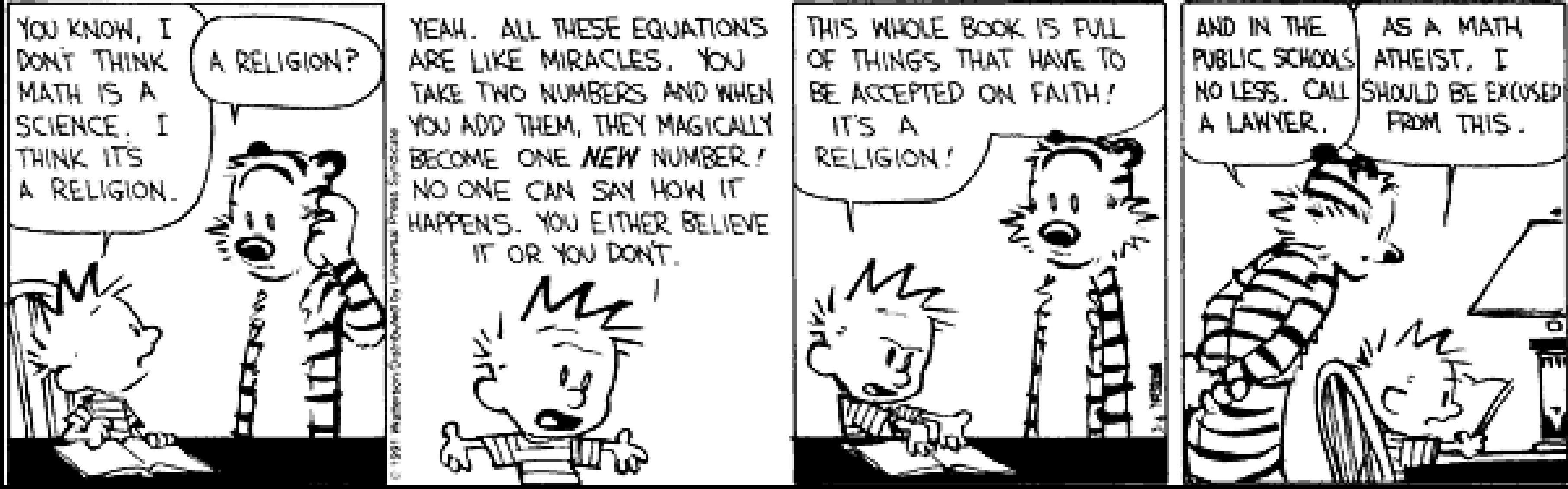

YOU KNOW, I DON'T THINK MATH IS A SCIENCE. I THINK IT'S A RELIGION.
A RELIGION?
YEAH. ALL THESE EQUATIONS ARE LIKE MIRACLES. YOU TAKE TWO NUMBERS AND WHEN YOU ADD THEM, THEY MAGICALLY BECOME ONE *NEW* NUMBER! NO ONE CAN SAY HOW IT HAPPENS. YOU EITHER BELIEVE IT OR YOU DON'T.
© 1991 Watterson/Distributed by Universal Press Syndicate
THIS WHOLE BOOK IS FULL OF THINGS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND IN THE PUBLIC SCHOOLS NO LESS. CALL A LAWYER.
AS A MATH ATHEIST, I SHOULD BE EXCUSED FROM THIS.

INFO 251: Applied Machine Learning

Nonexperimental Methods: Regression Discontinuity

Course Outline

- Causal Inference and Research Design
  - Experimental methods
  - **Non-experiment methods**
- Machine Learning
  - Design of Machine Learning Experiments
  - Linear Models and Gradient Descent
  - Non-linear models
  - Neural models
  - Unsupervised Learning
  - Practicalities, Fairness, Bias
- Special topics

Key Concepts (previous lecture)

- Conditional exogeneity
- Instrumental variables
- First Stage
- Second Stage
- Reduced Form
- Exclusion restriction
- Instrument relevance

Today's Outline and Key Concepts

- Regression Discontinuity
  - Motivation and intuition
  - Regression discontinuity
  - Example: Graphical Analysis
  - Running variables
  - Estimation
  - Examples
- Econometrics Summary / Exercise

Regression Discontinuity: Context

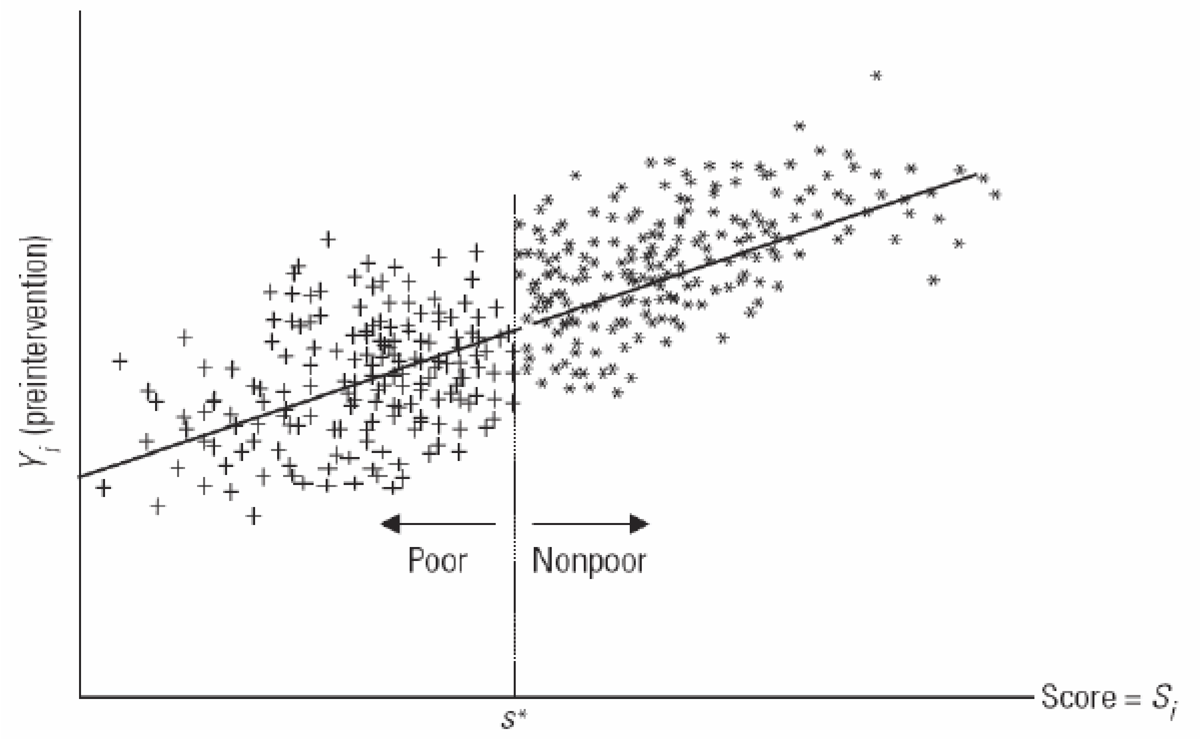
- Randomization is the cleanest way to identify causal effects
  - But randomization is not always feasible. So while we can imagine the “ideal experiment”, that rarely happens
- **Instrumental Variables:** looks for factors that create quasi-random variation in a treatment variable
- **Regression Discontinuity**: Exploits convenient discontinuities in treatment assignment

Regression Discontinuity: Intuition

- Many treatments are assigned based on an index or score
  - **Vaccines**: Everyone older than a threshold age is eligible
  - **Promotions**: Targeted to people who generate more than a threshold value of revenue per month
  - **Progresa**: Targeted to households below a wealth threshold
  - **Education**: Scholarships for students who score above a threshold
- Key: Treatment (or treatment eligibility) is quasi-random close to the discontinuity

