ECON 293/MGTECON 634: Machine Learning and Causal Inference

Susan Athey and Stefan Wager
Stanford University

Lecture 8: Regression Discontinuity Designs, and the Role of Optimization in Causal Inference

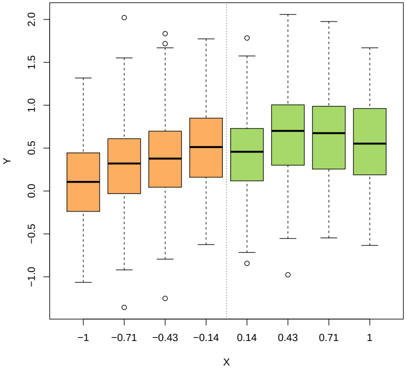
25 May 2018

Example of a regression discontinuity design:

- We want to understand the effect of supplementary feeding on future growth among under-nourished children.
- The current protocol treats children whose weight-for-age Z-score (WAZ) falls below $c = -2.5$.

Identification strategy: We can measure causal effects by comparing trajectories of children whose WAZ score falls just above/below the cutoff c.

Identification in regression discontinuity designs

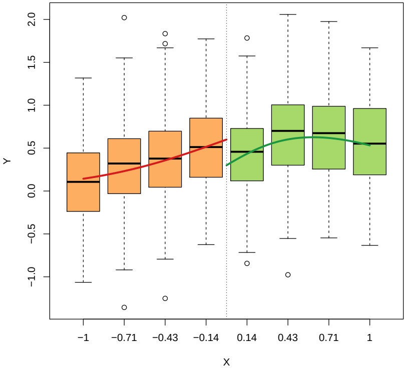


Identifying causal effects via **regression discontinuities** is increasingly popular (Hahn, Todd, and van der Klaauw, 2001):

$$\tau = \lim_{h \downarrow 0} \left(\mathbb{E}\left[Y \,\middle|\, X = h \right] - \mathbb{E}\left[Y \,\middle|\, X = -h \right] \right).$$

NB: In many applications, we only observe X over a **discrete grid**, and so we have a **partial identification problem**.

Identification in regression discontinuity designs



Identifying causal effects via **regression discontinuities** is increasingly popular (Hahn, Todd, and van der Klaauw, 2001):

$$\tau = \lim_{h \downarrow 0} \left(\mathbb{E}\left[Y \,\middle|\, X = h \right] - \mathbb{E}\left[Y \,\middle|\, X = -h \right] \right).$$

NB: In many applications, we only observe X over a **discrete grid**, and so we have a **partial identification problem**.

Estimation in regression discontinuity designs

We use the Neyman-Rubin **potential outcomes model**, with data

$$\{X_i,\, Y_i,\, W_i\}_{i=1}^n\,, \;\; Y_i = Y_i(W_i),\;\; \tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right].$$

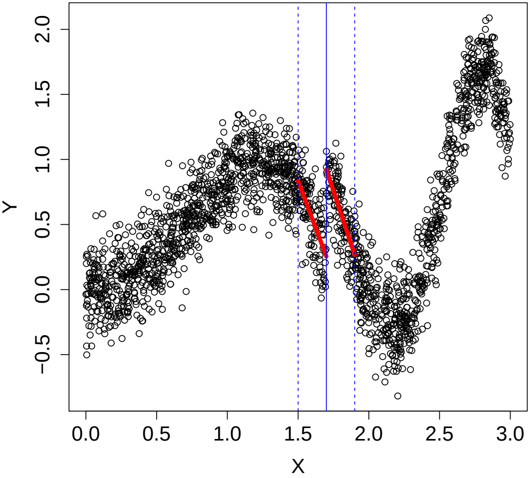
In the simplest case, $X_i \in \mathbb{R}$, and $W_i = 1\left(\{X_i \geq 0\}\right)$ is determined by a **single cutoff**.

There are several approaches for estimating $\tau(0)$, often framed in terms of **models estimated on both sides of the boundary**; see Imbens and Lemieux (2008) for a review:

- **Local linear/polynomial regression**.
- **Weighted local linear/polynomial regression**.

Consistency is verified via local estimation theory.

