1 Introduction to Non-Parametric Estimators

- Suppose we want to estimate a highly non-linear relation between two variables. How would we do it?
  - Estimate relation with set of orthogonal functions?
    - High order polynomials.
    - Trigonometric functions.
    - Problems?
      - Very sensitive to outliers.
      - Non-local impact of outliers
  - Splines (linear, quadratic, ...)

* Divide X into I different sections.. $1, 2, ...I$

$$\min_{\hat{\alpha}_i, \hat{\beta}_i} \sum_i \left(Y_{it} - \hat{\alpha}_i - \hat{\beta}_i X_{it} \right)$$

$$s.t.\ \hat{\alpha}_{i-1} + \hat{\beta}_{i-1} X_I = \hat{\alpha}_i + \hat{\beta}_i X_I$$

* Linear, Quadratic, Quartic, Trigonometric
* More local impacts but non-differentiable

- Non-parametric estimators
  * Require tons of data - especially problematic with high dimensionality of estimation
  * Must choose how locally to estimate (bandwidth)
- Semi-parametric estimators
  * Requires greater functional form assumptions
  * Better at dealing with high dimensional estimation

2 Histograms

- The histogram is a probability mass function which is usually an approximation to the probability density function (pdf) of a random varialbe.

- To create a histogram for a variable X, divide up X into K parts $[0, X_k)$ (could be equal portion of X-space or any other division of the X-space).

- Then the histogram is:

$$f(x_k) = \sum_j \frac{I\left(X_{k-1} \leq x_j < X_k\right)}{X_k - X_{k-1}}$$

where $I(\cdot)$ is the indicator function.

- If the histograms are of equal length in X-space, then we can write the density as:

$$f(x_k) = \sum_j \frac{I\left(X_k - h \leq x_j < X_k + h\right)}{2h}$$

- In the limit as $h \rightarrow 0$, if the density (pdf) is differentiable, then you will recover the density.

3 Kernel Density Estimation

- The kernel density estimator is a generalization of the histogram - it is in general smoother.

- The histogram density is for a sample from the population. Often the sample is a noisy estimate of the population. Therefore, kernel densities smooth the density estimates between points using functions called kernel functions.

- The value of the estimator at a point x_o is

$$\hat{f}(x_o) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x_i - x_o}{h}\right)$$

where N is the number of total points being used in the estimation of the density.

- $K(\cdot)$ is called the kernel function and it is what smooths the density. It must satisfy 4 conditions

  1. $K(z)$ is symmetric around zero and continuous
  2. $\int K(z)\,dz = 1$, $\int zK(z)\,dz = 0$, and $\int |z|\,K(z)\,dz < \infty$
  3. Either
     (a) $\exists z_0$ such that $K(z_0) = 0\ \forall z$ such that $|z| \geq z_0$
     (b) $\lim_{z\to\infty} |z|\,K(z) = 0$
  4. $\int z^2 K(z)\,dz = c < \infty$

- Usually kernel functions satisfy ($3a.$) not just ($3b.$)

- Usually $\frac{\partial K}{\partial |z|} \leq 0$ so that the impact of data points z_k on the value of the non-parametric estimator at a point z_0 decline with distance between z_0 and z_k.

- h is called the bandwidth parameter; it roughly gives the size of the histogram bins.
  - Tradeoff: h large $\implies$ density estimate is smoother
  - h small $\implies$ less functional form bias

- Different kernels
  - Uniform: $\frac{1}{2} \cdot I\left(|z| < 1\right)$
  - Triangular: $(1 - |z|) \cdot I\left(|z| < 1\right)$
  - Epanechnikov: $\frac{3}{4}\left(1 - z^2\right) \cdot I\left(|z| < 1\right)$
  - Quartic: $\frac{15}{16}\left(1 - z^2\right)^2 \cdot I\left(|z| < 1\right)$

- Gaussian: $(2\pi)^{-\frac{1}{2}} e^{-\frac{z^2}{2}}$

- Most popular: Epanechnikov and Uniform
  - Kernels with higher order terms fit better (lower bias)
  - Kernels with lower order terms are smoother

- The kernel density estimator is biased as $N \longrightarrow \infty$ keeping h fixed but not if $h \longrightarrow 0$ as $N \longrightarrow \infty$
  - Since inference is done with a fixed h, assymptotic statistical inference is complicated by an assymptotic bias term.
  - Often densities don't have error bars on them

- Note that there are two types of convergence we can discuss since we discussing convergence to a density not just a paramter:

  - Convergence in distribution

  - Pointwise convergence

  - Most inference is pointwise

- One choice for an optimal bandwidth can come from minimizing mean integrated square error (between the density and the data).

- Two choices: kernel and bandwidth. Choice of kernel doesn't usually have a large impact on the estimation. Choice of bandwidth, however, is crucial.

4 Non-parametric Regression

- Can we use local regression methods to characterize the relationship between two variables as opposed to the density of a variable and the variable itself? Yes!!! Its called non-parametric regression.

- Definition of the estimator:

$$\hat{m}(x_0) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)}$$

where again $K(\cdot)$ is the kernel and h is the bandwidth.

- Basically you are averaging $Y(X_0)$ with $X's$ close to X_0 and in a weighted fashion.

- Special case of Local Weighted Average Estimator

$$\hat{m}(x_0) = \sum_{i=1}^{N} w_{i0,h} y_i$$

where $w_{io,h} = w(x_i, x_0, h)$

- K-Nearest Neighbor Estimator

$$\hat{m}(x_0) = \frac{1}{k}\left(y_{i-\left(\frac{k-1}{2}\right)} + ... + y_{i+\left(\frac{k+1}{2}\right)}\right)$$

- Generalization of kernel regression as local constant:

- Local linear regression estimator

$$\min_{a_0,b_0} \sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)\left(y_i - a_0 - b_0\left(x_i - x_0\right)\right)^2$$

then

$$\hat{m}(x) = \hat{a}_0 + \hat{b}_0\left(x - x_0\right)$$

- Regular kernel is local linear with b_0 constrained to be zero. We can generalize this approach to higher order polynomials.

- One particularlly popular kernel for non-parametric regression is: LOcally WEighted Scatterplot Smoothing (LOWESS) Estimator

$$K(z) = \frac{70}{81}\left(1 - |z|^3\right)^3 I(|z| < 1)$$

  where

  1. $h_{o,i}$ varies - it depends upon the distance of the point x_0 to the k^{th} nearest neighbor and

  2. observations with large residuals, $(y_i - \hat{m}(x_i))$, are downweighted as in a quasi-GLS type estimator.

- Problems with non-parametric regression:

  – Requires alot of data, especially for multi-dimensional density estimation

5 Semi-parametric Regression

- Sometimes better to combine parametric and non-parametric - where along some dimensions you know the structure or where you don't care as much if you don't know the structure. Structure reduces the curse of dimensionality as with propensity score matching. This combination of parametric and non-parametric regression is called semi-parametric regression.