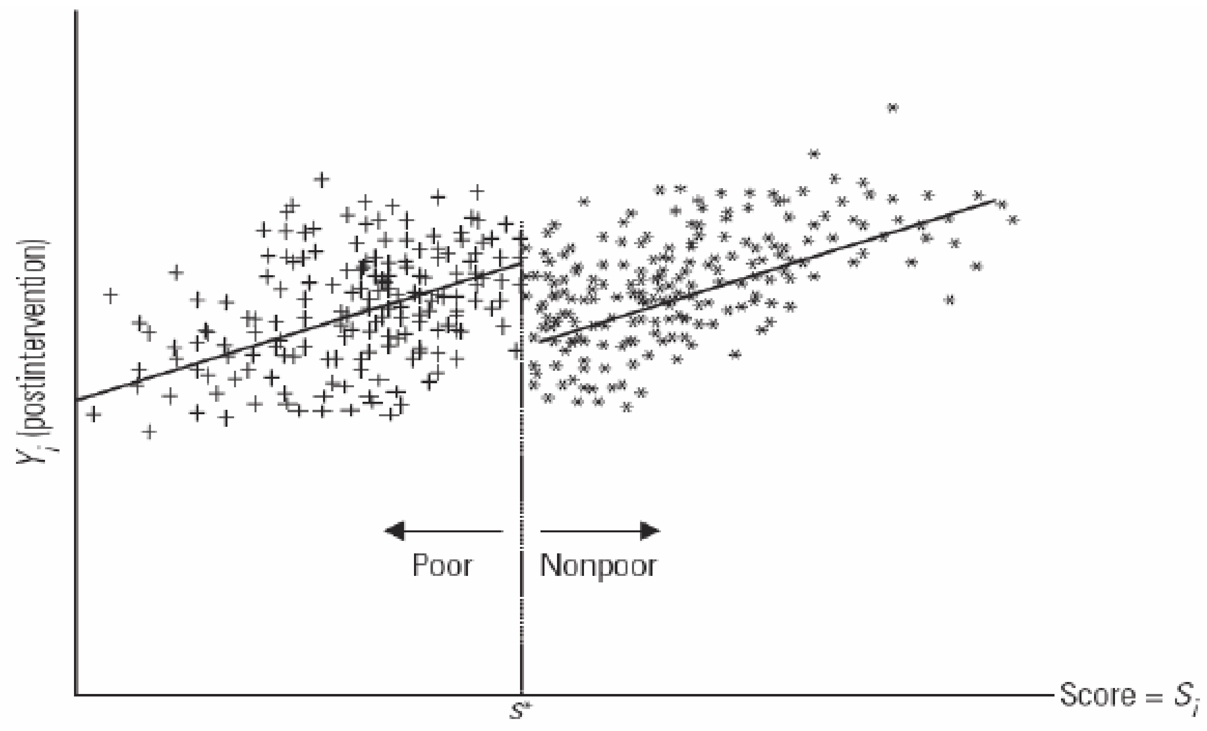
Regression Discontinuity: Intuition

- Outcomes before the program



Regression Discontinuity: Intuition

- Outcomes after the program

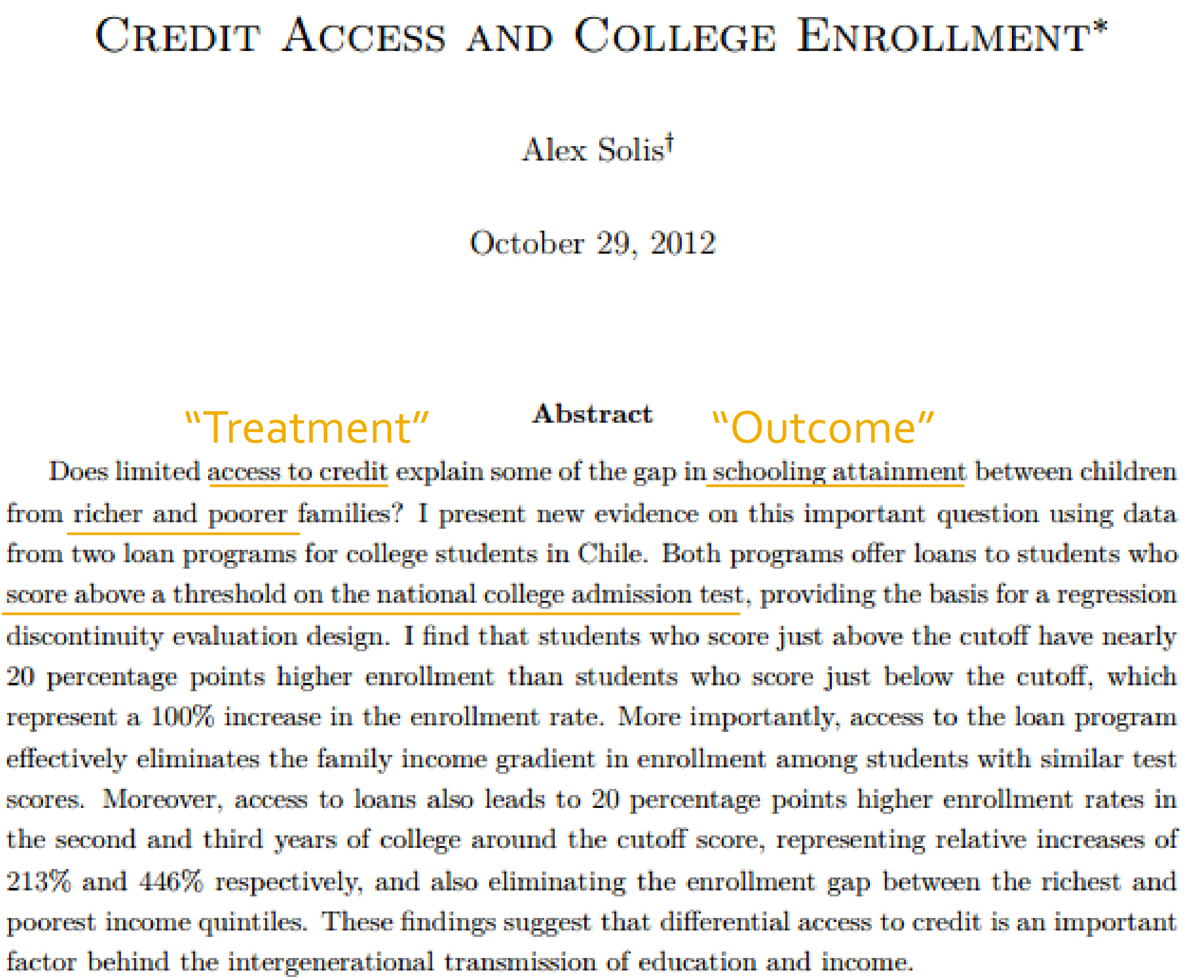


Regression Discontinuity: Intuition

- RDD uses individuals just below (or above) the threshold as a counterfactual for the treated individuals just above (or below) the threshold

- This requires that:
  - Other factors that determine the outcome are not discontinuous at the threshold point (smoothness assumption)
  - Running variable has not been "manipulated"

Regression Discontinuity: Example



CREDIT ACCESS AND COLLEGE ENROLLMENT*

Alex Solis†

October 29, 2012

Abstract

Does limited access to credit explain some of the gap in schooling attainment between children from richer and poorer families? I present new evidence on this important question using data from two loan programs for college students in Chile. Both programs offer loans to students who score above a threshold on the national college admission test, providing the basis for a regression discontinuity evaluation design. I find that students who score just above the cutoff have nearly 20 percentage points higher enrollment than students who score just below the cutoff, which represent a 100% increase in the enrollment rate. More importantly, access to the loan program effectively eliminates the family income gradient in enrollment among students with similar test scores. Moreover, access to loans also leads to 20 percentage points higher enrollment rates in the second and third years of college around the cutoff score, representing relative increases of 213% and 446% respectively, and also eliminating the enrollment gap between the richest and poorest income quintiles. These findings suggest that differential access to credit is an important factor behind the intergenerational transmission of education and income.

