is intuitive: without any outside information, one should not be able to provide an informative estimate of the treatment effect.

What would be reasonable values for B_1 , B_0 , σ_1^2 , and σ_0^2 ? One possibility is to use the data away from the discontinuity point to justify the parameters of the prior distribution. For example, a reasonable B_1 might be $x_k\widehat{\gamma} + \widehat{\beta}$, the predicted value of $E[Y_1|X = x_k]$ using all data to the right of the k th cell in a parametric regression. A reasonable value for σ_1^2 could be the variance in that prediction, namely, $V\left(x_k\widehat{\gamma} + \widehat{\beta}\right) + \sigma_a^2$. Analogously, a regression using all data to the left of the k th cell, would produce $x_k\widehat{\gamma}$ and $V(x_k\widehat{\gamma}) + \sigma_a^2$ as reasonable values for the location and scale of the prior on $E[Y_0|X = x_k]$. Choosing these values for the prior, in fact, yields a posterior given by $N(\beta^*, V(\beta^*))$.[12] Therefore, the confidence interval described in the previous section can be equivalently interpreted as a Bayesian "confidence interval", using the regression predictions and its error in the prior distribution.

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 $\left[\lambda\overline{Y}_k + (1-\lambda)\left(x_k\widehat{\gamma} + \widehat{\beta}\right)\right]$ of $E[Y_1|X = x_k]$ is simply a specific application of the Empirical Bayes, "shrinkage"/Stein estimator (see Morris' review of the parametric Empirical Bayes approach). The confidence intervals provided in Sections IV and V -- in which the probability statement is with respect to the randomness in sampling and randomness in a -- are indeed, Empirical Bayes confidence regions.

VI. Summary

This paper draws attention to functional form issues in the estimation of regression discontinuity designs, when X is a discrete variable. In particular, in the discrete case, the conditions for non-parametric or semi-parametric methods are not applicable; indeed, even with an infinite amount of data, the treatment effect is not even identified without assuming a parametric form. Our goal was to formally incorporate uncertainty in the necessary parametric modeling of the underlying RD function.

We have proposed a procedure for inference that explicitly acknowledges errors in

[12] This is true when separate parametric forms are used to estimate the function on the left and the right.

whatever parametric functional form is chosen. Instead of assuming that the chosen functional form "correctly" describes the underlying regression function, we model the deviations from the true conditional means from the parametric function as random specification errors, allowing for an unknown variance. This relaxation of the model requires -- at a minimum -- the computation of cluster-consistent standard errors (clustered on the distinct values of X), as opposed to the conventional OLS standard errors. An even more flexible model of the RD counterfactual functions requires further adjustment; the resulting confidence intervals can be formally justified as Bayes' or Empirical Bayes' intervals.

Even though we allow for specification error, there still remains the issue of how to choose the functional form for the systematic part of the functional form (i.e. how many polynomial terms in X). On the other hand, we consider our approach to be superior to simply assuming the parametric form is correct; because the possibility that the functional form is correct (zero specification error) is included as a special case, and results in (asymptotically) identical standard errors. Useful from a practitioner's perspective, the adjustments that we propose are either provided automatically or can be easily computed from variances and covariances provided by regression routines in standard statistical packages.

Throughout the paper, we have assumed that specification errors are assumed to be homoskedastic and serially independent – from one cell to the next. A natural extension would be to formally test for these assumptions, and relax the model to allow for non-independent specification errors between adjacent cells; it is likely that this allowance would lead to tighter inferences, since one could take advantage of serially correlated errors to provide more precise forecasts of the counterfactual mean.

Appendix

Independent Counterfactual Specification Errors: Asymptotic Distribution

For the case of independent counterfactual specification errors, we establish conditions under which

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

After normalizing X_j so that the threshold is at $X = 0$, let w_j denote the variables (excluding the constant) in the parametric regression (and θ, the corresponding coefficient vector), with each variable deviated from its sample mean. Make the following assumptions: 1) $E[W_j] \neq 0$, 2) $w_j \equiv \frac{1}{\sqrt{J}} w_j^*$, with $\frac{1}{J}\sum_{j=1}^{J} n_j^* w_j^{*\prime} w_j^* \xrightarrow{p} V_w$, positive definite, 3) $n_j = Jn_j^*$, n_j^* a finite constant with $\frac{1}{J}\sum_j n_j^* = \overline{n}$ a positive constant, and 4) $\varepsilon_{ij} = \sqrt{n_j}\,\varepsilon_{ij}^*$, so that $\sigma_{\varepsilon j}^2 = n_j \sigma_{\varepsilon j}^{*2}$, $\sigma_{\varepsilon j}^{*2}$ finite constant. The reason for the first two assumptions is discussed below, while the reason for the third and fourth are discussed in the next section. Let $J \rightarrow \propto$.