RDDs via local linear regression



How not to use machine learning in RDDs

We use the Neyman-Rubin **potential outcomes model**, with data

$$\{X_i,\ Y_i,\ W_i\}_{i=1}^n,\ \ Y_i = Y_i(W_i),\ \ \tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = x\right].$$

In the simplest case, $X_i \in \mathbb{R}$, and $W_i = 1\left(\{X_i \geq 0\}\right)$ is determined by a **single cutoff**.

A simple idea (but **don't do this!**) is to estimate

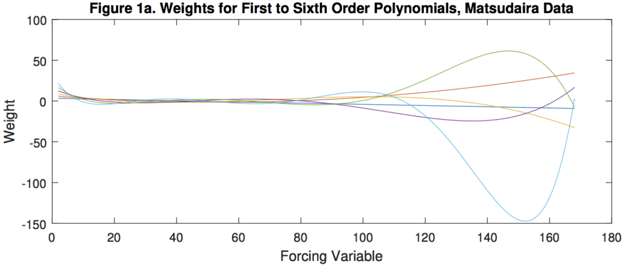
$$\hat{\mu}_w(x) = \widehat{\mathbb{E}}\left[Y \,\middle|\, X = x,\ W = w\right]$$

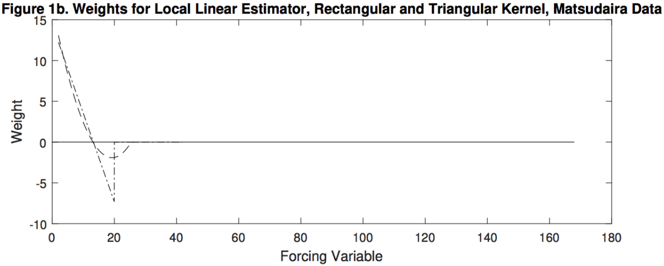
globally via a **non-parametric method**, and then set

$$\hat{\tau}(c) = \hat{\mu}_1(c) - \hat{\mu}_0(c).$$

Variants of this idea, especially using **higher-order polynomial** regression, are unfortunately quite common; see discussion in Imbens and Gelman (2017).

How not to use machine learning in RDDs





[Image source: Imbens and Gelman, 2017]

Estimation in regression discontinuity designs

The **conceptual justification** for **local linear regression** typically relies on smoothness assumptions of the form:

$$\left| \frac{d^2}{dx^2} \mathbb{E}\left[Y(w) \,\middle|\, X = x \right] \right| \leq B. \tag{1}$$

If X is continuous and univariate with a single threshold, and we use weighted linear regression, then a **triangular kernel** is optimal (Cheng, Fan, and Marron, 1997).

This type of assumption is often used for **bandwidth selection**; see, e.g., Imbens and Kalyanaraman (2012).

But if we are willing to assume (1), is local linear regression **the best we can do**? Also, how do we generalize to more complex problems such as **geographic discontinuities**?

Estimation in regression discontinuity designs

All (potentially weighted) local linear regression estimators can be written as **linear estimators**,

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i.$$

For example, standard OLS calculations imply that unweighted local linear regression with bandwidth h uses the following weights, where $\mathcal{S}_h^+ = \{i : c < X_i < c + h\}$, etc.

$$\hat{\gamma}_i = \begin{cases} \frac{\text{avg}_{\mathcal{S}_h^+}\left\{(X_i - c)^2\right\} - \text{avg}_{\mathcal{S}_h^+}\{X_i - c\}(X_i - c)}{\text{avg}_{\mathcal{S}_h^+}\{(X_i - c)^2\} - \text{avg}_{\mathcal{S}_h^+}\{X_i - c\}^2} & \text{if } i \in \mathcal{S}_h^+ \\ -\frac{\text{avg}_{\mathcal{S}_h^-}\left\{(X_i - c)^2\right\} - \text{avg}_{\mathcal{S}_h^-}\{X_i - c\}(X_i - c)}{\text{avg}_{\mathcal{S}_h^-}\{(X_i - c)^2\} - \text{avg}_{\mathcal{S}_h^-}\{X_i - c\}^2} & \text{if } i \in \mathcal{S}_h^- \\ 0 & \textit{else}. \end{cases}$$

These weights only depend on the X_i.