- Some semi-parametric estimators:

  – Partially Linear:

  $$E\left(Y|X,Z\right)=X\beta+\lambda\left(Z\right)$$

  parameters: β, non-parametric part: λ

  – Single Index:

  $$E\left(Y|X\right)=G\left(X\beta\right)$$

  parameters: β, non-parametric part: G

– Generalized Partial Linear:

$$E(Y|X,Z) = G(X\beta + \lambda(Z))$$

parameters: β, non-parametric parts: G, λ

6 Identification: IV + Non-parametrics

1. Almost nothing done here (a few recent papers such as by Blundell and Powell).

2. Hard because you need alot of data both for IV and for non-parametrics.

3. Even more difficult if you want your instruments to be non-parametric.

7 Overview

- Positive: Non-parametric methods can be very good data description techniques since they are very flexible.

- Negative: Require alot of data.

- Negative: Difficult to do inference.

- Negative: Difficult to get good indentification.

- Net: Often good complement (not substitute) for parametric analysis.

1 Regression Discontinuity

- Standard identification problem:

$$Y_i = (1 - W_i)Y_i(0) + W_iY_i(1)$$

where W_i is a binary treatment variable.

2 Sharp Design

- Regression Discontinuity is used when there is discontinuity in assignment of treatment
  - $\exists X$ (X is a variable) such that $\exists c$ such that $W_i = 1 \Leftrightarrow X \geq c$
  - W_i is called the treatment variable
  - X is called the running variable or the forcing variable

- Assume Continuity of the Conditional Regression Function:

$$E\left[Y\left(0\right)|X=x\right] \text{ and } E\left[Y\left(1\right)|X=x\right]$$

are continuous in X.

  - Where might this assumption fail?

    * Roll call voting in small groups as opposed to anonymous voting in large electorates

    * Announced policies with cutoffs (i.e. differential tax treatment based on firm size)

- Then

$$\begin{aligned} & \lim_{x\downarrow c} E\left[Y_i|X_i=x\right] - \lim_{x\uparrow c} E\left[Y_i|X_i=x\right] \\ = & \lim_{x\downarrow c} E\left[Y_i\left(1\right)|X_i=x\right] - \lim_{x\uparrow c} E\left[Y_i\left(0\right)|X_i=x\right] \end{aligned}$$

3 Fuzzy Design

- Using discontinuity techniques in two different settings:
  - Sharp: When assignment is sharp $W_i = 1 \Leftrightarrow x \geq c$
  - Fuzzy: When assignment is fuzzy: $\lim_{x \downarrow c} \Pr[W_i | X_i = x]$ $\lim_{x \uparrow c} \Pr[Y_i | X_i = x]$
- Estimate has to be scaled by differential percent taking-up treatment: W_i

$$\frac{\lim_{x \downarrow c} E[Y_i(1) | X_i = x] - \lim_{x \uparrow c} E[Y_i(0) | X_i = x]}{\lim_{x \downarrow c} E[W_i | X_i = x] - \lim_{x \uparrow c} E[W_i | X_i = x]}$$

- Analog between Fuzzy RD, IV, and LATE estimators.

4 Estimating Regression Discontinuity Models

- Closeness Method
  - Define a definition of close to the discontinuity.
  - Regress (using OLS):
$$Y_i = \alpha + \beta D_i + \epsilon_i$$
where $D_i = 1$ if $X \geq c$ and $D_i = 0$ if $X < c$
  - Test $\beta = 0$

- Fundamental Tradeoff
  - Smaller closeness window $\implies$ Control and treatment are more similar
  - Smaller closeness window $\implies$ Less statistical power

- Polynomial Fitting Method: Motivation
  - Problem: imperfect functional form fit
  - Closeness method assumes that outcome is constant below as well as above the discontinuity
  - Polynomial method "generalizes" the closeness method

- Polynomial Fitting Method: Estimation
  - Global higher older polynomial fitting
  - Local higher order polynomial fitting
  - Estimate separately two global or local polynomial regressions
  - Predict separately from each of the two regressions, the predicted value of the outcome at the discontinuity

  - The difference in the prediction at the discontinuity is the estimate.
  - Problems
    * Non-parametrics doesn't work well with point estimation
    * Non-parametrics has low power when estimations are one sided

- Common to graph outcome as a function of the running variable with line or equation fitted on the graph and a vertical line at the discontinuity.

5 Checking RD Validity

- Test for discontinuity in variables which are not impacted by the treatment variable

  - i.e. race, sex, parental income with cutoffs in eligibility for university entrance (i.e. SAT scores)

- Can test with closeness method, polynomial fitting method, and graphically

- Same as running RD with placebo variables

- Note that you should expect an impact on variables which are impacted by Y

- Problem: if W affects both Y and Z, you may not be able to figure out if W affects Y or only affects Y indirectly through Z

6 Problems with RD

- Imperfect Functional Form Assumptions: If you dont perfectly fit the functional form of the outcome variable in the running variable, you may estimate a significant "effect" just from better fitting the functional form in the running variable

- External validity: What ability do you have to extrapolate the impact of the treatment away from the discontinuity. Good internal validity, poor external validity.

Do parties matter? I
Lee et. al.

- Determinants of legislative voting:
  - Preferences of voters (affect of voters): politician can commitment
  - Preferences of politicians (election by voters): politician can not commit

$$RC_t = \alpha + \pi_0 P_t^* + \pi_1 D_t + \varepsilon_t$$

- Where

$P_t^* =$ Electoral Strength

$D_t =$ Party of Politician in Power

$RC_t =$ Role Call Voting

Do parties matter? II

- Causal inference problem:

$$Cov\left(P_t^*, D_t\right) \neq 0$$

and

P_t^* is not observed

- However, we can estimate π_1 if we can randomize D_t by running:

$$RC_t = \alpha + \pi_1 D_t + \varepsilon_t$$

Do parties matter? III

- This allows to to calculate the degree to which voters elect rather than affect policies by electing politicians. However it does not allow us to calculate the degree to which voters affect rather than elect policies.

- Nevertheless note that:

$$E\left(RC_{t+1} \middle| D_t\right) = \alpha + \pi_0 P^*_{t+1} + \pi_1 E\left(D_{t+1} \middle| D_t\right) + E\left(\varepsilon_{t+1} \middle| D_t\right)$$

- Moreover, if D is randomly assigned, then:

$$E\left(\varepsilon_{t+1} \middle| D_t\right) = 0$$

Now, calculating the differential voting record at time t+1 given that a democrat wins at time t versus a republican

$$E\left(RC_{t+1} \middle| D_t = 1\right) - E\left(RC_{t+1} \middle| D_t = 0\right) = \pi_0\left(P^{*D}_{t+1} - P^{*R}_{t+1}\right) + \pi_1\left(P^D_{t+1} - P^R_{t+1}\right)$$

- where: $P^D_t =$ the probability of a democrat winning at time t+1 given a democrat won at time t

Do parties matter? IV

- Now (if we randomize the election: D) we can calculate:
  - (1.) The degree to which voters elect policies

$$RC_t = \alpha + \pi_1 D_t + \varepsilon_t$$

  - (2.) The probability of that a Democrat wins an election in an electoral district given that a Democrat won the prior election:

$$E\left(D_{t+1} \middle| D_t = 1\right) - E\left(D_{t+1} 0\right) = P_{t+1}^{D} - P_{t+1}^{R}$$