We need to show two things: 1) $\left(\widehat{\beta} - E[Y_1 - Y_0|X = x_k]\right) \xrightarrow{d} N\left(0, V\left(\widehat{\beta}\right) + 2\sigma_a^2\right)$, and 2) $\left(\widehat{V\left(\widehat{\beta}\right)} + 2\widehat{\sigma}_a^2\right)$ is a consistent estimator of $V\left(\widehat{\beta}\right) + 2\sigma_a^2$.

First, note that $\widehat{\beta} - E[Y_1 - Y_0|X = x_k] = \left(\widehat{\beta} - \beta\right) - (a_{k1} - a_{k0})$. The second term, by assumption is distributed as $N(0, 2\sigma_a^2)$. The first term is simply the element of $\widehat{\theta} - \theta$ that corresponds to the discontinuity gap. We have $\frac{1}{J}\sum_j n_j w_j' w_j \xrightarrow{p} E\left[n_j^* w_j^{*\prime} w_j^*\right]$. Consequently,

$\widehat{\theta}-\theta = \left(\frac{1}{J}\sum_j n_j w_j' w_j\right)^{-1}\frac{1}{J}\sum_j n_j w_j'(a_j+\overline{\varepsilon}_j)$ converges in distribution to $N\left(0, V_w^{-1}E[n_j^{*2}w_j^{*\prime}w_j^{*}(a_j+\overline{\varepsilon}_j)^2]V_w^{-1}\right)$. Note that $\widehat{\theta}-\theta$ is $O_p(1)$, (and not $O_p\left(\frac{1}{\sqrt{J}}\right)$, as in the usual case).

Second, decompose $\frac{1}{J^2}\sum_j n_j^2 w_j' w_j \widehat{a}_j^2$ into $\frac{1}{J^2}\sum_j n_j^2 w_j' w_j (a_j+\overline{\varepsilon}_j)^2$ + $\frac{1}{J^2}\sum_j n_j^2 w_j' w_j$ • ($2(a_j+\overline{\varepsilon}_j)w_j\left(\theta-\widehat{\theta}\right)$ + $\left(w_j\left(\theta-\widehat{\theta}\right)\right)^2$), which has probability limit $E[n_j^{*2}w_j^{*\prime}w_j^{*}(a_j+\overline{\varepsilon}_j)^2]$, since the second summation is $o_p(1)$. Thus, $\widehat{V\left(\widehat{\theta}\right)}$ is consistent for $V\left(\widehat{\theta}\right)$. Finally, decompose $\frac{1}{N}\sum_j n_j\widehat{a}_j^2$ as $\frac{1}{J}\frac{1}{\overline{n}}\sum_j n_j^{*}(a_j+\overline{\varepsilon}_j)^2$ + $\frac{1}{J}\frac{1}{\overline{n}}\sum_j n_j^{*}(2(a_j+\overline{\varepsilon}_j)w_j\left(\theta-\widehat{\theta}\right)$ + $\left(w_j\left(\theta-\widehat{\theta}\right)\right)^2)$, which converges in probability to $\sigma_a^2+\frac{1}{J\overline{n}}\sum_j\sigma_{\varepsilon j}^{*2}$, since the second summation can be shown to be $o_p(1)$. $\frac{1}{N}\sum_j\widehat{\sigma}_{\varepsilon j}^2$ = $\frac{1}{J\overline{n}}\sum_j\frac{1}{n_j}\left(\sum_i\varepsilon_{ij}^{*2}\right)$ + $\frac{1}{J^2\overline{n}}\sum_j\frac{1}{n_j}\left(\sum_i\left(-2\varepsilon_{ij}(\overline{\varepsilon}_j)+\overline{\varepsilon}_j^2\right)\right)$, which converges to $\frac{1}{J\overline{n}}\sum_j\sigma_{\varepsilon j}^{*2}$, since the second term is $o_p(1)$. Thus, $\widehat{\sigma}_a^2 \equiv \frac{1}{N}\sum_j n_j\widehat{a}_j^2 \; -\frac{1}{N}\sum_j\widehat{\sigma}_{\varepsilon j}^2$ is consistent for σ_a^2.