Estimation in regression discontinuity designs

All (potentially weighted) local linear regression estimators can be written as **linear estimators**,

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i.$$

The weights underlying local linear regression can also be expressed more abstractly as

$$\hat{\gamma}_i = \operatorname{argmin}_\gamma \left\{ \|\gamma\|_2^2 : \sum_{X_i < 0} \gamma_i = -1, \sum_{X_i > 0} \gamma_i = 1, \right.$$
$$\left. \sum_{X_i < 0} X_i \gamma_i = 0, \sum_{X_i > 0} X_i \gamma_i = 0, \gamma_i 1\left\{|X_i > h|\right\} = 0 \right\}.$$

"Idea:" Try to **optimize error bounds** among linear estimators.

Optimizing regression discontinuity designs

Suppose we use an estimator of the form $\hat{\tau} = \sum_{i=1}^n \hat{\gamma}_i Y_i$, where the weights $\hat{\gamma}_i$ depend only on the X_i. Then, the conditionall **variance** of this estimator is

$$s^2 = \text{Var}\left[\hat{\tau} \,\middle|\, X_1, \ldots, X_n\right] = \sum_{i=1}^n \hat{\gamma}_i^2 \sigma_i^2, \quad \sigma_i^2 = \text{Var}\left[Y_i \,\middle|\, X_i, W_i\right].$$

Moreover, if $|\mu''_{(w)}(x)| \leq B$, we can bound the worst-case conditional **bias** as

$$\left|\mathbb{E}\left[\hat{\tau} \,\middle|\, X_1, \ldots, X_n\right] - \tau(c)\right| \leq \hat{t}$$

$$\hat{t} = \sup\left\{\left(\sum_{i=1}^n \hat{\gamma}_i \mu_{(W_i)}(X_i)\right) - \left(\mu_{(1)}(c) - \mu_{(0)}(c)\right) : \left|\mu''_{(w)}(x)\right| \leq B\right\}.$$

The worst-case **mean-squared error** is $s^2 + \hat{t}^2$.

Optimizing regression discontinuity designs

We can numerically derive the **minimax linear** estimator of the form $\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i$, by optimizing $s^2 + \hat{t}^2$. Below, c is the discontinuity point and $\sigma_i^2 = \text{Var}\left[Y_i(W_i) \,\middle|\, X_i\right]$:

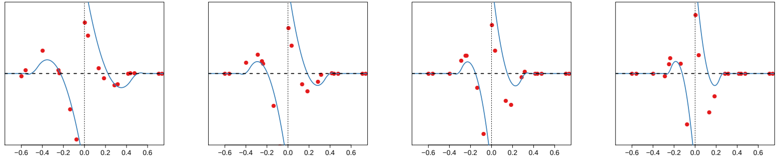
$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i, \quad \hat{\gamma} = \text{argmin}_{\gamma} \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + I_B^2(\gamma) \right\},$$

$$I_B(\gamma) := \sup_{\mu_0(\cdot), \mu_1(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_{W_i}(X_i) - (\mu_1(c) - \mu_0(c)) : \right.$$

$$\left. \left|\mu_w''(x)\right| \leq B \text{ for all } w, x \right\}.$$

This is **fully automatic** given a bound B on the second derivative, and does not require a choice of **bandwidth** or **weighting kernel**.

Optimized weighting functions



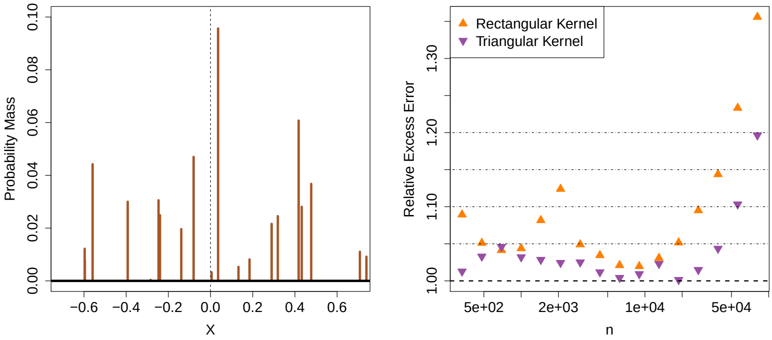
$n = 1,000$ $n = 3,000$ $n = 9,000$ $n = 27,000$

Comparison of **optimized** weighting functions $\hat{\gamma}_i$ for a **discrete design** and a **continuous design** with comparable amounts of data near the boundary.

- The shape of the optimal discrete weighting function **changes with sample size**.