- We can also estimate the effect of a democrat getting elected at date t on policy at date t+1:

$$\gamma = E\left(RC_{t+1} \middle| D_t = 1\right) - E\left(RC_{t+1} \middle| D_t = 0\right) = \pi_0\left(P_{t+1}^{*D} - P_{t+1}^{*R}\right) + \pi_1\left(P_{t+1}^{D} - P_{t+1}^{R}\right)$$

- Thus we can calculate the degree to which citizens affect policies (just the residual):

$$\gamma - \pi_1\left(P_{t+1}^{D} - P_{t+1}^{R}\right)$$

Do parties matter? V

- How do we acheive randomization of D?
- Regression Discontinuity: Two Approaches
  - (1.) Look at Close Elections (<2% vote margin of victory):

$$E\left(RC_t \middle| D_t = 1\right) - E\left(RC_t \middle| D_t = 0\right) = \pi_1$$

$$E\left(D_{t+1} \middle| D_t = 1\right) - E\left(D_{t+1} \middle| D_t = 0\right) = P_{t+1}^{D} - P_{t+1}^{R}$$

$$E\left(RC_{t+1} \middle| D_t = 1\right) - E\left(RC_{t+1} \middle| D_t = 0\right) = \gamma$$

Do parties matter? VI

- Regression Discontinuity: Two Approaches (continued)
  - (2.) Look at Polynomial fits in the vote share before and after the discontinuity and test for equality at the discontinuity:

$$E\left(RC_t \middle| D_t = 1\right) - E\left(RC_t \middle| D_t = 0\right) =$$
$$D_t\left[\alpha_0 + \alpha_1 V_t + \alpha_2 V_t^2 + \alpha_3 V_t^3 + \alpha_4 V_t^4\right] +$$
$$\left(1 - D_t\right)\left[\beta_0 + \beta_1 V_t + \beta_2 V_t^2 + \beta_3 V_t^3 + \beta_4 V_t^4\right]$$

    - The estimate is then

$$E\left(RC_t \middle| D_t = 1, V_t = .5\right) - E\left(RC_t \middle| D_t = 0, V_t = .5\right)$$

Do parties matter? VII

- Data:
  - Dependent Variable
    - Democratic Two-Party Vote Share from House of Representatives Elections
  - Independent Variable
    - ADA Score (weighted measure of liberalness based on 20 key votes every year)
    - Nominate and DW-Nominate (Rosenthal and Poole)
    - Measures of Party Loyalty
    - Measures by Interest Groups (Unions, Christian Groups, etc...)

Do parties matter? VII

- Problems:
  - Identification:
    - (1.) Close Elections: We dont know how close is close. A narrower definition of close election leads to better identification but less precision and less external validity.
      - Hahn, Todd, and Van Der Klaauw (Econometrica, Jan. 2001): Regression discontinuity as non-parametric estimator (remaining optimal bandwidth problem)
    - (2.) Polynomial Fitting: We dont know the functional form of the polynomial in the vote share. If we get it wrong, we may estimate an effect just due to poor fitting of the polynomial.
      - Solution: monte carlo selection of placebo discontinuity points
    - (3.) General Problem: How do we know that there isnt selection around the discontinuity (i.e. firm size regulations)
      - (a.) institutional knowledge (i.e. small committee elections with publicly observed votes versus general elections)
      - (b.) empirical verification that there is no selection around the discontinuity using other variables (i.e. David Lee paper, Jason Snyder paper)

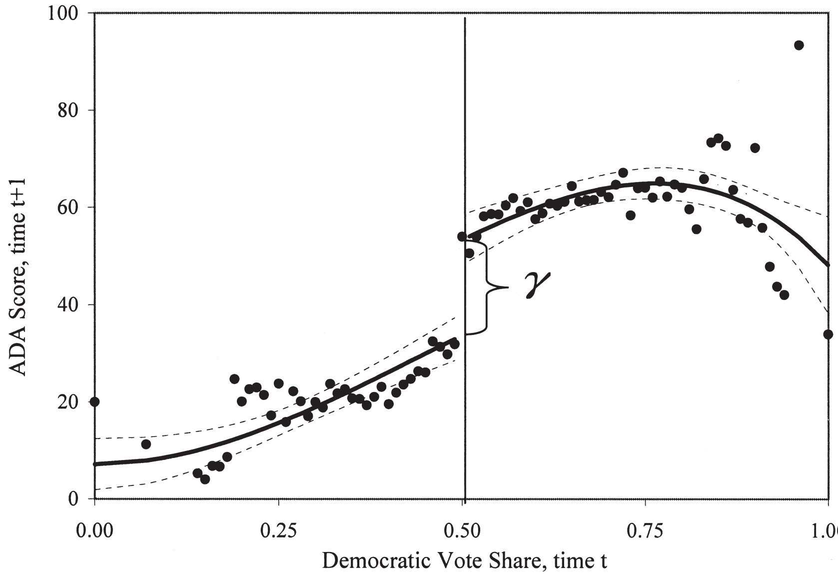

FIGURE I
Total Effect of Initial Win on Future ADA Scores: γ

This figure plots ADA scores after the election at time $t + 1$ against the Democrat vote share, time t. Each circle is the average ADA score within 0.01 intervals of the Democrat vote share. Solid lines are fitted values from fourth-order polynomial regressions on either side of the discontinuity. Dotted lines are pointwise 95 percent confidence intervals. The discontinuity gap estimates

$$\gamma = \underbrace{\pi_0(P^{*D}_{t+1} - P^{*R}_{t+1})}_{\text{“Affect”}} + \underbrace{\pi_1(P^{*D}_{t+1} - P^{*R}_{t+1})}_{\text{“Elect”}}.$$

be a continuous and smooth function of vote shares everywhere, except at the threshold that determines party membership. There is a large discontinuous jump in ADA scores at the 50 percent threshold. Compare districts where the Democrat candidate barely lost in period t (for example, vote share is 49.5 percent), with districts where the Democrat candidate barely won (for example, vote share is 50.5 percent). If the regression discontinuity design is valid, the two groups of districts should appear ex ante similar in every respect—on average. The difference will be that in one group, the Democrats will be the incumbent for the next election ($t + 1$), and in the other it will be the Republicans. Districts where the Democrats are the incumbent party for election $t + 1$ elect representatives who have much higher ADA scores, compared with districts where the Republican candidate