We do not consider the assumption of $w_j = \frac{1}{\sqrt{J}}w_j^{*}$ to be a literal description of the data generating process. Rather, we invoke the assumption in order to justify the asymptotic approximation. Without such a modification, $\left(\widehat{\beta}-\beta\right)$ – $(a_{k1}-a_{k0})$ is $O_p\left(\frac{1}{\sqrt{J}}\right)$ – $O_p(1)$, and would converge in distribution to $N(0,2\sigma_a^2)$. This approximation would incorporate the "prediction" error at the expense of ignoring the estimation error in $\widehat{\beta}$. On the other hand, $\sqrt{J}\left(\left(\widehat{\beta}-\beta\right)-(a_{k1}-a_{k0})\right)$ has an infinite variance. By letting the scale of w

shrink as the sample size grows, the distribution of $\widehat{\beta}$ is stabilized, allowing our asymptotic approximation to incorporate both estimation error and extrapolation error. Note that the error in the intercept is $\widehat{\alpha} - \alpha = (\overline{W})\left(\theta - \widehat{\theta}\right) + \frac{1}{J}\sum_{j=1}^{J}(a_j + \overline{\varepsilon}_j)$. Therefore, in order for the intercept to converge in distribution at the same rate as the slope coefficients, the mean of W_j cannot be zero.

Asymptotic Distribution of the "shrinkage" estimator.

We now establish the conditions under which

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

Maintain the four assumptions as specified above, and normalize so that x_k, the point of the threshold is zero. We will prove the result in two steps: 1) $\widehat{\beta^*} - E[Y_1 - Y_0|X = x_k] \overset{d}{\to} N(0, V(\beta^*))$, and 2) and $\widehat{V\left(\widehat{\beta^*}\right)}$ is consistent for $V(\beta^*)$.

First, re-write $\frac{1}{\sqrt{J}}\left(\widehat{\beta^*} - E[Y_1 - Y_0|X = 0]\right)$ as $\frac{1}{\sqrt{J}}(\widehat{\beta} - E[Y_1 - Y_0|X = 0] + \widehat{\lambda}\left(\overline{Y}_k - \left(\widehat{\alpha} + \widehat{\beta}\right)\right))$. Define $b_J^{'}$ as the vector $\left(\ \frac{1}{\sqrt{J}}\left(\widehat{\beta} - E[Y_1 - Y_0|X = 0]\right)\ \frac{1}{\sqrt{J}}\left(\overline{Y}_k - \left(\widehat{\alpha} + \widehat{\beta}\right)\right)\ \widehat{\lambda}\ \right)^{\prime}$, so that $\frac{1}{\sqrt{J}}\left(\widehat{\beta^*} - E[Y_1 - Y_0|X = 0]\right) = f(b_J)$.

We need to show b_J has probability limit $b' = \left(\begin{array}{ccc} 0 & 0 & \lambda \end{array} \right)'$, and that $\sqrt{J}(b_J - b)$ converges in distribution to $N(0, V^*)$. If true, then $\sqrt{J}(f(b_J) - f(b))$ will converge in distribution to $N\left(0, \left(\begin{array}{ccc} 1 & \lambda & 0 \end{array} \right) V^* \left(\begin{array}{ccc} 1 & \lambda & 0 \end{array} \right)'\right)$, 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 $\widehat{\beta^*} - E[Y_1 - Y_0 | X = 0] \overset{d}{\to} N(0, V(\beta^*))$.