Software implementation is available on CRAN: `optrdd` for `R`.

Optimized weighting functions



In this example, the **optimized design** improves over **local linear regression**.

What about confidence intervals?

We **estimate** the regression discontinuity parameter as

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i, \quad \left\{\hat{\gamma},\, \hat{t}\right\} = \text{argmin}_{\gamma,\, t} \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + t^2 : I_B(\gamma) \leq t \right\},$$

$$I_B\left(\gamma\right) := \sup_{\mu_0(\cdot), \mu_1(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_{W_i}(X_i) - \left(\mu_1(c) - \mu_0(c)\right) : \right.$$

$$\left. \left|\mu_w''(x)\right| \leq B \text{ for all } w,\, x \right\}.$$

The optimization problem thus provides us with an explicit bound for the **worst-case bias** as $\hat{t}$. Can we use it for confidence intervals?

What about confidence intervals?

If $Y_i \,\big|\, X_i,\, W_i$ is Gaussian, then (and in large samples, this is approximately true thanks to the central limit theorem)

$$\hat{\tau} \,\big|\, X \sim \mathcal{N}\left(\tau + b,\, s^2\right), \quad \text{for some} \quad |b| \leq \hat{t}.$$

In this setup, we can build **bias-aware confidence intervals** via the construction of Imbens and Manski (2004),

$$\tau \in \hat{\tau} \pm \ell_\alpha, \quad \ell_\alpha = \min\left\{\ell : \mathbb{P}\left[|b + sZ| \geq \ell\right] \leq \alpha \text{ for all } b \leq \hat{t}\right\},$$

where $s^2 = \sum_{i=1}^n \sigma_i^2 \gamma_i^2$ and $Z \sim \mathcal{N}(0,\, 1)$. Note that, in practice, we can also estimate the noise as

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \widehat{\mathbb{E}}\left[Y_i \,\big|\, X_i,\, W_i\right]\right)^2.$$

What about partial identification?

With a **discrete running variable**, treatment effects are only partially identified.

Because we **account for bias** in finite sample, the optimized method automatically gives valid confidence intervals for **partially identified** treatment parameters in the sense of Imbens and Manski (2004). We cover any point in the identification interval with probability at least $1 - \alpha$.

Point identification is just an **asymptotic statement** about whether the length of our confidence intervals goes to zero in large samples.

The effect of compulsory schooling

We consider a dataset from Oreopoulos (2006), who studied the effect of raising the **minimum school-leaving** age on earnings as an adult.

- The effect is identified by the UK changing its minimum school-leaving age from 14 to 15 in 1947.
- The response is log-earnings among those with non-zero earnings (in 1998 pounds).

This dataset exhibits notable **discreteness in its running variable**, i.e., the year in which a person turned 14.

The effect of compulsory schooling

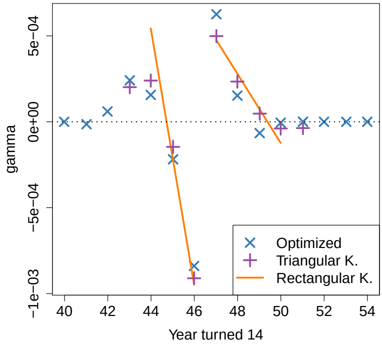
B	rect. kernel	tri. kernel	optimized
0.003	0.0213 ± 0.0761	0.0321 ± 0.0737	0.0302 ± 0.0716
0.006	0.0578 ± 0.0894	0.0497 ± 0.0867	0.0421 ± 0.0841
0.012	0.0645 ± 0.1085	0.0633 ± 0.1037	0.0557 ± 0.1003
0.03	0.0645 ± 0.1460	0.0710 ± 0.1367	0.0710 ± 0.1329

95% **confidence intervals** for $\tau(c)$ given different choices of B.

A global quadratic fit for the treated/controls separately suggests a **curvature** around $B = 0.006$ away from the cutoff.

All confidence intervals are **bias-aware** (even for local linear regression, one can use numerical optimization to derive the worst-case bias).

The effect of compulsory schooling



The plot above shows weights from local linear regression with a **rectangular** and **triangular** kernel, as well as **optimized** weights. In all cases, we use the weights to **estimate** $\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i$.