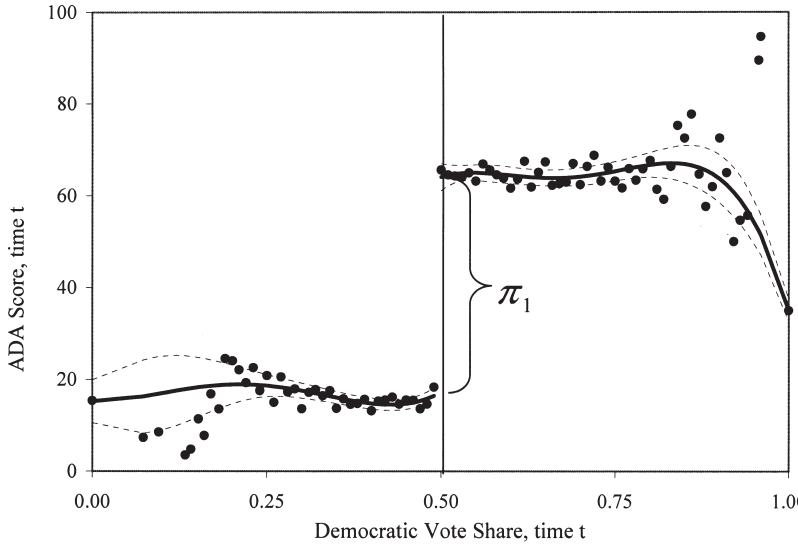

FIGURE IIa
Effect of Party Affiliation: π_1

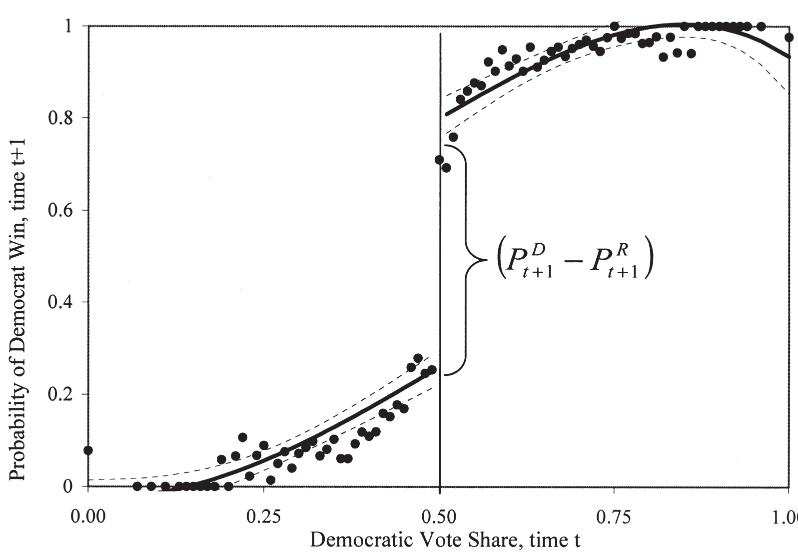

FIGURE IIb
Effect of Initial Win on Winning Next Election: $(P^D_{t+1} - P^R_{t+1})$

Top panel plots ADA scores after the election at time t against the Democrat vote share, time t. Bottom panel plots probability of Democrat victory at $t + 1$ against Democrat vote share, time t. See caption of Figure III for more details.

TABLE I
RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

	Total effect			Elect component	Affect component
	γ	π_1	$(P^D_{t+1} - P^R_{t+1})$	$\pi_1[(P^D_{t+1} - P^R_{t+1})]$	$\pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}]$
Variable	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. (2)*(col. (3))	(col. (1)) − (col. (4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2	47.6	0.48		
	(1.9)	(1.3)	(0.02)		
				22.84	−1.64
				(2.2)	(2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

primarily elect policies (full divergence) rather than affect policies (partial convergence).

Here we quantify our estimates more precisely. In the analysis that follows, we restrict our attention to "close elections"—where the Democrat vote share in time t is strictly between 48 and 52 percent. As Figures I and II show, the difference between barely elected Democrat and Republican districts among these elections will provide a reasonable approximation to the discontinuity gaps. There are 915 observations, where each observation is a district-year.[20]

Table I, column (1), reports the estimated total effect γ, the size of the jump in Figure I. Specifically, column (1) shows the difference in the average ADA_{t+1} for districts for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and districts for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. The estimated difference is 21.2.

In column (2) we estimate the coefficient π_1, which is equal to the size of the jump in Figure IIa. The estimate is the difference in the average ADA_t for districts for which the Democrat vote

20. In 68 percent of cases, the representative in period $t + 1$ is the same as the representative in period t. The distribution of close elections is fairly uniform across the years. In a typical year there are about 40 close elections. The year with the smallest number is 1988, with twelve close elections. The year with the largest number is 1966, with 92 close elections.

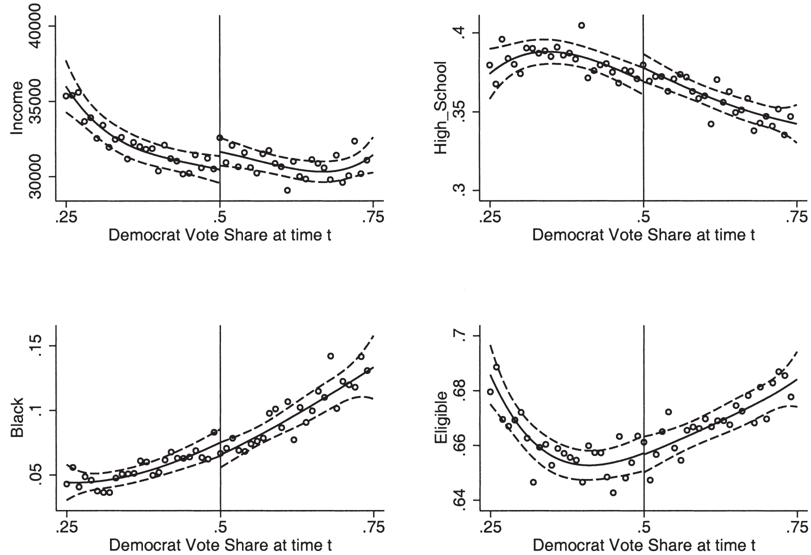

FIGURE III

Similarity of Constituents' Characteristics in Bare Democrat and Republican Districts–Part 1

Panels refer to (from top left to bottom right) the following district characteristics: real income, percentage with high-school degree, percentage black, percentage eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democrat vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval.

share. The coefficient reported in column (6) is the predicted difference at 50 percent. The table confirms that, for many observable characteristics, there is no significant difference in a close neighborhood of 50 percent. One important exception is the percentage black, for which the magnitude of the discontinuity is statistically significant.[23]

As a consequence, estimates of the coefficients in Table I from regressions that include these covariates would be expected to produce similar results—as in a randomized experiment—since

23. This is due to few outliers in the outer part of the vote share range. When the polynomial is estimated including only districts with vote share between 25 percent and 75 percent, the coefficients becomes insignificant. The gap for percent urban and open seats, while not statistically significant at the 5 percent level, is significant at the 10 percent level.

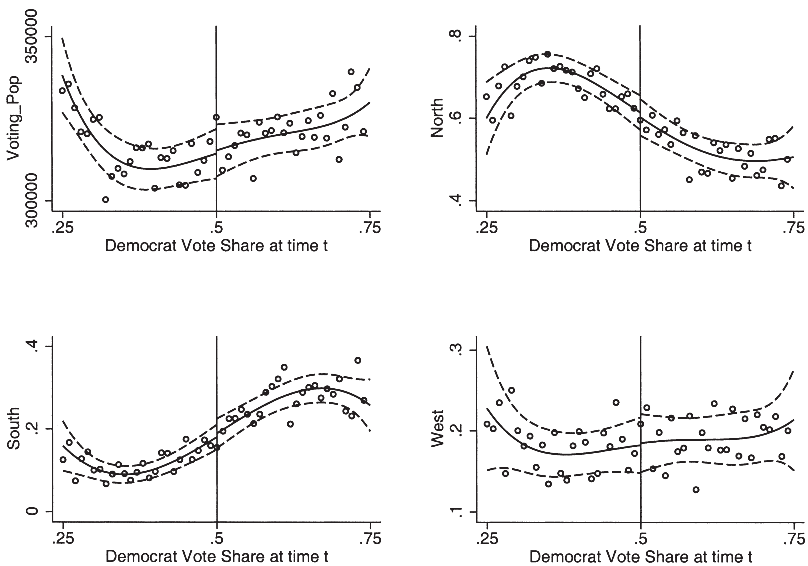

FIGURE IV

Similarity of Constituents' Characteristics in Bare Democrat and Republican Districts—Part 2

Panels refer to (from top left to bottom right) the following district characteristics: voting population, North, South, West. Circles represent the average characteristic within intervals of 0.01 in Democrat vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval.

all predetermined characteristics appear to be orthogonal to D_t. We have reestimated all the models in Table I conditioning on all of the district characteristics in Table II, and found estimates that are virtually identical to the ones in Table I.