To show $b_J \overset{p}{\to} \left(\begin{array}{ccc} 0 & 0 & \lambda \end{array} \right)'$, recall that $\widehat{\beta} - E[Y_1 - Y_0 | X = 0]$ is $O_p(1)$; multiplying by $\frac{1}{\sqrt{J}}$, yields $o_p(1)$. Similarly, $\overline{Y}_k - \left(\widehat{\alpha} + \widehat{\beta}\right) = (\alpha - \widehat{\alpha}) + \left(\beta - \widehat{\beta}\right) + a_k + \overline{\varepsilon}_k$ is also $O_p(1)$; multiplying by $\frac{1}{\sqrt{J}}$ yields $o_p(1)$. $\widehat{\lambda}$ is consistent for λ, because the sample analogs to each of its parts are consistent. For example, by the same argument as in the previous section, the standard estimators for $C(\widehat{\beta}, \widehat{\alpha} + \widehat{\beta})$ and $V\left(x_k \widehat{\gamma} + \widehat{\beta}\right)$ are consistent; $\widehat{\sigma}_a^2$ is consistent as shown above. Also, $\frac{1}{n_k^2} \sum_{i=1}^{n_k} (Y_{ik} - \overline{Y}_k)^2 = \frac{1}{n_k^* n_k} \sum_{i=1}^{n_k} \varepsilon_{ik}^{*2} + \frac{1}{J^2 n_k^{*2}} \sum_{i=1}^{n_k} (2\varepsilon_{ik} \overline{\varepsilon}_k + \overline{\varepsilon}_k^2) \overset{p}{\to} \sigma_{\varepsilon k}^{*2} / n_k^* = \sigma_{\varepsilon k}^2 / n_k$, because the second summation can be shown to be $o_p(1)$.

To show $\sqrt{J}(b_J - b) \overset{d}{\to} N(0, V^*)$, we will show that $\sqrt{J}(b_J - b)$ converges in probability to a sum of random vectors, each either normally distributed or converges to a normal distribution.

$$\sqrt{J}(b_J - b) = \begin{pmatrix} \widehat{\beta} - \beta \\ (\alpha - \widehat{\alpha}) + \left(\beta - \widehat{\beta}\right) \\ \sqrt{J}\left(\widehat{\lambda} - \lambda\right) \end{pmatrix} + \begin{pmatrix} 0 \\ \overline{\varepsilon}_k \\ 0 \end{pmatrix} + \begin{pmatrix} a_{k1} - a_{k0} \\ a_{k1} \\ 0 \end{pmatrix}$$

The element in the second vector is $\frac{1}{n_k}\sum_{i=1}^{n_k} \varepsilon_{ik} = \frac{1}{\sqrt{n_k^*}}\frac{1}{\sqrt{J}}\sum_{i=1}^{n_k} \varepsilon_{ik}^*$, which converges to a normal. The third vector is normal, by assumption.

The first vector: $\widehat{\theta} - \theta$ converges to $(V_w)^{-1}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^{J} n_j^* w_j^{*\prime}(a_j + \overline{\varepsilon}_j)\right)$. $\widehat{\beta} - \beta$ can thus be expressed as a summation of the form $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} z_j$, where z_j is mean zero i.i.d. $(\alpha - \widehat{\alpha}) = (\overline{W})\left(\theta - \widehat{\theta}\right) + \frac{1}{J}\sum_{j=1}^{J}(a_j + \overline{\varepsilon}_j)$ which converges in probability to $(E[W_j])\left(\theta - \widehat{\theta}\right)$. Thus, $(\alpha - \widehat{\alpha}) + \left(\beta - \widehat{\beta}\right)$ is simply a linear function of the elements of $\left(\theta - \widehat{\theta}\right)$, and therefore can be expressed in the form of $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} z_j$.

Finally, we must show that $\sqrt{J}\left(\widehat{\lambda} - \lambda\right)$ can also be expressed as a summation in the form of $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} z_j$.

$$\sqrt{J}\left(\frac{\widehat{\sigma}_a^2 + \widehat{C(\widehat{\beta}, \widehat{\alpha} + \widehat{\beta})}}{\widehat{\sigma_a^2} + \widehat{V\left(\widehat{\alpha} + \widehat{\beta}\right)} + \widehat{\frac{\sigma_{\varepsilon k}^2}{n_k}}} - \frac{\sigma_a^2 + C(\widehat{\beta}, \widehat{\alpha} + \widehat{\beta})}{\sigma_a^2 + V\left(\widehat{\alpha} + \widehat{\beta}\right) + \frac{\sigma_{\varepsilon k}^2}{n_k}} \right)$$

converges in probability to

$$\sqrt{J}\left(\frac{\widehat{\sigma}_a^2 - \sigma_a^2 + \widehat{C(\widehat{\beta}, \widehat{\alpha} + \widehat{\beta})} - C(\widehat{\beta}, \widehat{\alpha} + \widehat{\beta})}{\sigma_a^2 + V\left(\widehat{\alpha} + \widehat{\beta}\right) + \frac{\sigma_{\varepsilon k}^{*2}}{n_k^*}} \right)$$

The numerator can be shown to be a summation in the form of $\frac{1}{\sqrt{J}}\sum_{j=1}^{J} z_j$. 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 $V(\beta^*)$.