"Treatment"

"Outcome"

"covariates"

"Running variable"

Regression Discontinuity: Example

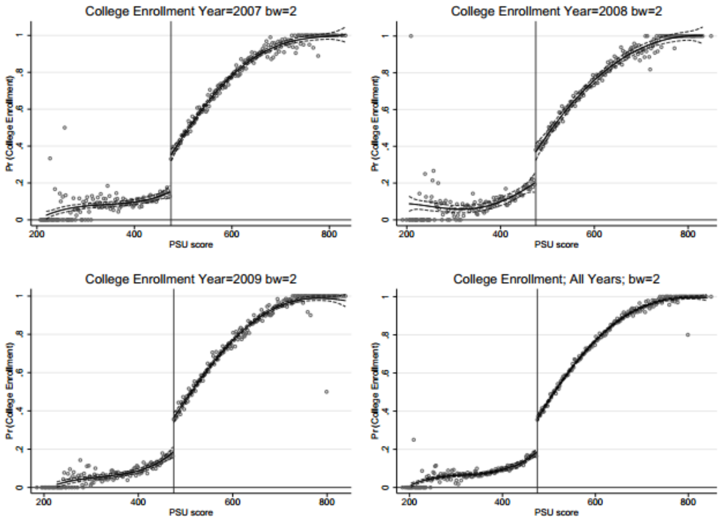
A key strength of RD is that it is often possible to test the presence (or absence) of effects with a few simple figures:

1. Outcomes (Y_i , T_i) vs. Running variable (Z_i)
2. Covariates (X_i) vs. Running variable
3. Density of Running variable

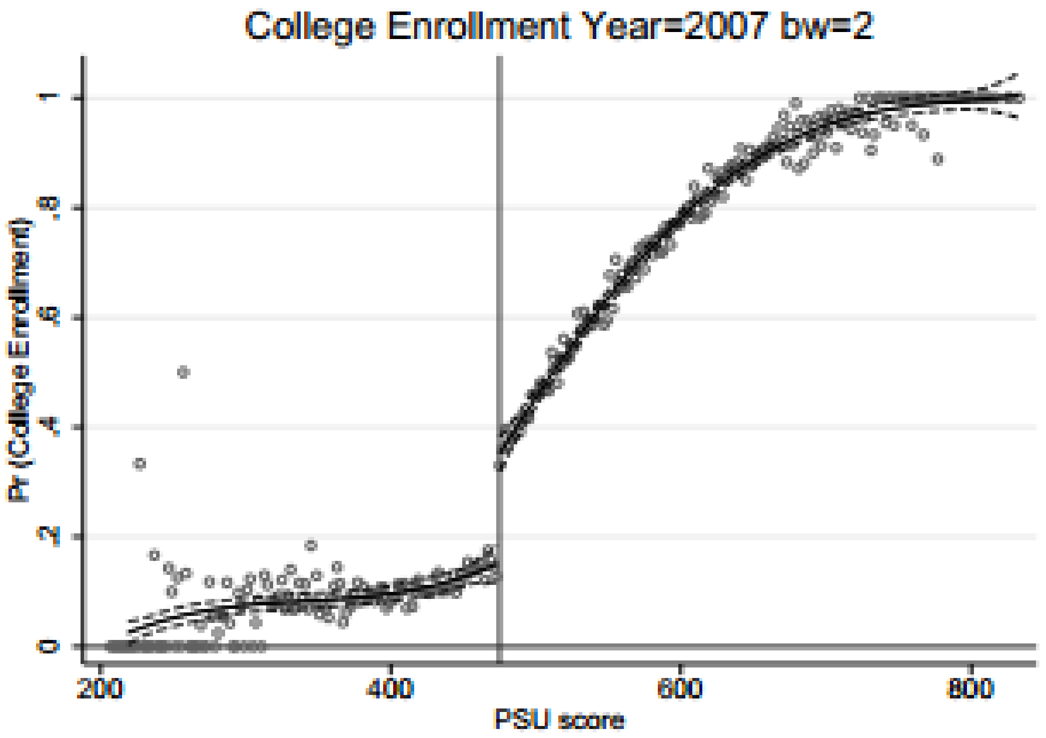
Regression Discontinuity: Example

1. **Outcomes (Y_i , T_i) vs. Running variable (Z_i)**
2. Covariates (X_i) vs. Running variable
3. Density of Running variable

Figure 6: RD for College enrollment. Full sample.



Note: Each dot represents average college enrollment in an interval of 2 PSU points.
The dashed lines represent fitted values from a 4th order spline and 95% confidence intervals for each side.
The vertical line indicates the cutoff (475).
These graphs show the full sample of students fulfilling all requirements to be eligible for college loans and taking the PSU immediately after graduating from high school.



Regression Discontinuity: Example

1. **Outcomes (Y_i , T_i) vs. Running variable (Z_i)**
2. Covariates (X_i) vs. Running variable
3. Density of Running variable

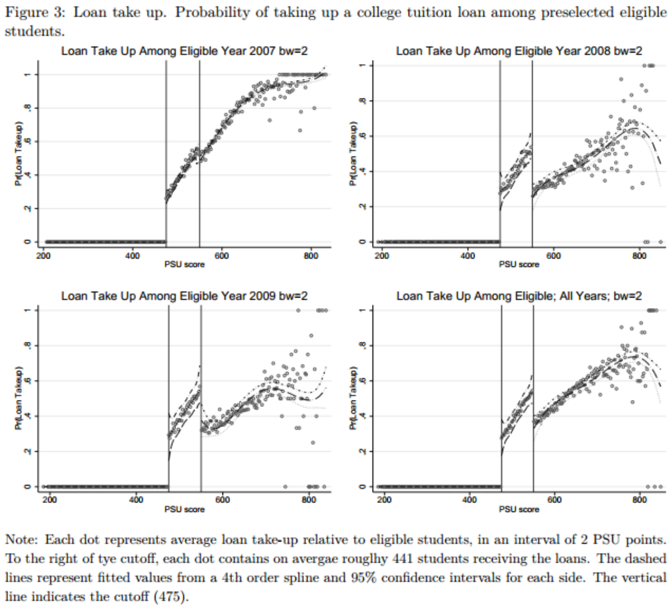
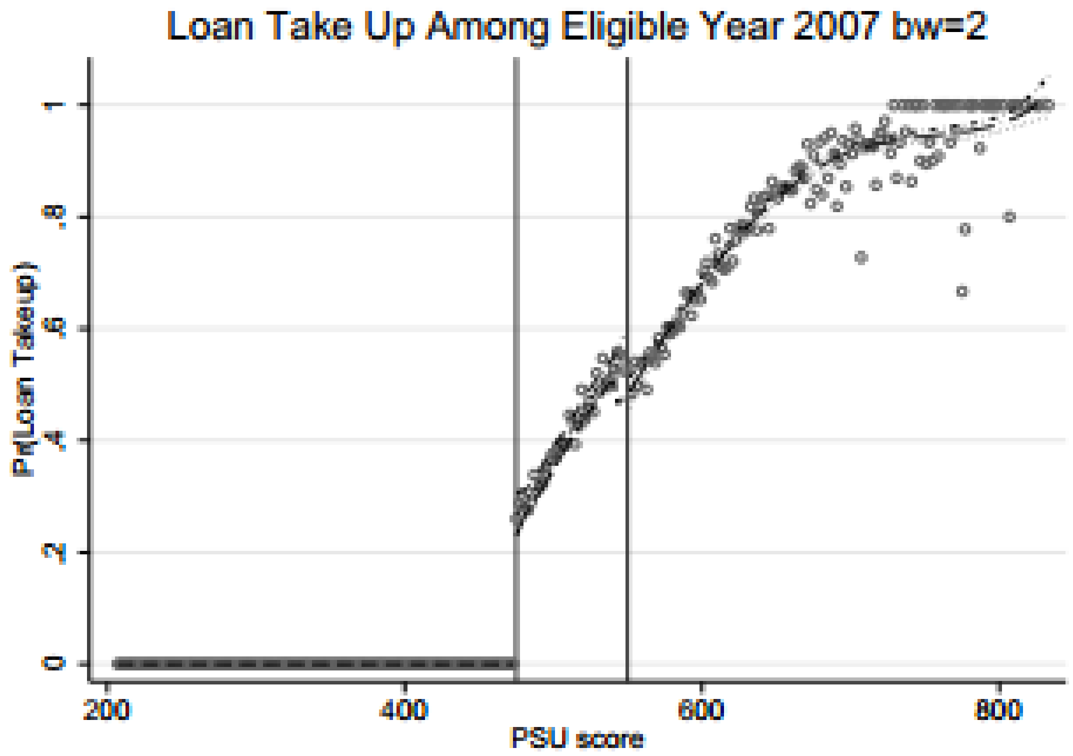