As a similar empirical test of our identifying assumption, in Figure V we plot the ADA scores from the Congressional sessions that *preceded* the determination of the Democratic two-party vote share in election t. Since these past scores have already been determined by the time of the election, it is yet another predetermined characteristic (just like demographic composition, income levels, etc.). If the RD design is valid, then we should observe no discontinuity in these lagged ADA scores—just as we would expect, in a randomized experiment, to see no systematic differences in any variables determined prior to the experiment. The

TABLE II
DIFFERENCE IN DISTRICT CHARACTERISTICS BETWEEN DEMOCRAT AND REPUBLICAN DISTRICTS, BY DISTANCE FROM 50 PERCENT

	All (1)	+/− 25 (2)	+/− 10 (3)	+/− 5 (4)	+/− 2 (5)	Polynomial (6)
North	−0.211	−0.156	−0.096	−0.054	−0.059	−0.041
	(0.018)	(0.019)	(0.021)	(0.024)	(0.036)	(0.045)
South	0.250	0.145	0.093	0.053	0.009	0.015
	(0.015)	(0.014)	(0.016)	(0.019)	(0.028)	(0.036)
West	−0.031	−0.012	−0.036	−0.003	0.001	0.001
	(0.013)	(0.015)	(0.020)	(0.017)	(0.020)	(0.036)
Log income	−0.086	−0.036	0.014	0.026	0.030	0.052
	(0.013)	(0.012)	(0.014)	(0.017)	(0.026)	(0.033)
Percentage high-school grad.	−0.035	−0.024	−0.008	−0.001	0.001	0.008
	(0.003)	(0.003)	(0.004)	(0.004)	(0.007)	(0.008)
Percentage urban	0.070	0.065	0.053	0.053	0.056	0.053
	(0.011)	(0.011)	(0.012)	(0.014)	(0.023)	(0.028)
Percentage black	0.082	0.042	0.013	0.003	−0.003	−0.053
	(0.005)	(0.004)	(0.004)	(0.005)	(0.009)	(0.013)
Manufacturing employment	−0.002	0.000	0.004	0.004	0.005	0.003
	(0.001)	(0.001)	(0.002)	(0.002)	(0.004)	(0.005)
Total population	−1817.9	3019.2	4961.5	3211.4	8640.4	2007.5
	(3517.3)	(3723.0)	(4562.4)	(5524.2)	(8427.9)	(10483.0)
Percentage eligible to vote	0.005	0.010	0.007	0.006	−0.003	−0.003
	(0.002)	(0.002)	(0.003)	(0.004)	(0.006)	(0.007)
Open seats	0.070	0.065	0.053	0.053	0.056	0.053
	(0.011)	(0.011)	(0.012)	(0.014)	(0.023)	(0.028)
Number of observations	13413	10229	4174	2072	910	13413

Standard errors are in parentheses. The unit of observation is a district-congressional session. Columns (1) to (5) report the difference in average district characteristics between Democrat and Republican districts. Column (1) includes the entire sample. Columns (2) to (5) include only districts with Democrat vote share between 25 percent and 75 percent, 40 percent and 60 percent, 45 percent and 55 percent, and 48 percent and 52 percent, respectively. The model in column (6) includes a fourth-order polynomial in Democrat vote share that enters separately for vote share above and below 50 percent. The coefficient reported in column (6) is the predicted difference at 50 percent. All standard errors account for district-decade clustering.

lack of discontinuity in the figure lends further credibility to our identifying assumption.[24]

Overall, the evidence strongly supports a valid regression discontinuity design. And as a consequence, it appears that among close elections, who wins appears virtually randomly assigned, which is the identifying assumption of our empirical strategy.

24. The estimated gap is 3.5 (5.6).

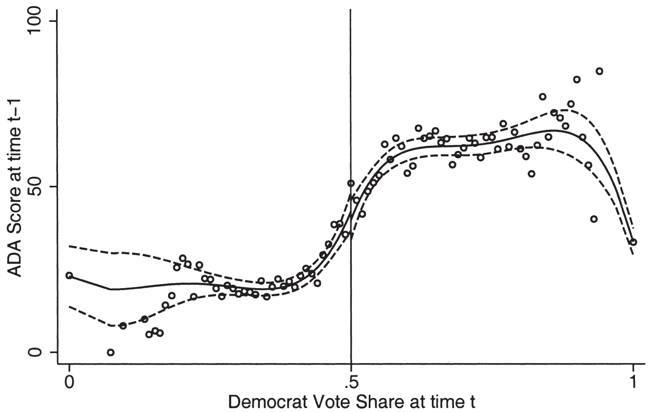

FIGURE V

Specification Test: Similarity of Historical Voting Patterns between Bare Democrat and Republican Districts

The panel plots one time lagged ADA scores against the Democrat vote share. Time t and $t - 1$ refer to congressional sessions. Each point is the average lagged ADA score within intervals of 0.01 in Democrat vote share. The continuous line is from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line is the 95 percent confidence interval.

V.C. Sensitivity to Alternative Measures of Voting Records

Our results so far are based on a particular voting index, the ADA score. In this section we investigate whether our results generalize to other voting scores. We find that the findings do not change when we use alternative interest groups scores, or other summary measures of representatives' voting records.

Table III is analogous to Table I, but instead of using ADA scores, it is based on two alternative measures of roll-call voting. The top panel is based on McCarty, Poole, and Rosenthal's DW-NOMINATE scores. The bottom panel is based on the percent of individual roll-call votes cast that are in agreement with the Democrat party leader. All the qualitative results obtained using ADA scores (Table I) hold up using these measures. When we use the DW-NOMINATE scores, γ is -0.36, remarkably close to the corresponding estimate of $\pi_1[P^D_{t+1} - P^R_{t+1}]$ in column (4), which is -0.34. The estimates are negative here because, unlike ADA scores, higher Nominate scores correspond to a more conservative voting record. When we use the measure "percent voting with the Democrat leader," γ is 0.13, almost indistinguishable from the

TABLE III
RESULTS BASED ON NOMINATE SCORES AND ON PERCENT VOTED LIKE DEMOCRAT LEADERSHIP—CLOSE ELECTIONS SAMPLE

Variable	Total effect γ Z_{t+1} (1)	π_1 Z_t (2)	$(P^D_{t+1} - P^R_{t+1})$ DEM_{t+1} (3)	Elect component $\pi_1[(P^D_{t+1} - P^R_{t+1})]$ (col. (2)*(col. (3)) (4)	Affect component $\pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}]$ (col. (1)) − (col. (4)) (5)
(a) Results based on Nominate scores					
Estimated gap	−0.36	−0.58	0.62		
	(0.03)	(0.02)	(0.04)		
				−0.34	−0.02
				(0.04)	(0.04)
(b) Results based on percent voted like Democrat leadership					
Estimated gap	0.13	0.29	0.46		
	(0.01)	(0.006)	(0.02)		
				0.13	0.00
				(0.02)	(0.02)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. The top panel uses the DW-NOMINATE score constructed by McCarty, Poole, and Rosenthal. Higher Nominate scores correspond to more conservative roll-call voting records. The bottom panel uses the percent of a representative's votes that agree with the Democrat party leader. Sample size is 276 in top panel and 1010 in bottom panel.

estimate $\pi_1[P^D_{t+1} - P^R_{t+1}]$ in column (4), which is 0.13. We show the graphical analysis for the estimate of π_1 in Figure VI.

