Segment 6: Quasi Experiments

Section 02: Regression Discontinuity

Quasi-Experimental Methods

- Most of the methods we've discussed for analyzing observational studies rely in large part on:
  - Knowing (and observing) the "right" X to believe the conditional ignorability assumption
  - Using those X to "restructure" the data to recreate the conditions of a randomized experiment

- A different class of methods aims to directly leverage circumstances of the study to *avoid* conditional ignorability:
  - If the circumstances are right, we may not need to "restructure" with X
  - We may be lucky that we can otherwise leverage the conditions of an "as if randomized" study design
  - Even without conditional ignorability
  - "Automatically" adjust for **unobserved** confounders

Regression Discontinuity Designs

Setting: Treatments are assigned according to some known, nonrandom rule, but that rule is dictated by some known variable falling above or below a certain threshold

The key idea of a *regression discontinuity* design is that units falling *just* above and *just below* that threshold are essentially identical on background characteristics $\rightarrow$ "quasi randomization"

Example: Hypothetical School Tutoring

- Wish to evaluate an after school tutoring program
- Due to program constraints, the program only admitted students who scored below 60 on a reading test
- Complete lack of overlap! Terrible balance!
  - $Pr(Z = 1|\text{pre-test} < 60) = 1$ and $Pr(Z = 1|\text{pre-test} \geq 60) = 0$
- Pre-test score would be called a *forcing* or *assignment variable*
  - The variable for which there is a threshold dictating treatment

- But maybe students who scored *close* to 60 on the test are really comparable
  - E.g., $57 <$pre-test< 63, sd(pre-test) = 15
- Could possible ground an assumption that program participation is "as though random" right around the pre-test threshhold

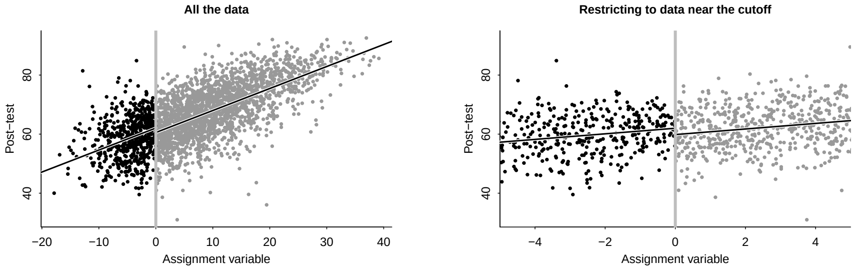
Example: Educational Program in Chile

from Gelman, Hill, Vehtari

- Chilean government introduced a program to improve struggling public schools
- Resources were provided to only to schools with certain test scores below a set cutoff (the "forcing variable")
- Some other complications, as described in text

Example: Educational Program in Chile

from Gelman, Hill, Vehtari



Example: Educational Program in Chile

from Gelman, Hill, Vehtari

$$y_i = \beta_0 + \tau Z_i + \beta_1 X_i + error_i$$

fit only to units within a certain range of the threshold

```
stan_glm(read92 ~ eligible + rule2, data=chile,
subset = abs(rule2)<5)
```

(can also include additional adjustment variables)

Example: Geographic Regression Discontinuity

Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

Yuyu Chen[a,1], Avraham Ebenstein[b,1], Michael Greenstone[c,d,1,2], and Hongbin Li[e,1]

[a]Applied Economics Department, Guanghua School of Management, Peking University, Beijing 100871, China; [b]Department of Economics, Hebrew University of Jerusalem, Mount Scopus 91905, Israel; [c]Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02142; [d]National Bureau of Economic Research, Cambridge, MA 02138; and [e]China Data Center and Department of Economics, School of Economics and Management, Tsinghua University, Beijing 100084, China

Edited by William C. Clark, Harvard University, Cambridge, MA, and approved May 28, 2013 (received for review January 2, 2013)

This paper's findings suggest that an arbitrary Chinese policy that greatly increases total suspended particulates (TSPs) air pollution is causing the 500 million residents of Northern China to lose more than 2.5 billion life years of life expectancy. The quasi-experimental empirical approach is based on China's Huai River policy, which provided free winter heating via the provision of coal for boilers in cities north of the Huai River but denied heat to the south. Using a regression discontinuity design based on distance from the Huai River, we find that ambient concentrations of TSPs are about 184 μg/m^3 [95% confidence interval (CI): 61, 307] or 55% higher in the north. Further, the results indicate that life expectancies are about 5.5 y (95% CI: 0.8, 10.2) lower in the north owing to an increased incidence of cardiorespiratory mortality. More generally, the analysis suggests that long-term exposure to an additional 100 μg/m^3 of TSPs is associated with a reduction in life expectancy at birth of about 3.0 y (95% CI: 0.4, 5.6).

causes). The estimates suggest that the 500 million residents of Northern China during the 1990s experienced a loss of more than 2.5 billion life years owing to the Huai River policy.