Figure 3: Loan take up. Probability of taking up a college tuition loan among preselected eligible students.

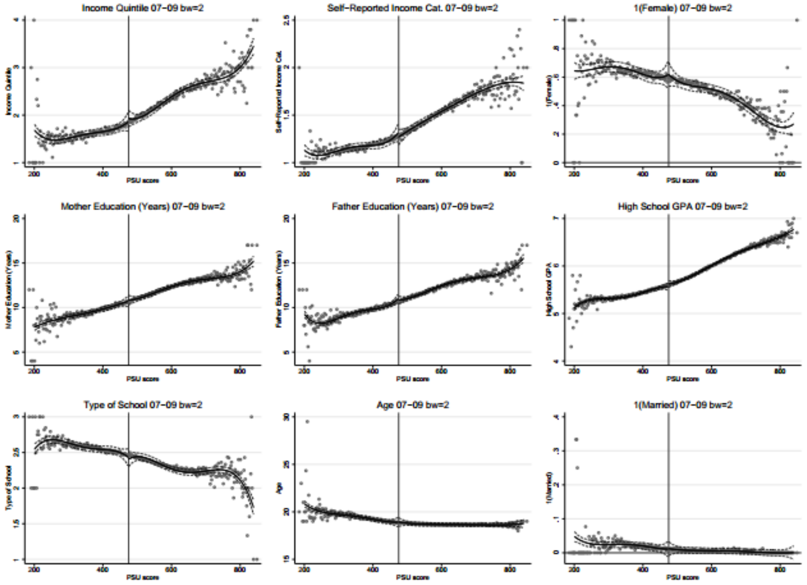
Note: Each dot represents average loan take-up relative to eligible students, in an interval of 2 PSU points. To the right of tye cutoff, each dot contains on avergae rouglhy 441 students receiving the loans. The dashed lines represent fitted values from a 4th order spline and 95% confidence intervals for each side. The vertical line indicates the cutoff (475).

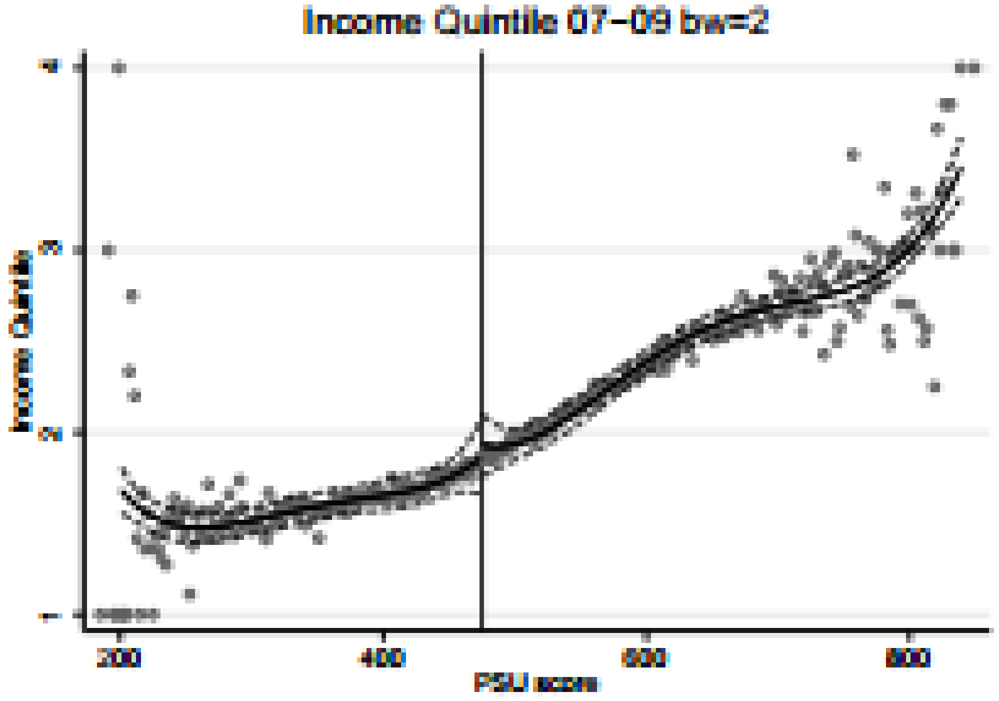


Regression Discontinuity: Example

1. Outcomes (Y_i , T_i) vs. Running variable (Z_i)
2. **Covariates (X_i) vs. Running variable**
3. Density of Running variable

Figure 5: RD for base line characteristics. Full sample.

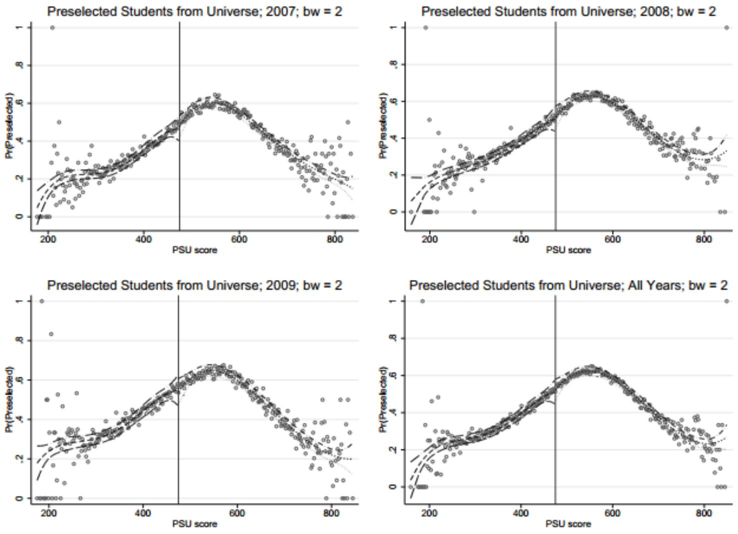




Regression Discontinuity: Example

1. Outcomes (Y_i , T_i) vs. Running variable (Z_i)
2. Covariates (X_i) vs. Running variable
3. **Density of Running variable**

Figure 2: Probability of being *Preselected* for loans.