References

Angrist, J. and A. Krueger (1999), “Empirical Strategies in Labor Economics,” Handbook of Labor Economics, Volume 3, Ashenfelter, A. and D. Card, eds., Amsterdam: Elsevier Science.

Angrist, J.D. and V. Lavy, 1999, “Using Maimondes' rule to estimate the effect of class size on scholastic achievement,” *Quarterly Journal of Economics* 114, 533-575.

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

Card, David & Lara D. Shore-Sheppard, 2002. "Using Discontinuous Eligibility Rules to Identify the Effects of the Federal Medicaid Expansions on Low Income Children," NBER Working Papers 9058, National Bureau of Economic Research, Inc.

Chamberlain, Gary (1994), “Quantile Regression, Censoring and the Structure of Wages,” in Sims, C., ed., *Advances in Econometrics: Proceedings from the Sixth World Congress,* Cambridge U. Press.

Dinardo, John, and David S. Lee, “The Impact of Unionization on Private Sector Employers,” UC Berkeley Manuscript, February 2004.

Hahn, Jinyong & Todd, Petra & Van der Klaauw, Wilbert, 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," Econometrica, Vol. 69 (1) pp. 201-09.

Thomas J. Kane, “A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going”, National Bureau of Economic Research Working Paper No. 9703, May 2003.

Lee, David S. “Randomized Experiments from Non-Random Selection in U.S. House Elections”, UC Berkeley Manuscript, 2003.

Morris, C (1983), “Parametric empirical Bayes inference: Theory and applications,” *Journal of the American Statistical Association*, 78, 47--55.

Moulton, Brent R, 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit," The Review of Economics & Statistics, Vol. 72 (2) pp. 334-38.

Porter, Jack, “Estimation in the Regression Discontinuity Model,” Harvard University Manuscript, 2003.

Shore-Sheppard, Lara, “The precision of Instrumental Variables Estimates with Grouped Data,” Princeton University Industrial Relations Section Working Paper #374, December 1996.

Thistlethwaite, Donald, and Donald Campbell, (1960) “Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment,” *Journal of Educational Psychology*, 51, 309-17.

White, H., *Asymptotic Theory For Econometricians*. New York: Academic Press (1984)