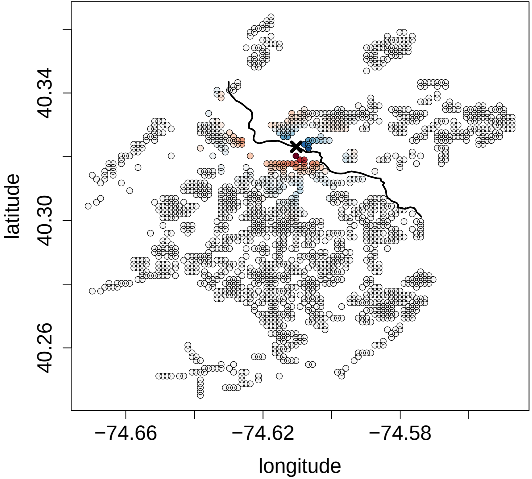
Multivariate regression discontinuity designs

There are many problems where the **treatment/control boundary** is more complicated than a single cutoff.

Example. Keele and Titiunik (2014) study the effect of **television advertising** on **voter turnout** in presidential elections

- For **identification**, they examine a school district in New Jersey, half of which belongs to the Philadelphia media market (= many ads) and the other half to the New York media market (= no ads).
- This is a **geographic RDD**, where the "cutoff" corresponds to the media-market boundary.

Application: Effect of political advertising



Data set of Keele and Titiunik (2014), with $n = 24,460$ samples over a school district. These weights estimate $\tau(c)$ at the point marked with $\times$.

Optimizing multivariate regression discontinuity designs

We now have a multivariate running variable $X \in \mathbb{R}^k$, and treatment is assigned as $W_i = 1\left(\{X_i \in \mathcal{A}\}\right)$ for some set $\mathcal{A}$. Generalizing our previous approach, we bound **curvature** via

$$\left\|\nabla^2 \mu_w(x)\right\| \leq B \text{ for all } w,\, x.$$

Then, for any **focal point** c along the boundary, we can estimate $\tau(c) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, X_i = c\right]$ as

$$\hat{\tau} = \sum_{i=1}^n \hat{\gamma}_i Y_i, \quad \hat{\gamma} = \text{argmin}_\gamma \left\{\sum_{i=1}^n \gamma_i^2 \sigma_i^2 + I_B^2(\gamma)\right\},$$

$$I_B\left(\gamma\right) := \sup_{\mu_0(\cdot), \mu_1(\cdot)} \left\{ \sum_{i=1}^n \gamma_i \mu_{W_i}(X_i) - \left(\mu_1(c) - \mu_0(c)\right) : \left\|\nabla^2 \mu_w(x)\right\| \leq B \text{ for all } w,\, x \right\}.$$

This provides an estimator for the **conditional average treatment effect** $\tau(c)$.

Optimizing multivariate regression discontinuity designs

Restricting our analysis to the neighborhood of a single focal point c may cost us **power**.

If we are willing to assume a **constant treatment effect**, then we can seamlessly use data anywhere along the boundary.

In the constant effect model, we have $\mu_{(1)}(x) = \mu_{(0)}(x) + \tau$ with

$$\left\|\nabla^2 \mu_0(x)\right\| \leq B \text{ for all } x,$$

and the optimization problem **simplifies** to

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i, \quad \hat{\gamma} = \operatorname{argmin}_{\gamma} \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + I_B^2(\gamma) : \sum_{i=1}^{n} \gamma_i W_i = 1 \right\},$$

$$I_B\left(\gamma\right) := \sup_{\mu_0(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_0(X_i) : \left\|\nabla^2 \mu_0(x)\right\| \leq B \text{ for all } x \right\}.$$

This provides an estimator for the **constant treatment effect** τ.

Optimizing multivariate regression discontinuity designs

We can also interpret the output of the constant treatment effect estimator under **treatment heterogeneity**.