Our empirical findings are also not sensitive to the use of ratings from various liberal and conservative interest groups. Liberal interest groups include the American Civil Liberties Union, the League of Women Voters, the League of Conservation Voters, the American Federation of Government Employees, the American Federation of State, County, and Municipal Employees, the American Federation of Teachers, the AFL-CIO Building and Construction, and the United Auto Workers. Conservative groups include the Conservative Coalition, the U. S. Chamber of Commerce, the American Conservative Union, and the Christian Voice. All the ratings range from 0 to 100. For liberal groups, low ratings correspond to conservative roll-call votes, and high ratings correspond to liberal roll-call votes. For conservative groups the opposite is true.

These alternative ratings yield results that are qualitatively similar to our findings in Table I and III. Instead of presenting these results in a table format as we did in Table I and III, we present the main results in graphical form. We summarize our

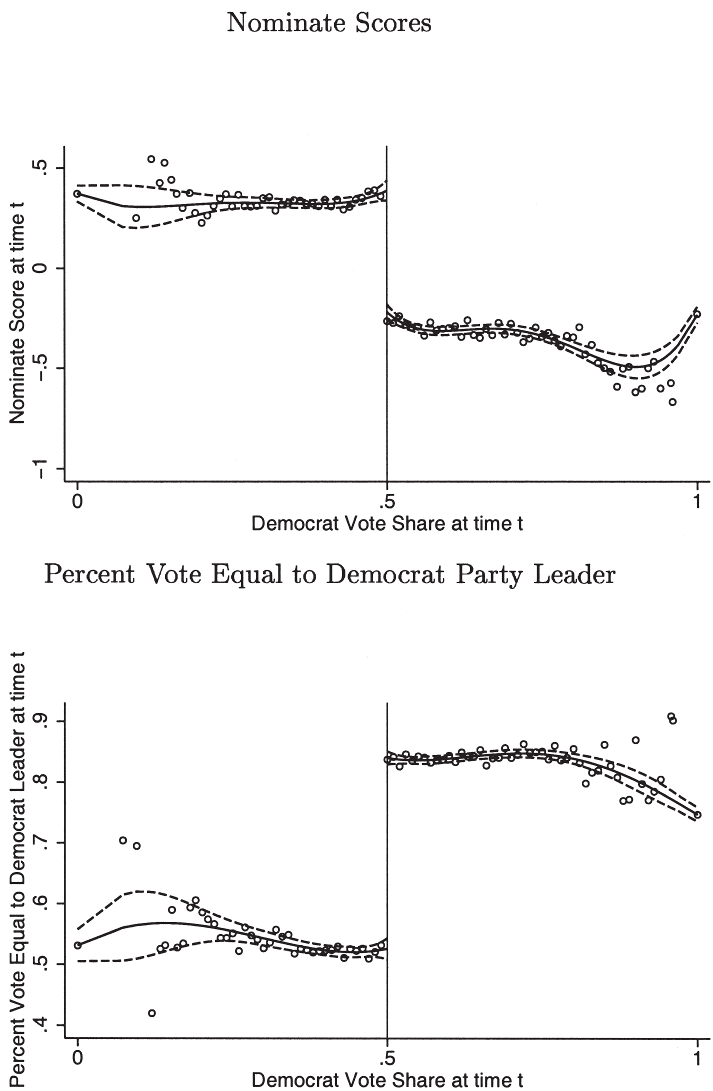

FIGURE VI

Nominate Scores, by Democrat Vote Share; and Percent Voted with Democrat Leader, by Democrat Vote Share

The top panel plots DW-Nominate scores at time *t* against the Democrat vote share at time *t*. Circles represent the average Nominate score within intervals of 0.01 in Democrat vote share. The bottom panel plots the fraction of a Representative's votes that agree with the Democrat party leader at time *t* against the Democrat vote share at time *t*. Circles represent the percent voted with Democrat leader within intervals of 0.01 in Democrat vote share. The continuous line is from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line is the 95 percent confidence interval.

TABLE IV
RESULTS BASED ON ADA SCORES, BY DECADE—CLOSE ELECTIONS SAMPLE

	(1) Total effect	(2)	(3)	(4) Elect component	(5) Affect component
	γ	π_1	$(P^D_{t+1} - P^R_{t+1})$	$\pi_1[(P^D_{t+1} - P^R_{t+1})]$	$\pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}]$
Variable	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. (2)*(col. (3))	(col. (1)) − (col. (4))
1946–1958	14.2	41.7	0.41		
	(3.2)	(2.3)	(0.05)		
				17.0	−2.8
				(4.8)	(4.0)
1960–1968	23.5	49.5	0.51		
	(3.5)	(2.7)	(0.05)		
				25.2	−1.7
				(4.9)	(4.1)
1970–1978	11.5	46.6	0.40		
	(4.7)	(3.1)	(0.06)		
				18.6	−7.1
				(5.1)	(5.1)
1980–1996	46.8	56.6	0.76		
	(3.7)	(2.8)	(0.05)		
				43.0	3.8
				(4.9)	(4.5)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample sizes are 322 in 1946–1958; 245 in 1960–1968; 183 in 1970–1978; 164 in 1980–1996.

VI. RELATION TO PREVIOUS EMPIRICAL LITERATURE

A number of empirical studies have directly or indirectly examined the policy convergence issue.[28] Typically, the studies examine whether party affiliation matters for the observed voting records of the legislator. Most studies find evidence of this, which is strictly inconsistent with the *complete* policy convergence result. For example, Poole and Rosenthal [1984] show that senators from the same state belonging to different parties have significantly different voting records.

28. An example of early empirical work in this area is Miller and Stokes [1963]. The literature is too large to be summarized here. Other examples include, but are not limited to, Snyder and Ting [2001a], Fiorina [1999], Poole and Rosenthal [2001], Snyder and Ting [2001b], Lott and Davis [1992], Canes-Wrone, Brady, and Cogan [2002], Krehbiel [2000], Bender [1991], McArthur and Marks [1988], and McCarty, Poole, and Rosenthal [2000].

Do parties matter? VIII

- Problems:
  - Interpretation & External Validity
    - Benefit of Approach: Clean identification
    - Cost of Approach: Small Sample
      - Heterogeneous Treatment Effects and External Validity
      - Statistical Power

- Specific Problems with Lee et. Al. Paper
  - Estimates of electing policies: clean
  - Estimates of affecting policies: not so clean
  - Other identification problems
    - impact on composition of legislature and thus on what bills are voted in?
    - do we care about voting in the legislature? maybe the only difference in voting patterns is for votes which are sufficiently lop-sided that the differences are policy irrelevant?