Furthermore, a research design based on this policy allows for a unique opportunity to estimate the effect of TSPs on human health, which can be applied to other countries, time periods, and settings. The resulting estimates suggest that long-term exposure to an additional 100 μg/m^3 of TSPs is associated with a reduction in life expectancy at birth of about 3.0 y (95% CI: 0.4, 5.6). This estimate is more than five times larger than the estimated impact of TSPs on life expectancy from fitting a conventional ordinary least-squares equation on the same data.

The study addresses several shortcomings in our understanding about the health effects of air pollution. First, the research design provides estimates of the impact of long-run exposure to TSPs on life expectancy. The policy caused long-run differences in TSP

Example: Geographic Regression Discontinuity

Chen et al. (2013)

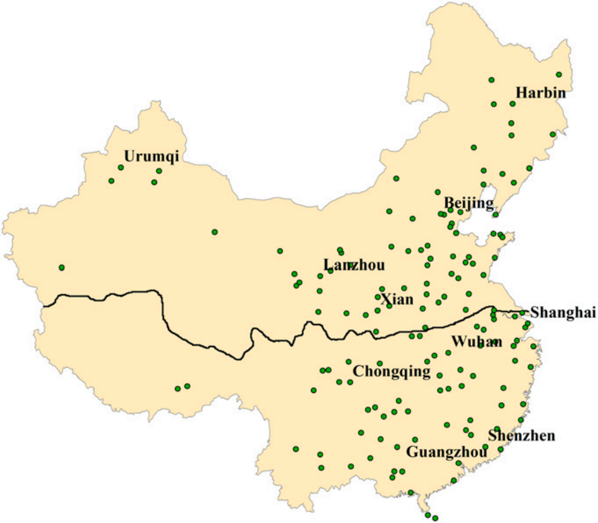


Fig. 1. The cities shown are the locations of the Disease Surveillance Points. Cities north of the solid line were covered by the home heating policy.

Example: Geographic Regression Discontinuity

Chen et al. (2013)

This paper's RD design exploits the discrete increase in the availability of free indoor heating as one crosses the Huai River line (with no availability to the south and, in principle, complete availability north of the line). Specifically, we separately test whether the Huai River policy caused a discontinuous change in TSPs at the river and a discontinuous change in life expectancy. The respective necessary assumptions are that any unobserved determinants of TSPs or mortality change smoothly as they cross the river. If the relevant assumption is valid, adjustment for a sufficiently flexible polynomial in distance from the river will remove all potential sources of bias and allow for causal inference.

Example: Geographic Regression Discontinuity

Chen et al. (2013)

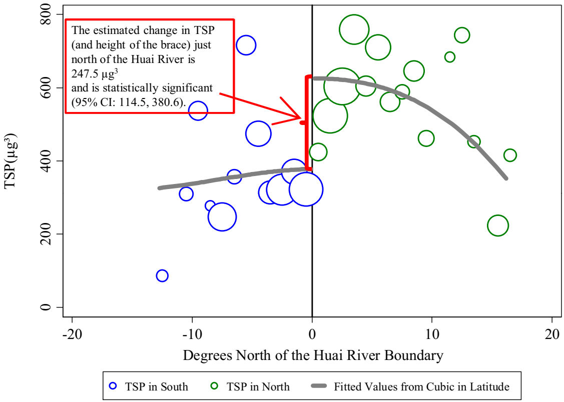


Fig. 2. Each observation (circle) is generated by averaging TSPs across the Disease Surveillance Point locations within a 1° latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1° latitude range. The plotted line reports the fitted values from a regression of TSPs on a cubic polynomial in latitude using the sample of DSP locations, weighted by the population at each location.

Example: Geographic Regression Discontinuity

Chen et al. (2013)

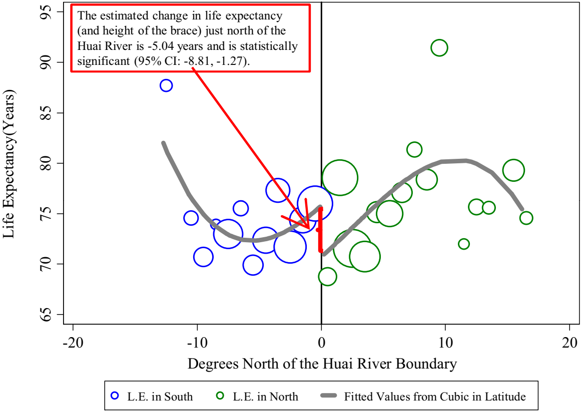


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Regression Discontinuity: Main Intuition

- Variation in "background" characteristics varies smoothly around the threshold of the forcing variable
  - $\Rightarrow$ Observed and unobserved factors are balanced between treatment groups
- Key question: are units below the threshold a good "control group" for those above the threshold?
  - $E[Y^0|Z=0] = E[Y^0|Z=1]$?
  - $E[Y^1|Z=0] = E[Y^1|Z=1]$?
- Often a very big tradeoff between *internal* and *external* validity
  - RD circumstances may produce high *internal* validity *for the population represented near the threshold*
  - Often comes at the cost of *external* validity if the population near the threshold is not representative of a broader population of interest