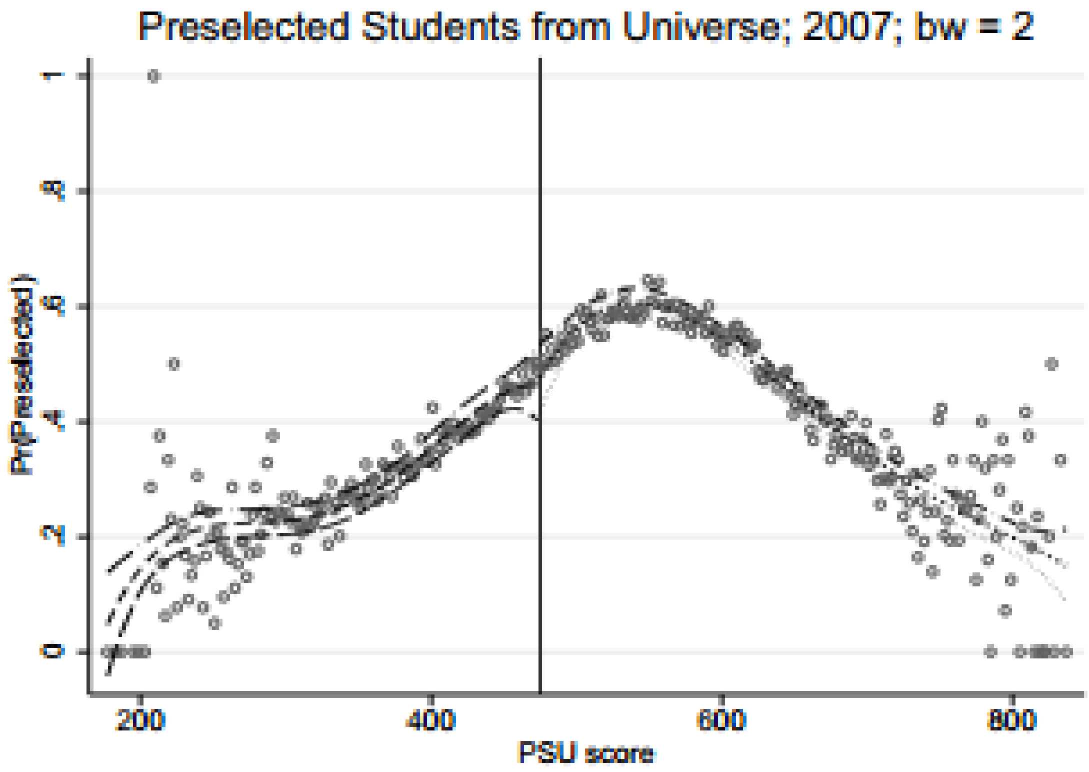

Preselected Students from Universe; 2007; bw = 2

Preselected Students from Universe; 2008; bw = 2

Preselected Students from Universe; 2009; bw = 2

Preselected Students from Universe; All Years; bw = 2

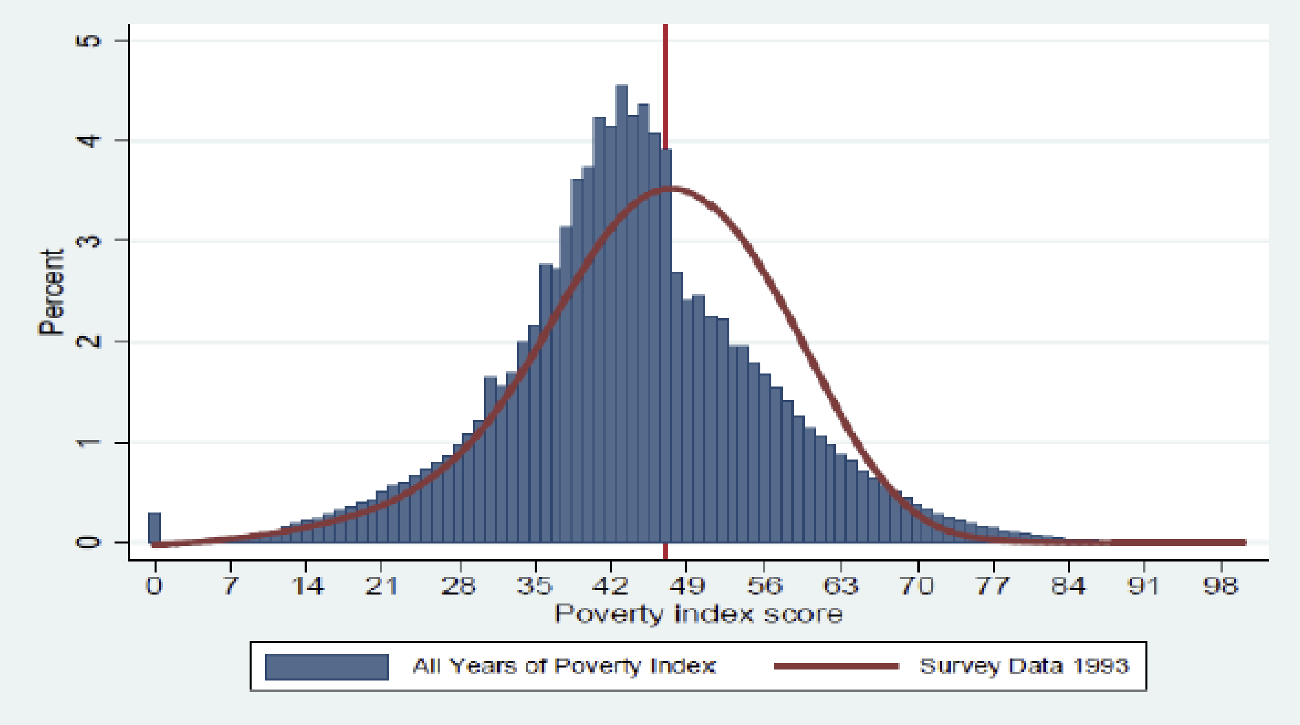
Note: Each dot indicates the preselection rate of students with scores in an interval of 2 PSU points (all students included). On average each dot contains 670 students. The dashed lines represent fitted values from a 4th order polynomial spline and 95% confidence intervals for each side. The vertical line indicates the cutoff (475).



Regression Discontinuity: Manipulation

- Density of Running variable
  - What if the running variable itself is discontinuous at the threshold?

Figure 3: Census of the Poor and 1993 Survey Data Score Distribution



Regression Discontinuity: Manipulation

The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations†

By THOMAS S. DEE, WILL DOBBIE, BRIAN A. JACOB, AND JONAH ROCKOFF*

- New York teachers inflate approximately 40 percent of test scores near the proficiency cutoffs
- Inflating a student's score to fall just above a cutoff increases his or her probability of graduating from high school by 27 percent
- Manipulation disappeared completely in 2012 when the Board ordered that exams be graded by teachers from other schools