1 Local Party Effects

- Big differences between parties at national level. What about local level?

- Would we expect to see the same results? Tiebout?

- Gyourko and Ferreira redid Lee, Moretti and Butler with cities (and with budget size rather than voting in the legislature): their answer - no party effects at the local level

- Per Petterson-Lidbom (JEEA) did Lee, Moretti and Butler with budget expenditures in Sweden and Finland and got positive impacts of expenditure by left wing parties. Why is there difference? Greater mobility in Sweden/Finland?

- Tried to collect data from all cities over 25,000 in population: 877 have an elected mayor (as opposed to a city manager appointed by local city councils). 413 cities responded to the survey.

TABLE I
SAMPLE REPRESENTATIVENESS

	Final sample (1)	All U.S. cities (2)	Cities with >25,000 population (3)	Cities with >25,000, elected mayor (4)	Cities >25,000, elected mayor, survey reply (5)
Number of cities	413	34,574	1,893	877	498
Population	126,364	7,666	86,245	112,392	113,104
	(256,768)	(62,732)	(255,000)	(346,409)	(234,874)
% west	0.20	0.12	0.24	0.18	0.21
	(0.40)	(0.33)	(0.42)	(0.39)	(0.41)
% south	0.26	0.24	0.24	0.25	0.29
	(0.44)	(0.43)	(0.43)	(0.44)	(0.45)
% north	0.16	0.23	0.25	0.23	0.14
	(0.36)	(0.42)	(0.43)	(0.42)	(0.35)
% white	0.68	0.88	0.75	0.69	0.68
	(0.22)	(0.20)	(0.19)	(0.23)	(0.22)
% black	0.13	0.05	0.11	0.13	0.12
	(0.17)	(0.14)	(0.16)	(0.17)	(0.17)
% college degree	0.26	0.17	0.28	0.26	0.26
	(0.13)	(0.13)	(0.15)	(0.13)	(0.13)
Median family income	$53,035	$46,916	$57,927	$53,334	$53,428
	(16,202)	(19,262)	(19,566)	(16,687)	(16,265)
Median house value	$132,622	$100,526	$156,718	$133,838	$134,067
	(66,582)	(86,412)	(100,769)	(70,988)	(66,961)

Notes. All variables are based on the 2000 Census. Standard deviations in parentheses. Column (1) presents descriptives for the mayoral election sample used in this paper. Column (2) reports descriptives for all cities in the United States. Column (3) restricts the sample to cities with more than 25,000 people as of year 2000. Column (4) additionally constrains the sample to cities that directly elect a mayor. Column (5) presents results for cities that replied to the survey with vote totals but no information about party affiliation. See the text for additional details.

some information on vote totals and candidate names for 57% of the 877 cities that elect mayors by popular vote. Summary statistics for this group of 498 places are displayed in column (5). Our final sample of 413 cities, which is 47% of those places that directly elect a mayor, also contains information on party affiliation, not just vote totals.[7]

7. Two factors made it difficult to collect information on candidates' party affiliations, even when we knew who they were and how many people voted for them. First, some cities and counties could not provide the data because this required gathering information from inaccessible voter registration records. Second, there is a large fraction of cities (59% as of year 2000) that are institutionally nonpartisan in that they prohibit party labels from being printed on election ballots or used in election campaigning. This certainly does not mean that nearly 60% of mayoral races literally had no partisan content. A quick review showed that elections in many such cities (e.g., Los Angeles, California) clearly were partisan in the standard use of that term. Hence, survey information was complemented

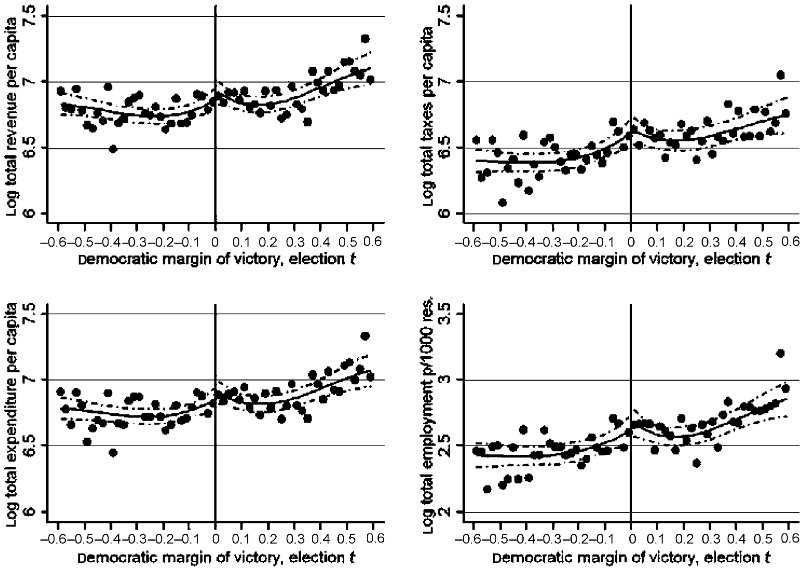

FIGURE II
Party Effect on Size-of-Government Measures

a Democrat rather than a Republican, so there is no evidence of any robust partisan impact on crime (or any other variable).

Because pictures often are illuminating in a regression discontinuity context, Figure II graphs the results for each size of government outcome. Each dot in a panel corresponds to the average outcome that follows election *t*, given the margin of victory obtained by Democrats in election *t*. The solid line in the figure represents the predicted values from the cubic polynomial fit without covariates as described in equation (1), with the dashed lines identifying the 95% confidence intervals. Visual inspection confirms that there always is a positive correlation between size of government and Democratic margin of victory, but there never are any significant discontinuities around the close election breakpoint for any revenue, tax, spending, or employment outcome.[13]

13. Figures for the composition of expenditures and crime rates show similar patterns. We also performed a formal test of political divergence as in Lee, Moretti, and Butler (2004). Given the very small partisan effects reported in Table II, it is not surprising that we cannot reject the conclusion of political convergence, that is, that it is local voters, not the political parties, who are determining policy outcomes. This is in stark contrast to Lee, Moretti, and Butler's conclusion that, for congressional representatives at the federal level, voters simply are choosing one party's bliss point. See our NBER working paper for those results.