If we run the following estimator,

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i, \quad \hat{\gamma} = \operatorname{argmin}_{\gamma} \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + I_B^2(\gamma) : \sum_{i=1}^{n} \gamma_i W_i = 1 \right\},$$

$$I_B\left(\gamma\right) := \sup_{\mu_0(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_0(X_i) : \left\| \nabla^2 \mu_0(x) \right\| \leq B \text{ for all } w,\, x \right\},$$

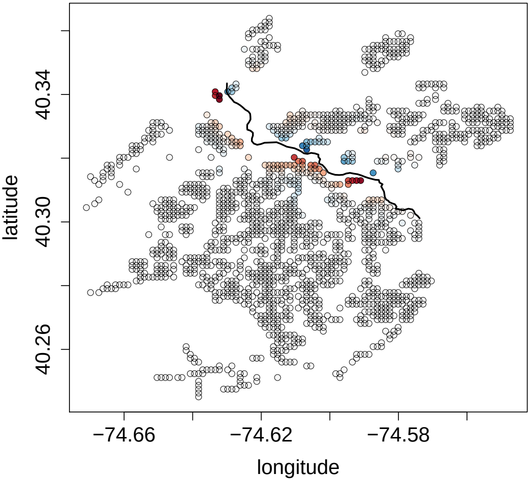
we are estimating the **weighted average treatment effect** $\bar{\tau}_\gamma$,

$$\bar{\tau}_\gamma = \sum_{i=1}^{n} \gamma_i W_i \tau(X_i),$$

where the weights have been chosen to maximize **precision**.

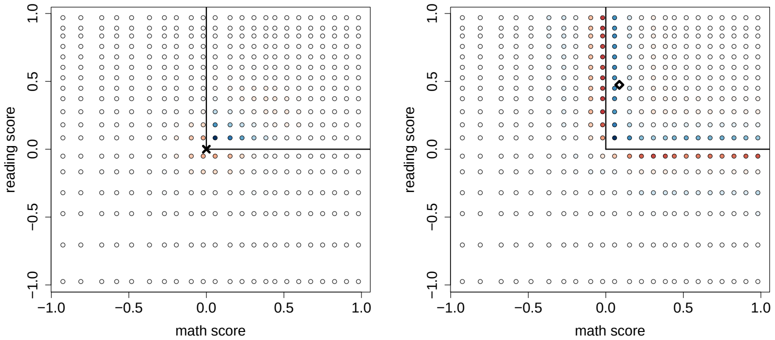
At a high level, these weights are connected to "overlap weights" discussed 2 weeks ago.

Application: Effect of political advertising



Data set of Keele and Titiunik (2014), with $n = 24,460$ samples over a school district. Weights allow for CATE averaging. We replicate null finding while directly controlling for spatial curvature.

Application: Effect of summer school



We want to measure the effect of **mandatory summer school** on **next year's grades**. Identification strategy: Students who fail either a year end math test or reading test need to go to summer school (Jacob and Lefgren, 2004; Matsudaira, 2008).

Existing analyses typically filter students who pass reading and then use math score as a **univariate discontinuity**.

Application: Effect of summer school

estimator:		unweighted CATE			weighted CATE		
subject	B	conf. int.	bias	s.e.	conf. int.	bias	s.e.
math	0.5	0.04 ± 0.093	0.03	0.04	0.08 ± 0.037	0.01	0.02
math	1.0	0.01 ± 0.126	0.04	0.05	0.07 ± 0.043	0.01	0.02
read	0.5	0.01 ± 0.098	0.03	0.04	0.04 ± 0.037	0.01	0.02
read	1.0	-0.01 ± 0.130	0.04	0.05	0.05 ± 0.043	0.01	0.02

Estimates for the effect of summer school on math and reading scores on the following year's test, using different estimators and choices of B. Reported are bias-adjusted 95% confidence intervals, a bound on the maximum bias given our choice of B, and an estimate of the sampling error conditional on $\{X_i\}$. Values of B are multiplied by 40^2.

Application: Effect of summer school

Code example with the `R` package `optrdd` (on CRAN). X denotes the (bivariate) running variable, and Y denotes the outcome. B is a bound on the curvature.

```
W = as.numeric((X[,1] < 0) | (X[,2] < 0))
optrdd.fit = optrdd(X = X, Y = Y, W = W,
                    max.second.derivative = B)
print(optrdd.fit)
[1] "95% CI for tau: 0.07 +/- 0.043"
```

See `www.github.com/swager/optrdd` for more examples.