Figure 1: Regression Discontinuity, Continuous Covariate

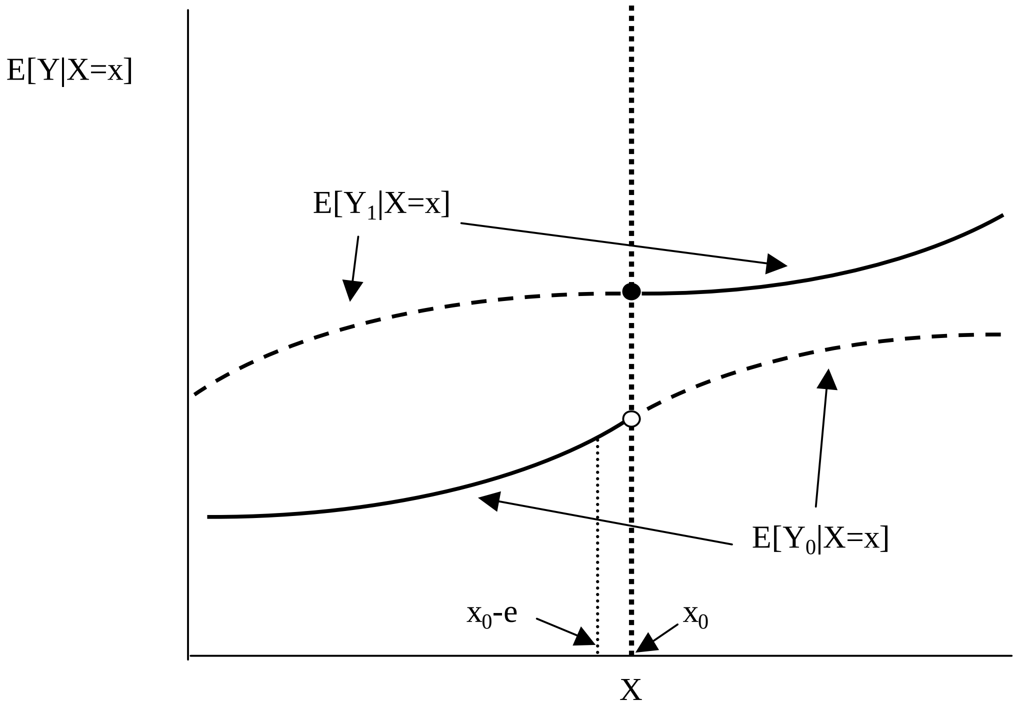


Figure 2: Regression Discontinuity, Discrete Covariate

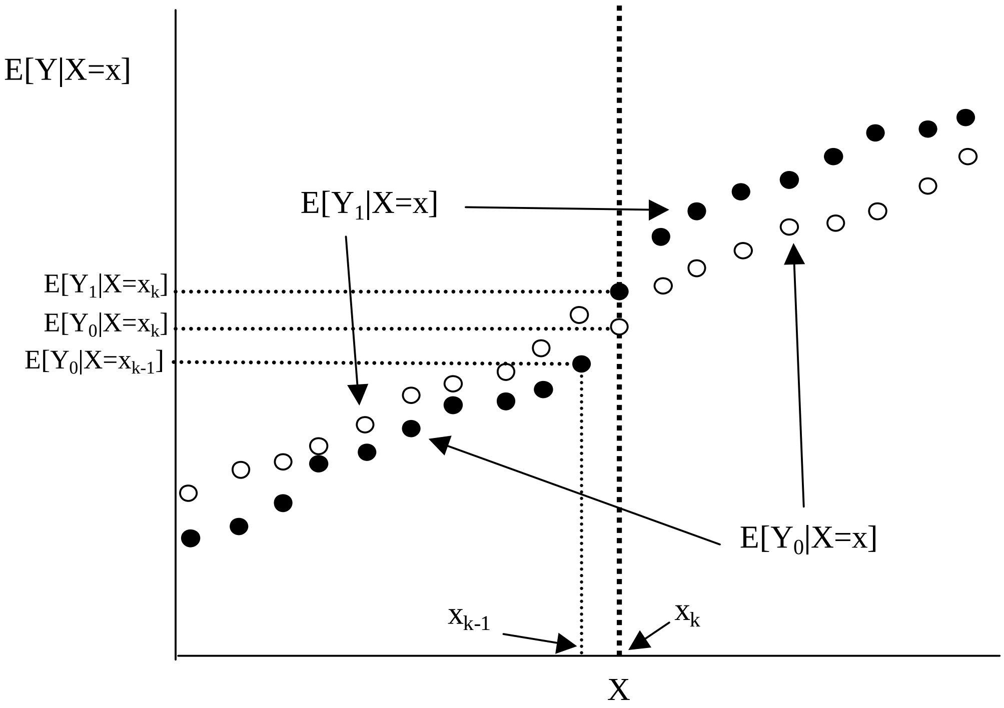