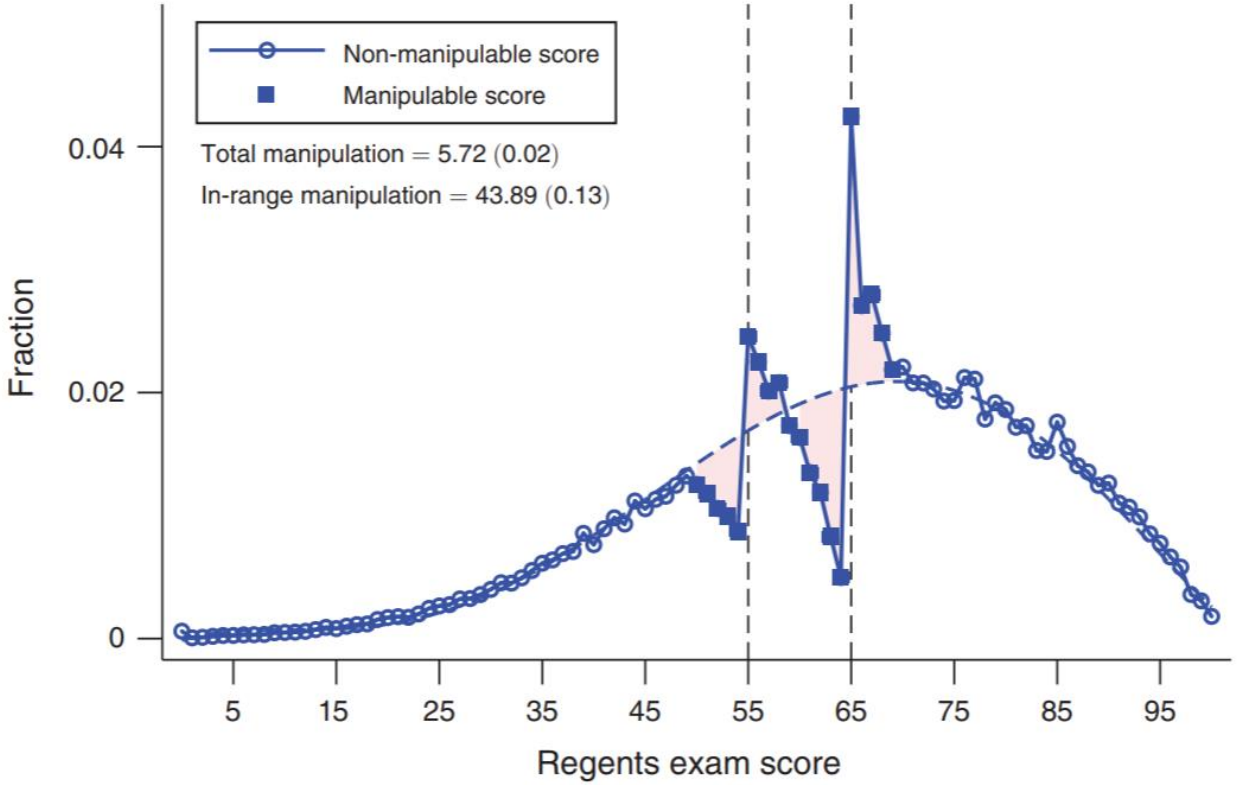


FIGURE 1. TEST SCORE DISTRIBUTIONS FOR CORE REGENTS EXAMS, 2004–2010

Regression Discontinuity: Example

- When the discontinuity precisely determines treatment, this is equivalent to quasi-random assignment *in a neighborhood*
- For instance:
  - Everyone older than 75 as of Jan 31 2021 is eligible for a Covid vaccine
    - (Let's assume that compliance is perfect)
  - We might compare rates of illness between people born in January 1946 and February 1946
    - Identifying assumption: Rates of illness in 2021 among people aged born in Jan and Feb 1946 *would have been the same* in the absence of the vaccine

Today's Outline

- Regression Discontinuity
  - Motivation and intuition
  - Regression discontinuity
  - Example: Graphical Analysis
  - Running variables
  - **Estimation**
  - Examples
- Econometrics Summary / Exercise

Regression Discontinuity: Estimation

- Quantifying the effect of the discontinuity
  - Instead of estimating: $GotSick_i = \alpha + \beta Vaccine_i + u_i$
  - We estimate: $GotFlu_i = \alpha + \beta(Over75_i) + \delta(AgeInDays_i) + u_i$
    - $Over75_i$ is a binary “treatment” variable
    - $AgeInDays_i$ is the individual’s age, in days
    - δ is a kernel (but just think of it as a constant, for now)
    - Estimated locally, for people with $\mathrm{s_{min}} < AgeInDays_i < s_{max}$
  - Note the similarity to Instrumental Variables!
    - $\beta(Over75_i)$ is an instrument for treatment status

Regression Discontinuity: Guidelines

- When to use this method?
  - The beneficiaries/non-beneficiaries can be ordered along a quantifiable dimension
  - This dimension can be used to compute a well-defined index or parameter
  - The index/parameter has a cut-off point for eligibility
  - The index value is what drives the assignment of a potential beneficiary to the treatment (or to non-treatment)

Regression Discontinuity: Placebo tests

- Show that the effects are zero when we use other arbitrary cutoff points
- Show that the effects are zero when using outcomes that the treatment should not influence

Regression Discontinuity: Examples

Figure I

Class Size in 1991 by Initial Enrollment Count, Actual Average Size and as Predicted by Maimonides' Rule

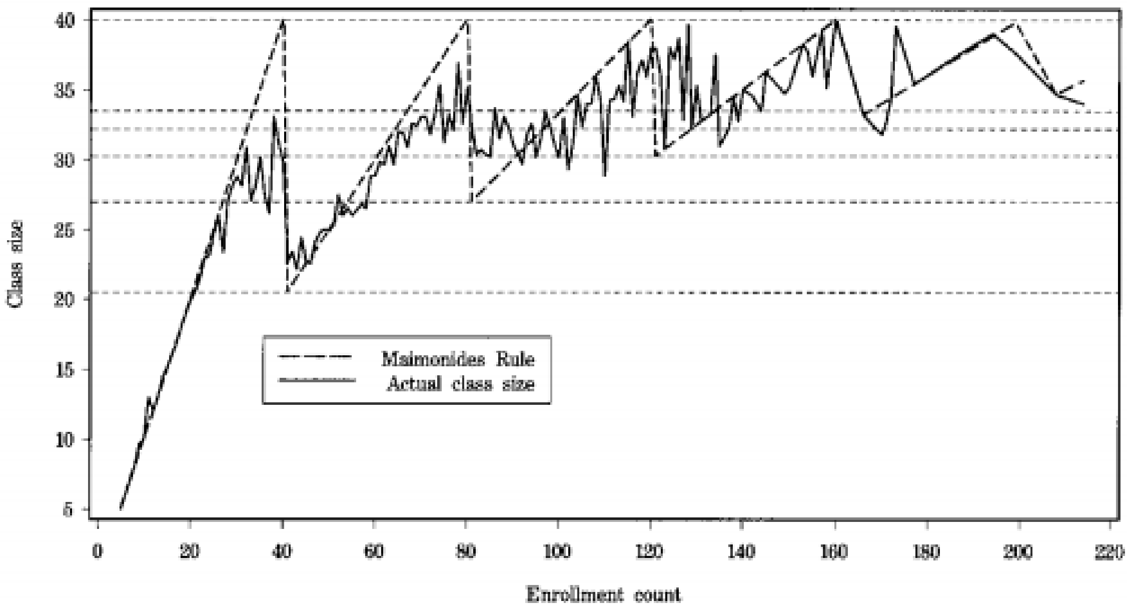
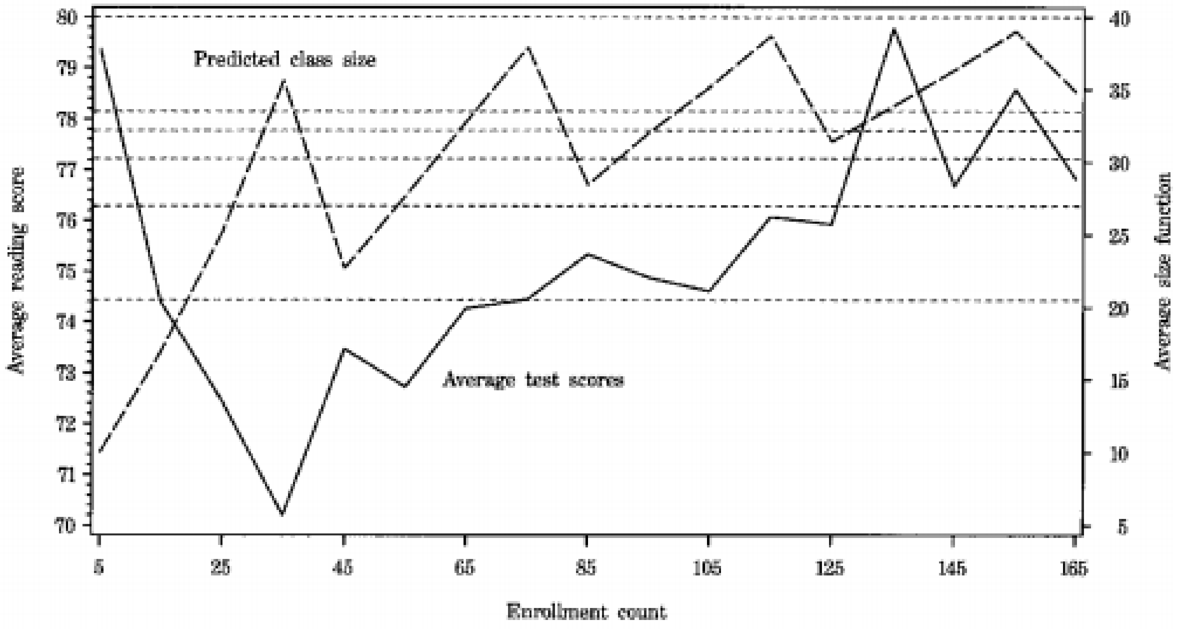


Figure II

Average Reading Scores by Enrollment Count, and the Corresponding Average Class Size Predicted by Maimonides' Rule



- "The twelfth century rabbinic scholar Maimonides proposed a maximum class size of 40. This same maximum induces a nonlinear and nonmonotonic relationship between grade enrollment and class size in Israeli public schools today. The estimates show that reducing class size induces a significant and substantial increase in test scores for fourth and fifth graders, although not for third graders."