Solution via convex duality

We can solve the underlying problem via **convex optimization**. To do so, consider the simplest case, where τ is constant and $X \in \mathbb{R}$. Then, we need to solve (recall that the first term is conditional **variance**; the second term bounds worst-case **bias**)

$$\text{minimize}_{\gamma,\, t} \quad \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + B^2 t^2$$

subject to:

$$\sum_{i=1}^{n} \gamma_i f(X_i) \leq t \text{ for all } f \text{ s.t. } f(c) = 0,\ f'(c) = 0,\ \left|f''(x)\right| \leq 1,$$

$$\sum_{i=1}^{n} W_i \gamma_i = 1,\ \sum_{i=1}^{n} (1 - W_i) \gamma_i = -1,\ \sum_{i=1}^{n} \gamma_i (X_i - c) = 0.$$

The first step is to re-write this via **convex duality**.

Solution via convex duality

By **duality**, the following problem is equivalent to the original:

$$
\begin{aligned}
&\text{maximize}_{f,\,\lambda} \\
&\quad \text{argmin}_{\gamma,\,t} \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + B^2 t^2 + \lambda_1 \left(\sum_{i=1}^{n} \gamma_i f(X_i) - t \right) \\
&\qquad + \lambda_2 \left(\sum_{i=1}^{n} W_i \gamma_i - 1 \right) + \lambda_3 \left(\sum_{i=1}^{n} (1 - W_i) \gamma_i + 1 \right) \\
&\qquad + \lambda_4 \sum_{i=1}^{n} \gamma_i (X_i - c),
\end{aligned}
$$

subject to:

$$
\begin{aligned}
&f(c) = 0, \;\; f'(c) = 0, \;\; \left| f''(x) \right| \leq 1 \text{ for all } x \in \mathbb{R}, \\
&\lambda_1 \geq 0, \quad \lambda_2, \, ..., \, \lambda_4 \in \mathbb{R}.
\end{aligned}
$$

The **inner minimization problem** is quadratic, and so can be solved in **closed form**, e.g., $t = 1/(2B^2)$, etc.

Solution via convex duality

Solving for γ and t in closed form, and some mild reparametrization, the problem **simplifies**:

$$\text{maximize}_{f,\lambda} \quad \frac{1}{4}\sum_{i=1}^{n}\frac{G_i^2}{\sigma_i^2} + \frac{\lambda_1^2}{4B^2} + \lambda_2 - \lambda_3$$

subject to:

$$G_i = f(X_i) + \lambda_2 W_i + \lambda_3(1 - W_i) + \lambda_4(X_i - c)$$

$$f(c) = 0,\ f'(c) = 0,\ \left|f''(x)\right| \le \lambda_1 \text{ for all } x \in \mathbb{R},$$

$$\lambda_1 \ge 0, \quad \lambda_2, \ldots, \lambda_4 \in \mathbb{R},$$

where the **original parameters** of interest are implicitly defined as

$$\hat{\gamma}_i = -\frac{\widehat{G}_i}{2\sigma_i^2}, \quad \hat{t} = \frac{\hat{\lambda}_1}{2B^2}.$$

This is just a **quadratic program** over the space of **twice differentiable functions**; can be solved via standard methods.

A remaining question

How should we select the **curvature parameter** B?

- **Impossible to be automatic**; see Armstrong and Kolesár (2018), as well as references therein.
- Requires collaborating with the **subject-matter expert** to exploit further (implicit?) regularity.

In the above examples, we tried the following strategies:

- Fit a **global quadratic** for both the treated and control samples. Set B to double the estimated curvature (used for the "effect of education" and "effect of summer school" problems)?
- Fit a flexible **non-parametric model** for $\mathbb{E}\left[Y \,\middle|\, X = x\right]$, and examine its worst-case curvature (used for "political advertising") example?

The first approach may fail by missing local effects not reflected in the global quadratic; the second approach may fail due to regularization in the non-parametric model. I use the first unless there is strong evidence the quadratic model doesn't fit the data.

Closing thoughts

Convex optimization presents a practical approach to **powerful** inference of causal effects in **complex RDD** problems.

I expect this to be a fruitful area for hybrid methodological/applied work that goes beyond classical regression-based approaches. Several challenges remain:

- How should one aggregate information across **multiple experiments** with RDDs?
- How do **fuzzy RDDs** interact with the methods discussed in this lecture?
- How should one add **covariates** to complex RDDs?
- What is the best way to estimate **treatment heterogeneity** along the boundary in an RDD?

When done correctly, the use of machine learning for causal inference can make the link between **identification** and **estimation** more explicit.