Figure 3A: Counterfactual Specification, Identical Errors

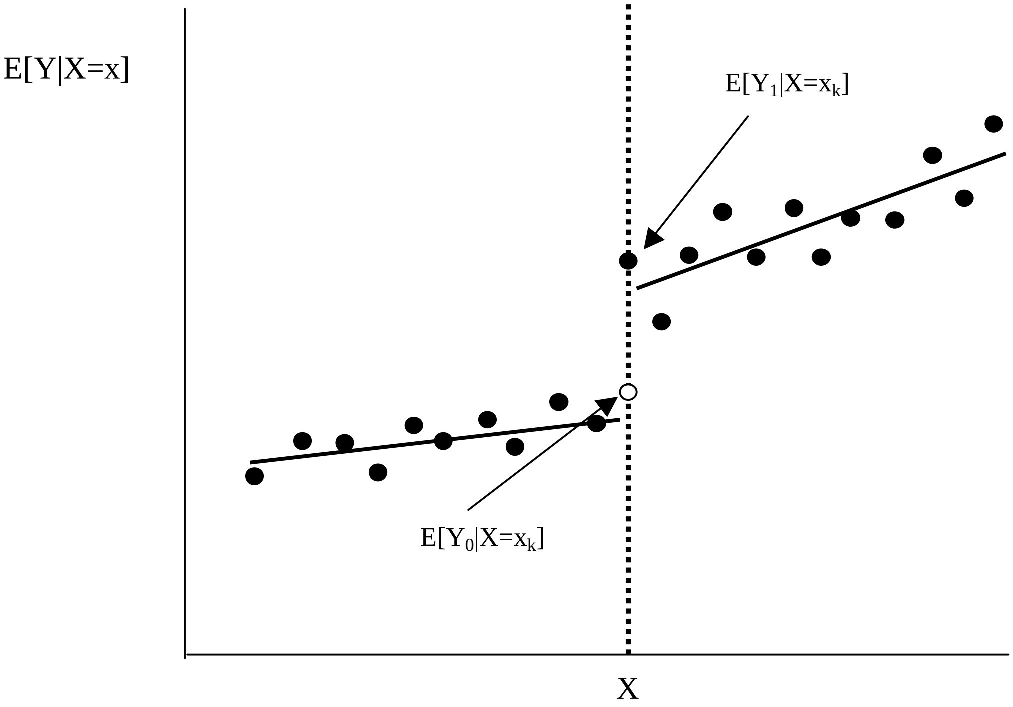
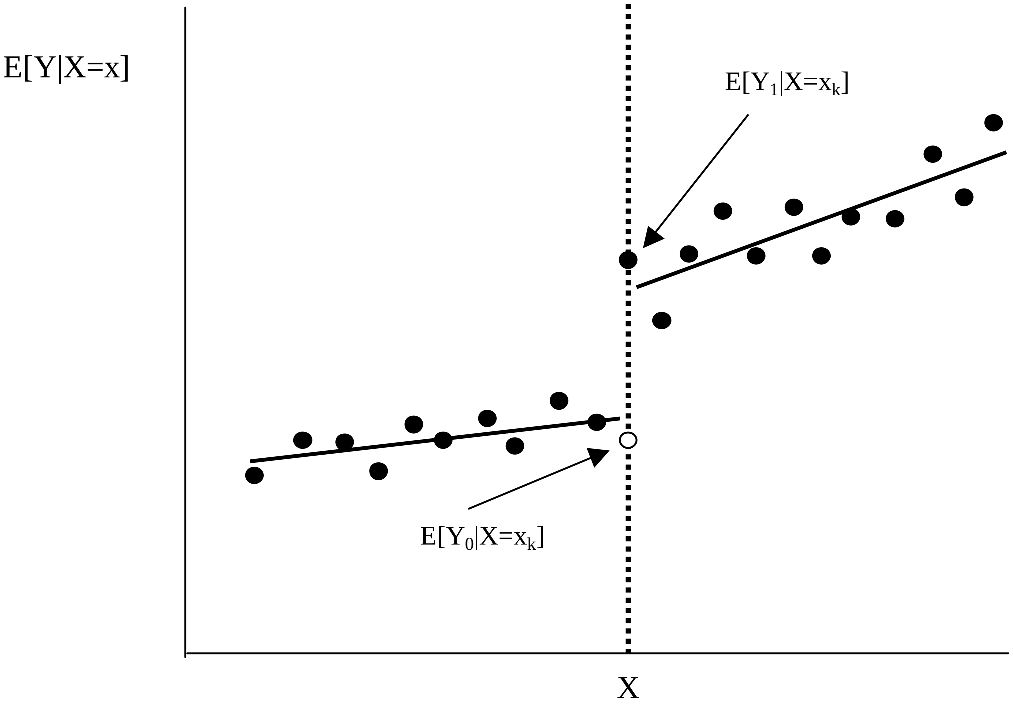


Figure 3B: Counterfactual Specification, Independent Errors


E[Y|X=x]
$E[Y_1|X=x_k]$
$E[Y_0|X=x_k]$
X