TABLE II
OLS AND RD ESTIMATES OF THE IMPACT OF A DEMOCRATIC MAYOR

		% diff. between Dem and Rep mayors			
Dependent variables	Average (std) (1)	OLS uncond. (2)	OLS conditional (3)	RD cubic (4)	RD linear (5)
		Size of government			
Total revenues per capita ($)	1,082	0.129	0.058	−0.016	−0.014
	(676)	(0.029)	(0.022)	(0.022)	(0.013)
Total taxes per capita ($)	852	0.160	0.091	−0.013	0.008
	(678)	(0.033)	(0.024)	(0.021)	(0.012)
Total expenditures	1,067	0.131	0.060	−0.009	−0.015
per capita ($)	(652)	(0.029)	(0.022)	(0.021)	(0.013)
Total employment per	15.25	0.169	0.087	0.017	0.014
1,000 residents	(9.52)	(0.035)	(0.028)	(0.016)	(0.011)
		Allocation of resources			
% spent on salaries and wages	0.61	0.007	0.012	0.020	0.007
	(0.12)	(0.006)	(0.006)	(0.014)	(0.008)
% spent on police department	0.20	−0.011	−0.003	−0.001	0.003
	(0.08)	(0.004)	(0.004)	(0.007)	(0.004)
% spent on fire department	0.13	−0.004	−0.001	0.006	0.006
	(0.05)	(0.003)	(0.003)	(0.005)	(0.003)
% spent on parks	0.19	−0.023	−0.009	0.011	0.009
and recreation	(0.17)	(0.009)	(0.007)	(0.014)	(0.009)
		Crime indices			
Murders per 1,000 residents	0.08	0.019	0.008	0.005	0.011
	(0.09)	(0.006)	(0.004)	(0.007)	(0.005)
Robberies per 1,000 residents	2.06	0.824	0.454	0.597	0.619
	(3.70)	(0.200)	(0.186)	(0.338)	(0.288)
Burglaries per 1,000 residents	15.54	0.948	0.194	0.572	1.579
	(12.40)	(0.780)	(0.732)	(1.024)	(0.735)
Larcenies per 1,000 residents	41.49	1.923	1.389	1.798	5.424
	(27.81)	(1.718)	(1.700)	(2.489)	(1.869)
Covariates		No	Yes	Yes	Yes

Notes. Column (1) presents averages and standard deviations for all independent variables, while Columns (2)–(5) report coefficients from OLS and RD regressions of each independent variable indicated in the table on an indicator variable for whether the mayor is a Democrat and other controls. The RD specification also has other controls for margin of victory as described in equation (1) in the text. All size-of-government variables were transformed to logs. The set of covariates includes city population, the type of election (partisan versus nonpartisan, length of term status), median income, percentage of white households, percentage of households with college degrees, homeownership rate, and median house value. Year and region fixed effects also are included. Columns (4) and (5) also include a control for the respective dependent variable at the year prior to the election. See the text for a more detailed explanation of the fiscal and crime variables. The numbers of observations for total employment and crime indices are 1,463 and 1,720, respectively, whereas 1,886 is the relevant number for all other variables. Reported standard errors are clustered by city and decade.

TABLE III
COMPARISONS BETWEEN CITIES AND CONGRESSIONAL DISTRICTS

	All cities		Cities >25,000 pop.		Congressional districts	
	Mean	Median	Mean	Median	Mean	Median
Number of jurisdictions	34,574		1,893		435	
Population	7,666	1,423	86,245	43,858	645,377	633,102
Income heterogeneity	0.18	0.16	0.33	0.32	0.43	0.41
Political heterogeneity	0.15	0.10	0.26	0.21	0.36	0.30
Number of newspapers	7.5	1.0	21.2	11.0	46.9	33.0
Herfindahl of newspapers	0.34	0.22	0.39	0.33	0.14	0.11

Notes. Number of jurisdictions and population are based on the 2000 Census. The income heterogeneity measure is based on the coefficient of variation for income that is calculated using block group mean and median incomes from the 2000 Census for the entire country. The political heterogeneity measure also is measured by the coefficient of variation based on the precinct level vote share for Bush in the 2000 presidential election. Voting precincts could be only accurately mapped to municipalities for the following states: CA, CT, IL, IN, MA, ME, NH, NJ, NY, OH, PA, RI, VT, WI, CO, DE, GA, HI, KS, MD, MN, NC, OK, TX, VA, WA, and WV. The number of newspapers is based on Burrele's Media Directory for the year 2000. The Herfindahl of newspapers is based on the circulation shares of all local and regional newspapers in a city. See the text for additional details.

diversity. The degree of income heterogeneity is measured by the standard deviation of all block group average family incomes (as of the year 2000) in a city, divided by the overall city mean family income. For political heterogeneity, the analogous statistic is computed for the proportion of Bush voters in the 2000 election using data at the precinct level that were mapped to city boundaries.[18] The median city is less heterogeneous than the median congressional district, whether one measures diversity along income or political lines. The difference is most stark with respect to income, where the median city's coefficient of variation is less than 40% that of the median congressional district (0.16/0.41$\sim$0.39 from row (3)). The difference is less great, but still quite apparent, if one looks only at larger communities with at least 25,000 residents (0.32/0.41$\sim$0.77). And this conclusion does not change if one uses means that reflect the influence of a few very large cities.

To help understand the diversity of media outlets at the local level, we gathered data on the number and type of newspapers within the metropolitan area. Row (5) of Table III reports the

18. Data limitations allowed us to compute this measure of political diversity in cities and congressional districts in only 27 states. A lack of reliable election data for small geographic areas in many parts of the county is the primary reason, but some states also do not have useful geographic identifiers for their voting precincts.

TABLE IV
TESTS OF LOCAL NONPARTISANSHIP MECHANISMS

Dependent variables	Total revenues >Median (1)	Total revenues >75% (2)	Total expenditures >Median (3)	Total expenditures >75% (4)	Total employment >Median (5)	Total employment >75% (6)
	Tiebout sorting					
Dummy for high income heterogeneity	−0.007	0.016	0.017	−0.007	−0.048	−0.033
	(0.035)	(0.032)	(0.033)	(0.035)	(0.027)	(0.031)
	Tiebout competition					
Dummy for high Herfindahl index	0.045	0.088	0.030	0.086	0.062	0.068
	(0.034)	(0.042)	(0.032)	(0.038)	(0.026)	(0.029)
	Strategic extremism					
Dummy for more newspapers	0.005	−0.009	−0.024	−0.024	0.017	0.012
	(0.033)	(0.039)	(0.032)	(0.038)	(0.026)	(0.033)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,854	1,854	1,854	1,854	1,351	1,351

Notes. All columns present RD coefficient estimates where each fiscal policy outcome was regressed on an indicator for Democratic victory in election t interacted with each of the three mechanisms that might explain the lack of partisanship at the local level, and other controls as described in equation (3). In the odd-numbered columns, each mechanism variable is generated as an indicator for cities that are above the median value of respective variable; in the even-numbered columns, indicators are for cities that are above the 75th percentile. All size of government variables were transformed to logs. Reported standard errors are clustered by city and decade.

indicator are most supportive of the hypothesis that it is Tiebout competition from communities within the metropolitan area that primarily disciplines partisan behavior at the local level of government. However, only one of those coefficients is statistically significant at standard confidence levels. The results in columns (2), (4), and (6), based on the 75th percentile indicator, provide stronger support. Each is statistically significantly different from zero at the 95% confidence level, and the magnitudes seem plausible. For example, estimates for total revenue per capita (column (2)) indicate that a community with so few (population-weighted) other local governments in its metropolitan area as to put it in the top quartile of our Herfindahl index for this constraint proxy raises nearly 9% more revenue if it barely elected a Democrat as its mayor than otherwise similar cities that also barely elected a Democrat but have Herfindahl values below the 75th percentile cutoff. That expenditures per capita are higher by virtually the

partisan impacts. Unfortunately, those data were quite noisy because unionization rates could only be measured at the county, not the city, level.