Angrist and Lavy (1999 Quarterly Journal of Economics)

Regression Discontinuity: Examples

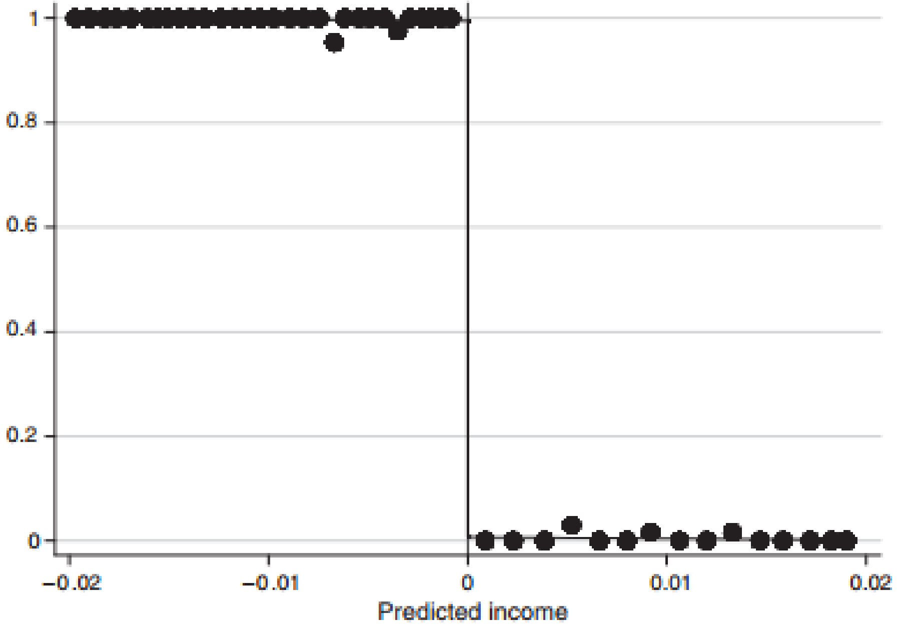


Figure 2. PANES Program Eligibility and Participation

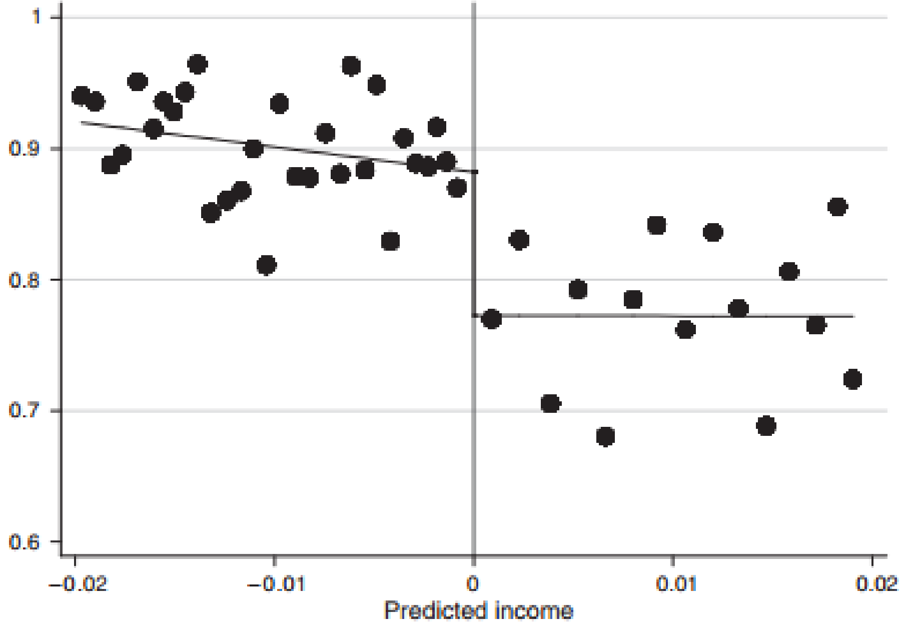


Figure 3. PANES Program Eligibility and Political Support for the Government,

- "This paper analyzes the effect of a large anti-poverty program, the Uruguayan Plan de Atención Nacional a la Emergencia Social (PANES), on expressed support for the government. We exploit the quasi-random assignment of applicants to the program based on a sharp discontinuity in a predicted income score in order to identify the effect of receiving transfers on support for the incumbent government and to ultimately advance understanding of voter decision making."

Manacorda et al (2011 AEJ: Applied Economics)

Regression Discontinuity: Examples

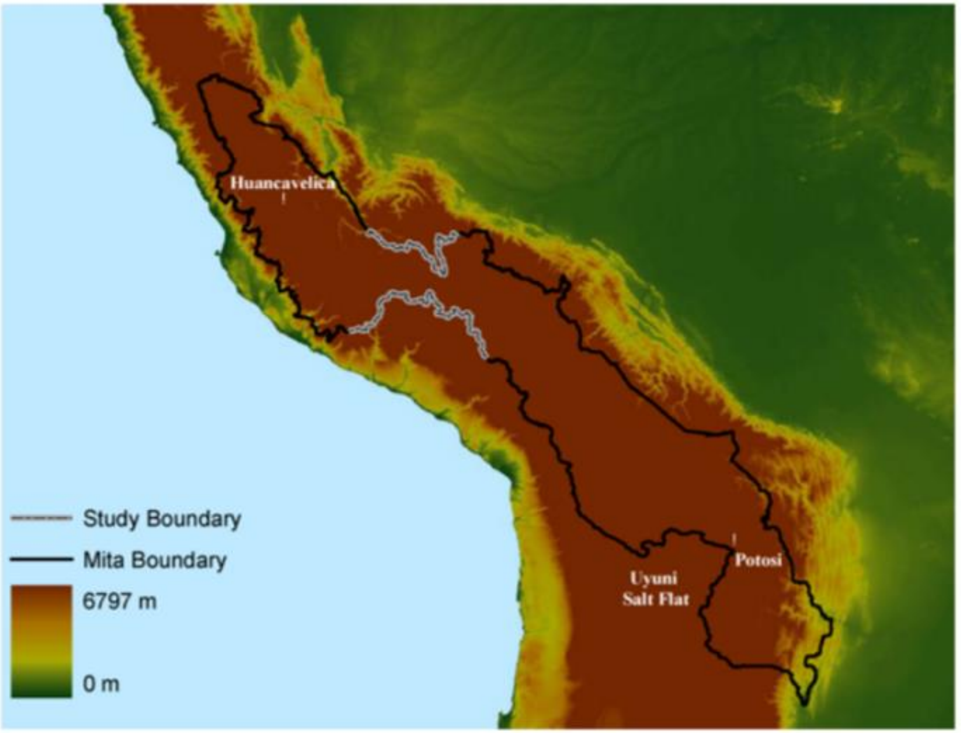
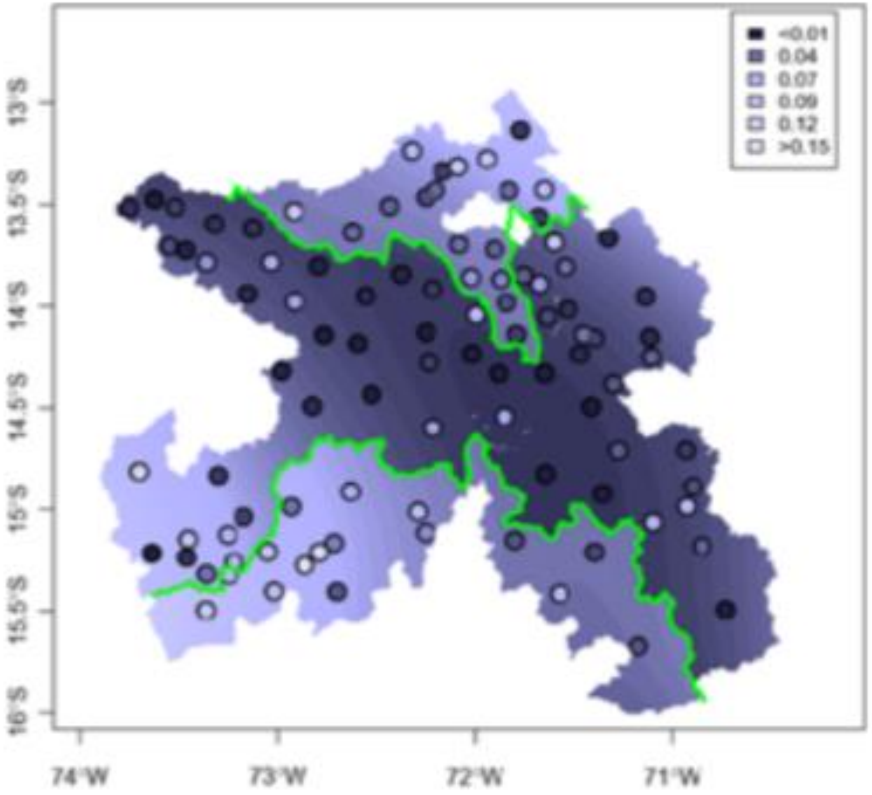


FIGURE 1.—The *mita* boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the *mita* boundary contributed to the *mita*. Elevation is shown in the background.



(f) Education (1876)

- "Results indicate that a *mita* effect lowers household consumption by around 25% and increases the prevalence of stunted growth in children by around 6 percentage points in subjected districts today"

Dell (2010 Econometrica)

Today's Outline and Key Concepts

- Regression Discontinuity
  - Motivation and intuition
  - Regression discontinuity
  - Example: Graphical Analysis
  - Running variables
  - Estimation
  - Examples
- **Econometrics Summary / Exercise**

Regression Discontinuity: Summary

- Advantages
  - Yields an unbiased estimate of treatment effect at the discontinuity
  - Takes advantage of a known rule for assigning the benefit that is common in the designs of social policy
  - A group of eligible households or individuals need not be excluded from treatment
  - Can be used in other settings
    - Spatial discontinuities
    - Temporal discontinuities (event studies)

Regression Discontinuity: Summary

- Disadvantages
  - Produces *local average treatment effects (LATE)* that are not always generalizable
  - Effect is estimated at the discontinuity, so generally, fewer observations exist than in a randomized experiment with the same sample size
  - Specification can be sensitive to functional form, including nonlinear relationships and interactions

Econometrics: Summary

- Wikipedia says:
  - **Econometrics** is the application of mathematics, statistical methods, and computer science, to economic data and is described as the branch of economics that aims to give empirical content to economic relations

- For the purposes of this class:
  - **Econometrics** is an enormously useful set of quantitative methods for understanding associations and causal relationships in data

Econometrics: What you've learned

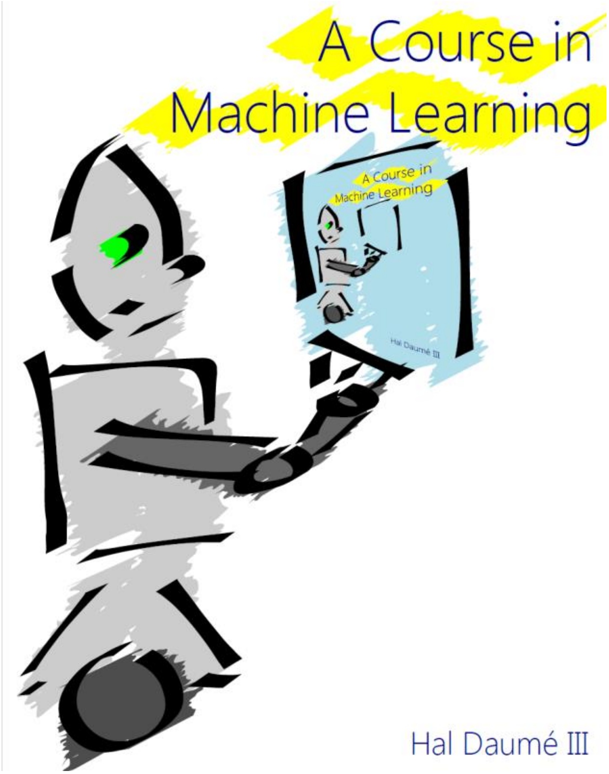
- Experimental methods
  - Design and randomization
  - Simple differences
  - Double differences
  - Regression
  - Fixed effects
- Non-experimental methods
  - All of the above and…
  - Instrumental variables
  - Regression discontinuity

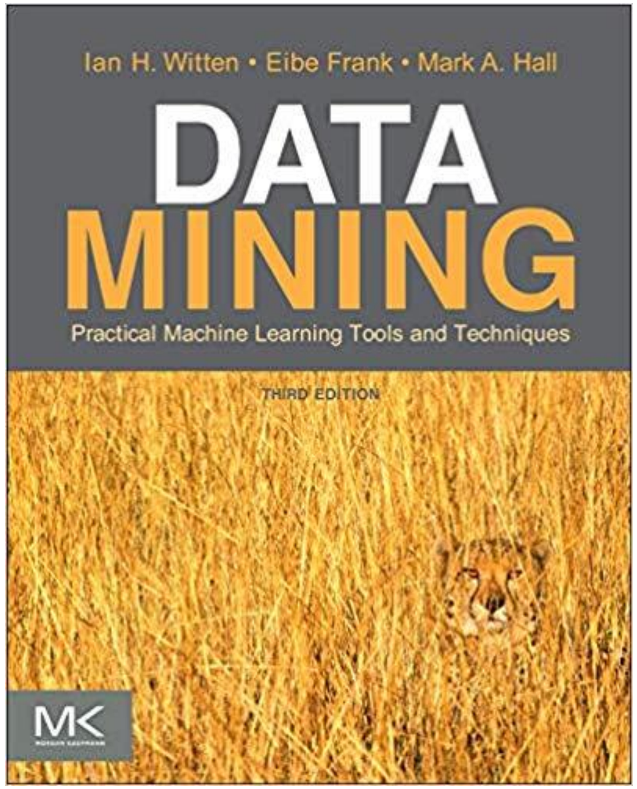
Econometrics: Key lesson

- No single method is “right” or “better”
- Each method requires a different identifying assumption, and implies a different counterfactual
- When deciding which method to use:
  - Determine which methods you *could potentially* use
  - For each candidate, articulate the identifying assumption
  - Brainstorm ways to possibly invalidate that assumption
  - Decide which assumption seems most reasonable, given your context, your data, and your situational knowledge

For Next Class: Intro to Machine Learning

- Read Daume (Chapters 1 & 2)
- Read Whitten et al (Chapter 5))




Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK