Econometrics

TA Session 8

Lucia Sauer

2025-11-20

Overview

- Regression Discontinuity Design
- Examples
- Sharp RD
- Fuzzy RD
- Assignment 7

The soul of RDD

RDD is all about finding "jumps" in the probability of treatment as we move along some running variable. So where do we find these jumps? Where do we find these discontinuities?

From **policies or rules** that switch treatment on/off at a specific threshold:

- Arrest for drunk driving jumps when **BAC 0.08**
  (Hansen, 2015)

- Medicare eligibility jumps at **Age 65**
  (Card, Dobkin & Maestas, 2008)

- Neonatal intensive care jumps when **birthweight $<$ 1500g**
  (Almond et al., 2010; Barreca et al., 2011)

- R&D subsidies jump when **project scores exceed funding threshold**
  (Bronzini & Iachini, 2014)

How RDD Works

RDD requieres a **running variable**, a **cutoff**, and a **treatment** to measure the causal effect on **outcome**.

- The **running variable** is a continuous variable assigning units to treatment.
- The **cutoff** is a specific value of the running variable that separates treated and control units.
- The **treatment** is an intervention that is expected to drive a change in outcome.
- The **outcome** is the variable of interest that we want to measure the effect on due to treatment.
- **Causal effect**: is known as the Local Average Treatment Effect (LATE) at the cutoff, since it is measured using the units around the cutoff.

Assumptions:

1. Local Relevance The probability of receiving treatment jumps at the cutoff.

$$lim_{z \to z_0^+} P(D = 1|Z = z) \neq lim_{z \to z_0^-} P(D = 1|Z = z)$$

Small differences in the running variable near the cutoff must change treatment uptake.

2. Exogeneity In the absence of treatment, potential outcomes must be smooth through the cutoff

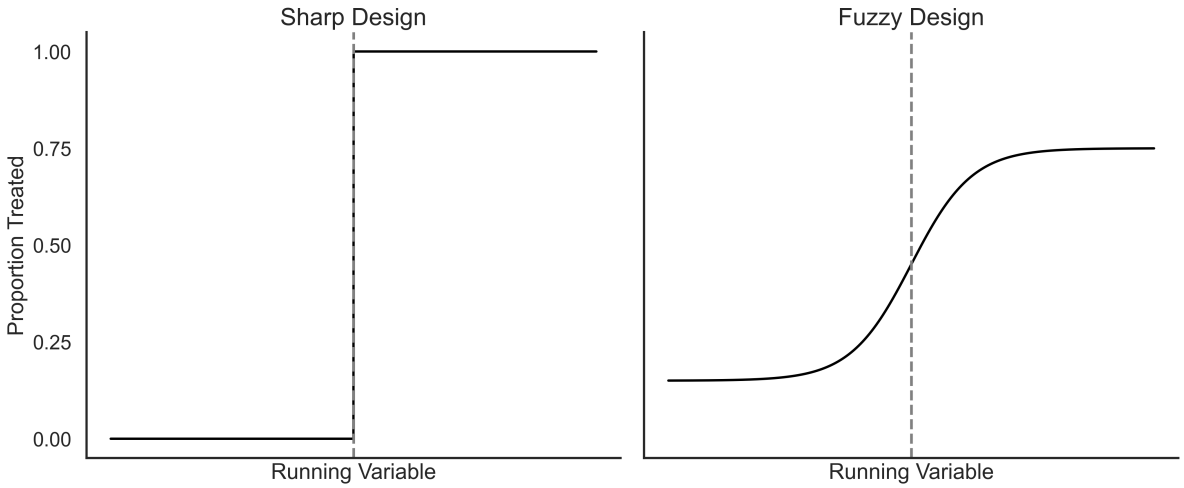
$$lim_{z \to z_0^+} P(Y_{ij} \leq r|Z_i = z_0) = lim_{z \to z_0^-} P(Y_{ij} \leq r|Z_i = z_0)$$

No other factor should change discontinuously at the cutoff.

---

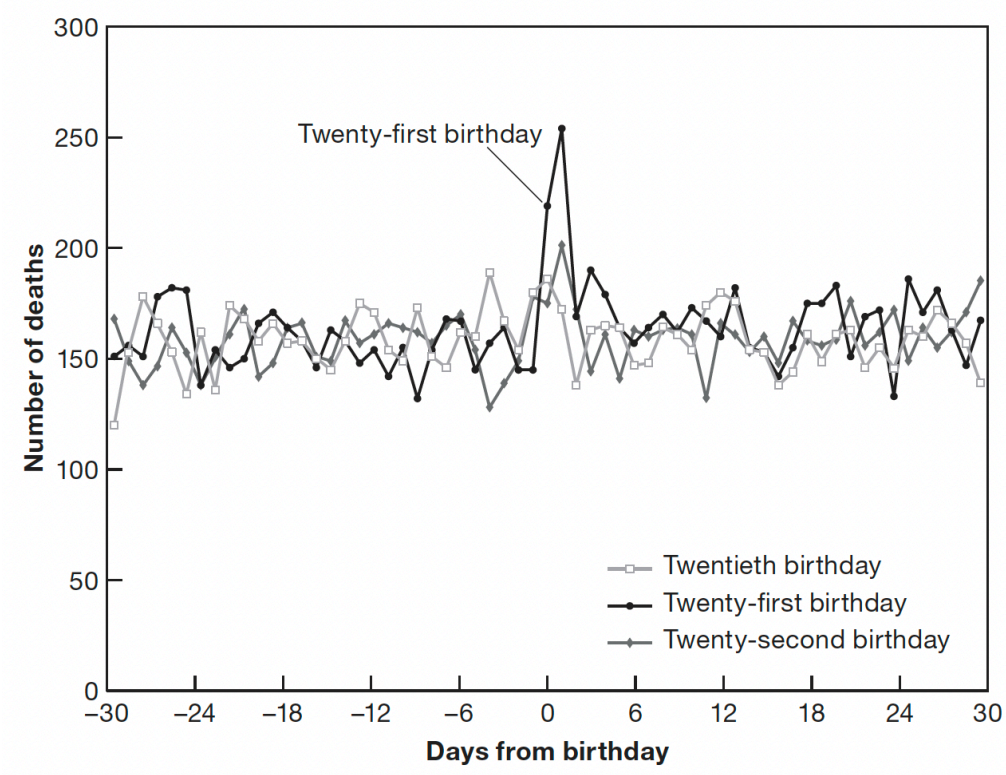
2 types of RDD

- In sharp RD, treatment assignment is **deterministic** at the threshold, so the RD estimand equals the **ATE** for all units at the cutoff.
- In fuzzy RD, assignment is only **probabilistic**, so we recover a **LATE** for compliers at the cutoff.



Mortality Risk and the MLDA

- Age 21 is the **legal drinking age** in the U.S.
- A tiny change in age → **big jump** in legal access
- Mortality **spikes exactly at 21**, not a generic "party-hardy" birthday effect.



Causal question:
Does legal access to alcohol **increase mortality**?

Sharp RD

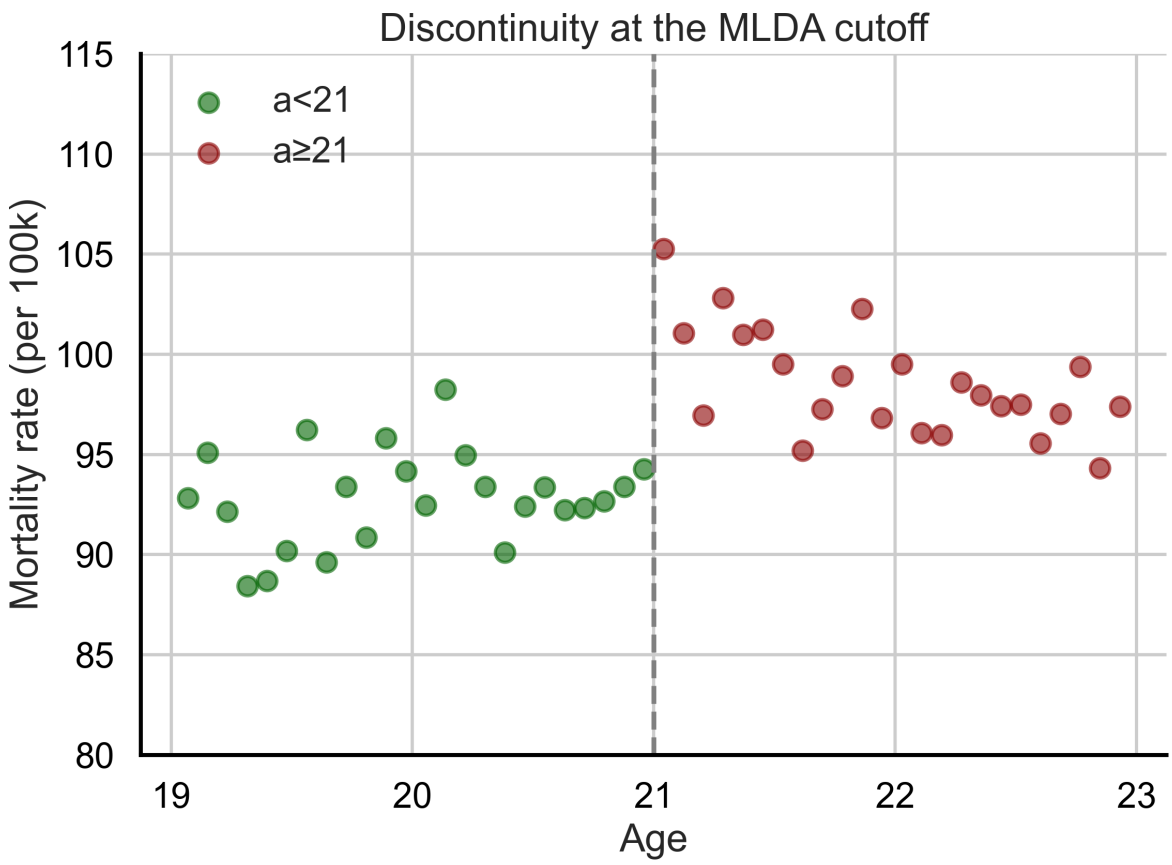
The effects of MLDA on mortality is a sharp RD, because the treatment swithches cleanly off or on as the running variable crosses the cutoff.

The treatment is defined as:

$$D_a = \begin{cases} 1 & \text{if } a \geq 21 \\ 0 & \text{if } a < 21 \end{cases}$$

where D_a is the treatment indicator at age a and 1 means that the individual is legally allowed to purchase alcohol.

Sharp RD visualization



Sharp RD equation:

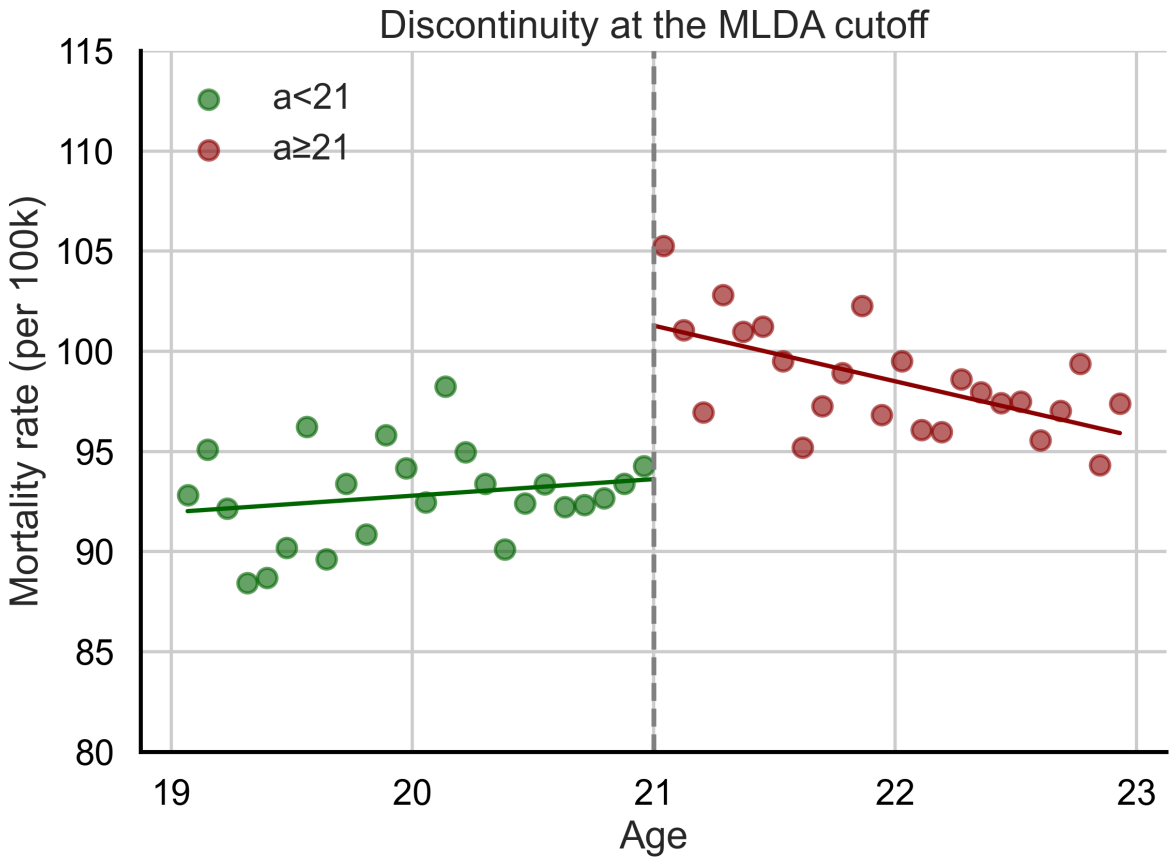
$$\bar{M}_a = \beta_0 + \beta_1(a - a_0) + \alpha_{RD} D_a + \beta_2(a - a_0)D_a + e_a$$

where:

- $\bar{M}_a$ is the average mortality rate in month a (each month is a 30-day interval relative to the 21st birthday),
- a is age in months and a_0 is the cutoff (the month in which individuals turn 21), so $(a - a_0) = 0$ at the threshold,
- $D_a = 1(a \geq a_0)$ is an indicator for being above the MLDA,

- α_{RD} is the RD parameter of interest, capturing the causal jump in mortality at the cutoff,
- β_1 captures the age-related trend in mortality below the cutoff,
- β_2 allows the slope of mortality with respect to age to differ above the cutoff.

Visualization of RD fit



- **Turning 21 increases mortality by ~7.7 deaths per 100k** at the cutoff.
- In a **sharp RD**, this jump is the **ATE at age 21**.

Is this causal?

- Yes — if age trends are correctly modeled.
- In a sharp RD, treatment is fully determined by the running variable (age), so once we control for age trends, there is no omitted variable bias.
- The causal identifying assumption is that mortality would have been smooth in age in the absence of the MLDA cutoff. The jump we see at 21 is then the ATE at the cutoff.

Linear age trend

- So the **key question** becomes whether our model captures the age–mortality relationship well enough.
- If the trend is truly linear near the cutoff, this is a credible causal estimate.

Dealing with Nonlinear Trends in RD

- So far we've assumed a **linear** age–mortality relationship.
- But if trends are nonlinear, a simple linear RD may **mistake nonlinearity for a discontinuity** → biased estimates.

Two common solutions:

1. Local linear regression

Focus on observations **close to the cutoff**, where linearity is more credible.

2. Flexible polynomials

Allow curvature using **quadratic or higher-order** functions of the running variable.

Quadratic RD

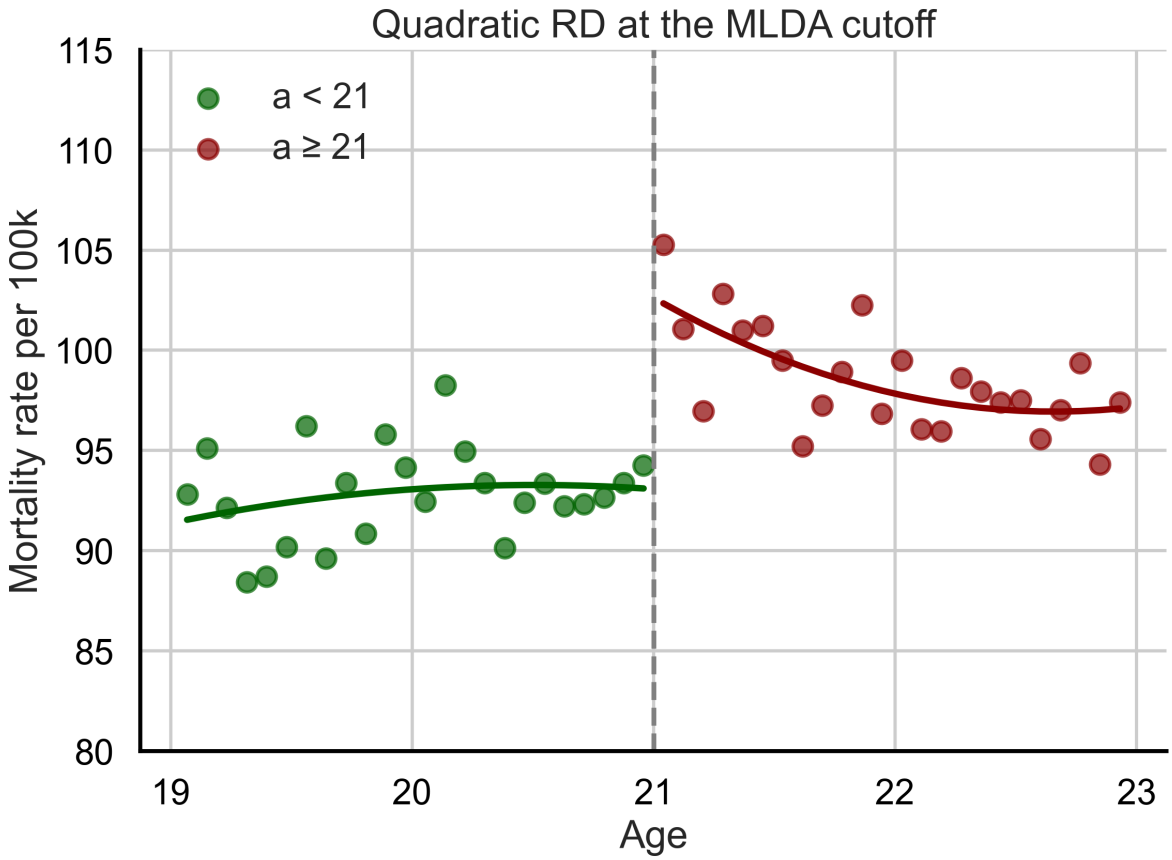
The figure shown before suggests the possibility of **mild curvature** in the relationship between $\bar{M}_a$ and a, at least for the points to the right of the cutoff.

$$\bar{M}_a = \beta_0 + \beta_1(a - a_0) + \beta_2(a - a_0)^2 + \alpha_{RD}D_a \qquad + \beta_3(a - a_0)D_a + \beta_4(a - a_0)^2 D_a + e_a$$

- Below 21: drinking is restricted → mortality tends to **decline** as youth mature.
- Above 21: legal access may **change the trend** (more drinking risk or faster maturity).
- RD lets the **slope differ** on each side of the cutoff instead of assuming a single trend.

Quadratic RD Fit

- The effect is still there and is stronger: turning 21 increases mortality by about 9.5 deaths per 100K.
- The estimated trend function generated has some curvature, mildly concave to the left of age 21 and markedly convex thereafter.



What drives the effect?

- If the MLDA truly causes the mortality jump at 21…
- …we should see the largest increase in **alcohol-related deaths**.

Validation strategy

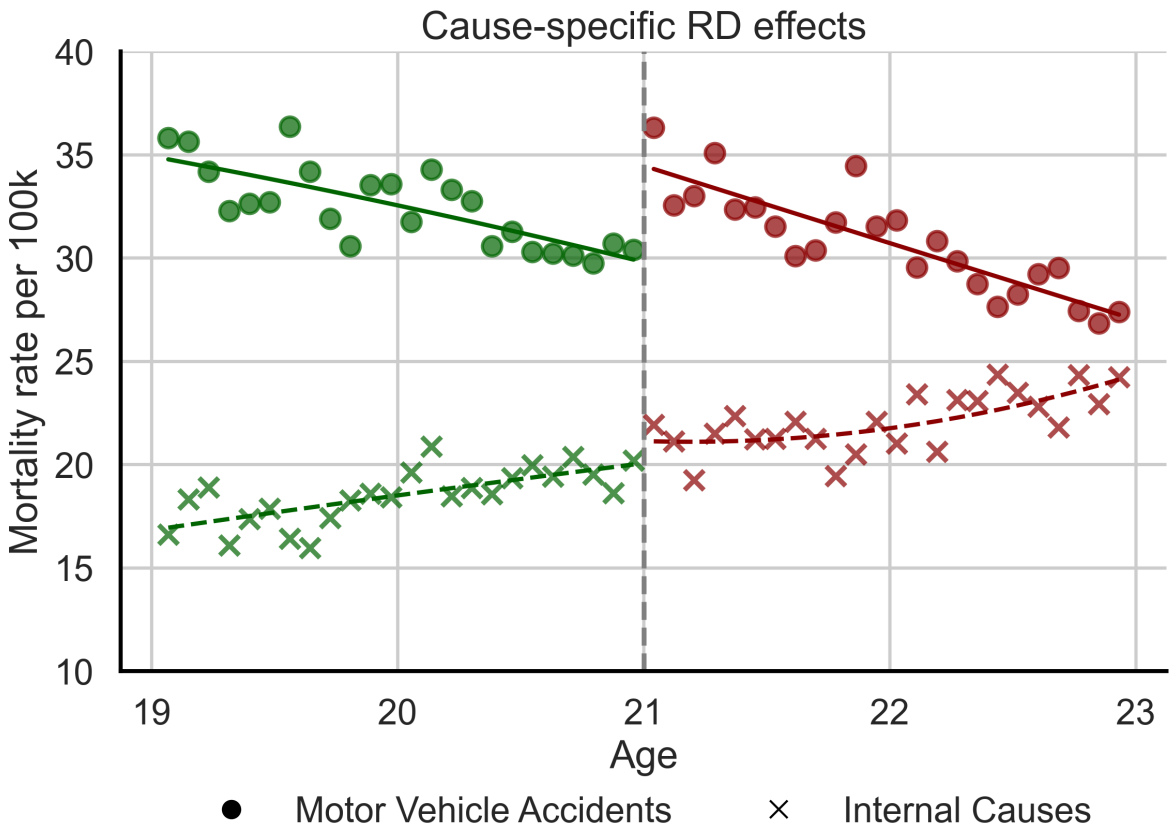
Compare causes:

- **Motor vehicle accidents** → strongly alcohol-related
- **Internal causes** → unrelated to drinking

Prediction: A real alcohol effect → a discontinuity only in MVA deaths.

Cause-specific RD effects

- Large jump in **motor vehicle deaths**
- Little to no jump in deaths from **internal causes**



Deaths rise exactly where alcohol matters most: on the road.

Robustness:

- **Reliable findings** must survive:
  - changes in functional form (linear → quadratic)
  - changes in the sample window (global → local)
  - changes in the outcome (alcohol-related vs. placebo)

> **Note**
>
> **Takeaway:**
> - Effects on **all causes** and **MVA deaths** are large & significant
> - **Internal causes** show ≈ 0 effect → placebo passed

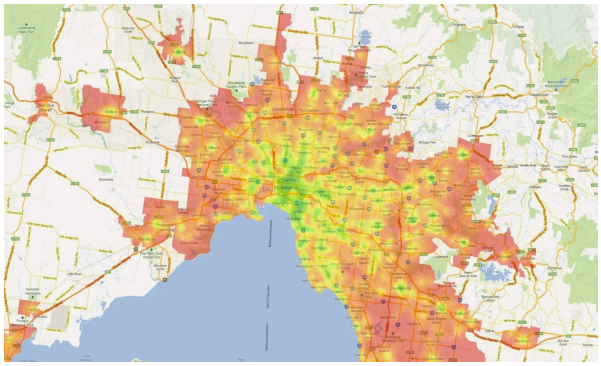
Results

Table 1: RD Estimates of MLDA on Mortality

	Cause	Linear (Global)	Quadratic (Global)	Linear (20-22)	Quadratic (20-22)
0	All Causes	7.66 (1.44)	9.55 (1.99)	9.75 (1.94)	9.61 (2.89)
1	Motor Vehicle	4.53 (0.77)	4.66 (1.15)	4.76 (1.12)	5.89 (1.56)
2	Suicide	1.79 (0.45)	1.81 (0.70)	1.72 (0.71)	1.30 (1.13)
3	Homicide	0.10 (0.39)	0.20 (0.52)	0.16 (0.52)	-0.45 (0.79)
4	Internal Causes	0.39 (0.60)	1.07 (0.91)	1.69 (0.76)	1.25 (1.20)
5	Alcohol	0.44 (0.16)	0.80 (0.22)	0.74 (0.23)	1.03 (0.33)

Uber Surge Pricing

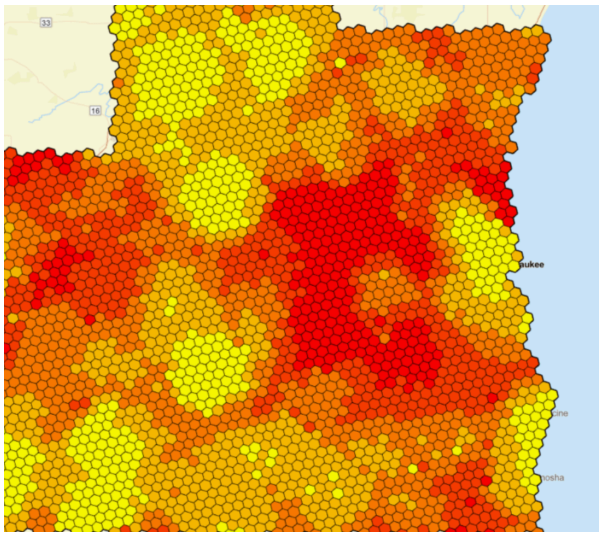
- Many riders request trips but too few drivers are active, the **market becomes imbalanced**.
- Uber activates **surge pricing** to restore equilibrium: fares increase temporarily.



Causal question:
Does surge pricing causally increase driver supply?

How Uber Activates Surge Pricing

- Uber's algorithms monitor supply and demand in real time, by dividing the city into **hyper-local zones**, and calculates a **market imbalance score** for each zone.
- If **demand > supply beyond a threshold**, surge activates in that cell → **price jump**.



The threshold creates **price discontinuities** that enable causal inference.

Why Fuzzy?

- Treatment: $D =$ **surge price received**
- Outcome: $Y =$ **driver supply** (number of drivers)
- Running variable: $Z =$ **market imbalance score**
- Cutoff: $z_0 =$ surge threshold
- The instrument: $1\{Z \geq z_0\}$ eligibility for higher prices

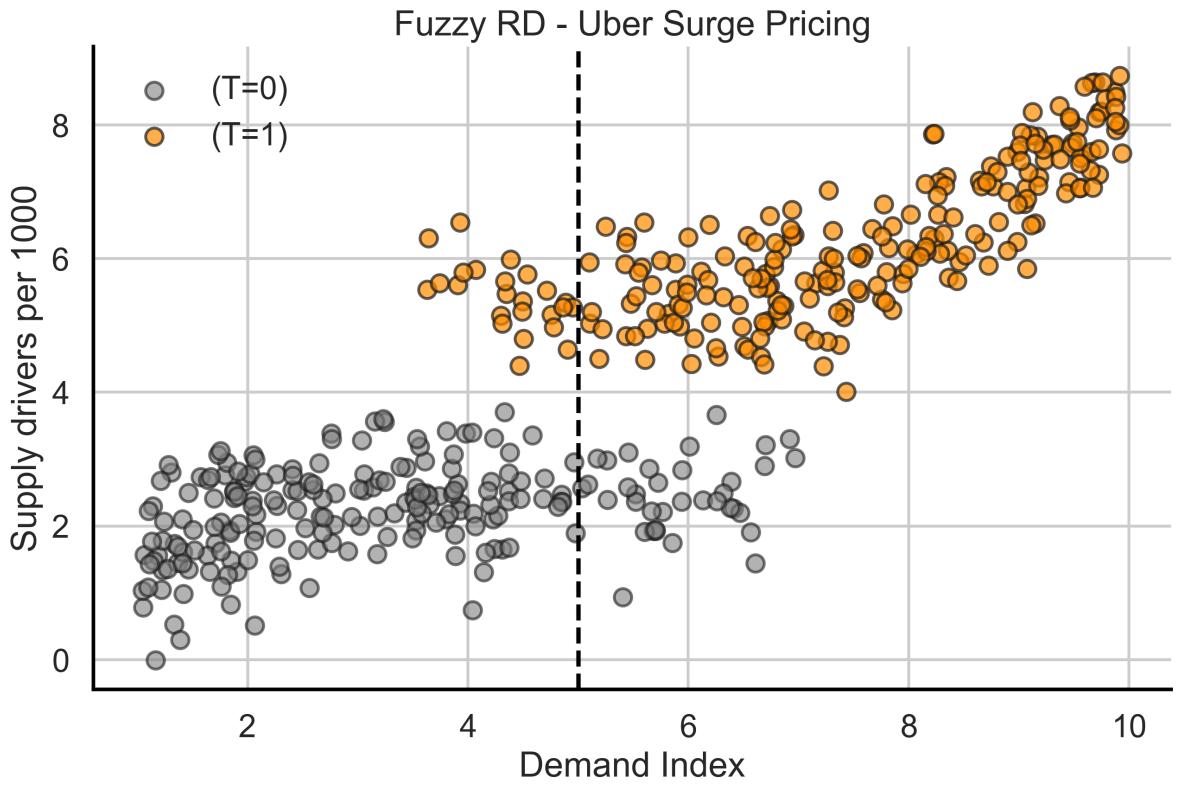
But the probability of receiving a surge price is not deterministic at the cutoff:

$$0 < P(D = 1|Z \geq z_0) < 1$$

- Not all drivers that receive surge work more (defiers).
- Some work regardless of surge (always-takers)
- Some never work (never-takers)

Fuzzy Illustration:

Orange dots: drivers that receive surge pricing (T=1) and gray dots: drivers that do not (T=0).



RD Fuzzy estimation

$$Y = \beta_0 + \beta_1 Z' + \alpha_{FRD} 1\{Z \geq z_0\} + \beta_2 Z' \cdot 1\{Z \geq z_0\} + \epsilon$$

```
cutoff = 5
df['Instr'] = (df['Z'] >= cutoff).astype(int)
# Center running variable at the cutoff
df['Z_centered'] = df['Z'] - cutoff

bandwidth = 2  # ej. ±2 units around cutoff
```

```
df_local = df[np.abs(df['Z_centered']) <= bandwidth]

# exogenous: constant + running variable centered
exog = sm.add_constant(df_local['Z_centered'])
# Endogenous: treatment T
endog = df_local['T']
# Instrument: crossing the cutoff
instr = df_local['Instr']

iv_model = IV2SLS(dependent=df_local['Y'],exog=exog,endog=endog,instruments=instr)
result = iv_model.fit(cov_type='robust')
print(result.summary)
```

Wald Estimator

$$\alpha_{FRD} = \frac{lim_{Z \to 5^+} E[Y|Z = z_0] - lim_{Z \to 5^-} E[Y|Z = z_0]}{lim_{Z \to 5^+} E[T|Z = z_0] - lim_{Z \to 5^-} E[T|Z = z_0]}$$

$$\alpha_{FRD} = \frac{\alpha_{SRD}}{\delta}$$

where δ is the proportion of drivers that receive surge pricing when demand crosses the threshold (compliers)

```
#First stage
fs = smf.ols("T ~ Instr + Z_centered", data=df_local).fit()
#Second stage
ss = smf.ols("Y ~ Instr + Z_centered", data=df_local).fit()

pi1 = fs.params["Instr"]
tau = ss.params["Instr"]
alpha_FRD = tau / pi1
```

Assignment 7

Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter, Oreopoulos (2006)

- What is the effect of **compulsory schooling laws** on **earnings** and **labor market outcomes**?
- Previous studies use compulsory schooling laws as IV, but they only affect small portion of the population because many students stayed in school beyond compulsory age.
- UK school-leaving age change in 1947 from 14 to 15 affected almost half of students.

Methodology

- Regression Discontinuity (RD): compares student cohort turning 15 just before and after the reforms in 1947 (UK).

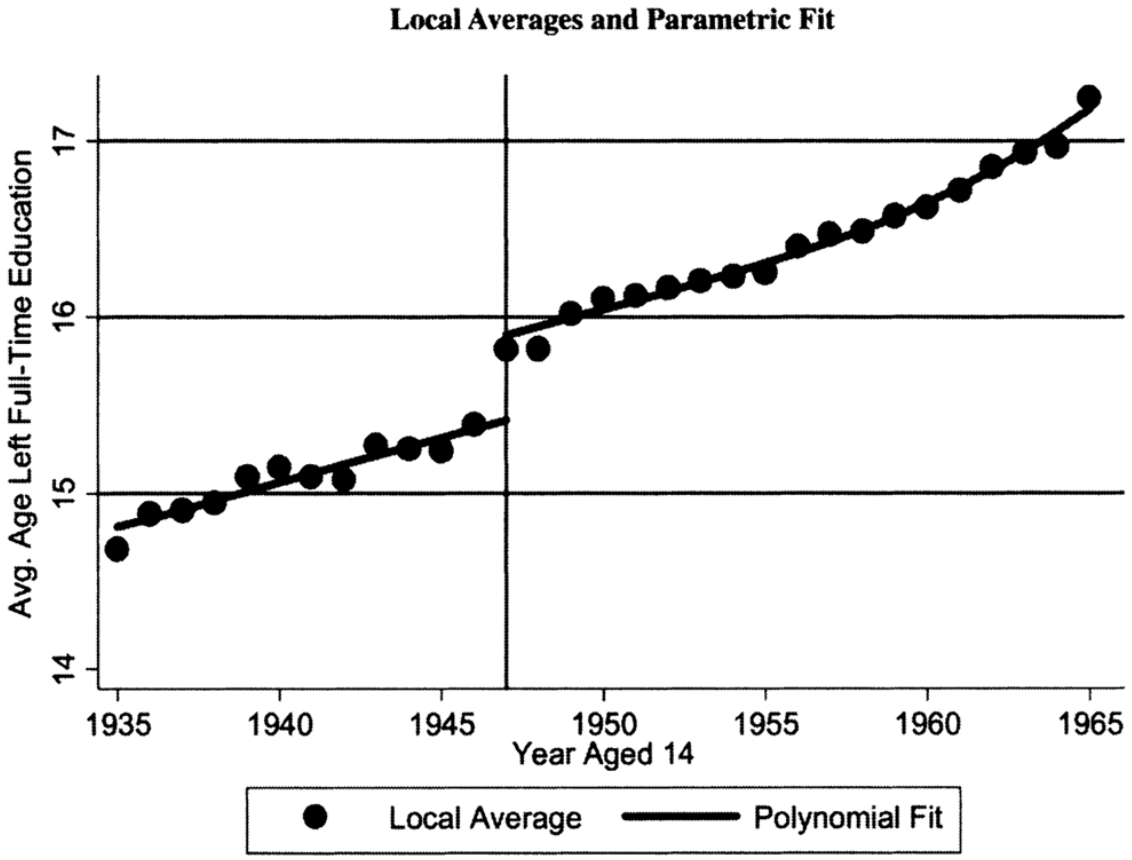


FIGURE 4. AVERAGE AGE LEFT FULL-TIME EDUCATION BY YEAR AGED 14
(*Great Britain*)

Note: Local averages are plotted for British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys. The curved line shows the predicted fit from regressing average age left full-time education on a birth cohort quartic polynomial and an indicator for the school-leaving age faced at age 14. The school-leaving age increased from 14 to 15 in 1947, indicated by the vertical line.

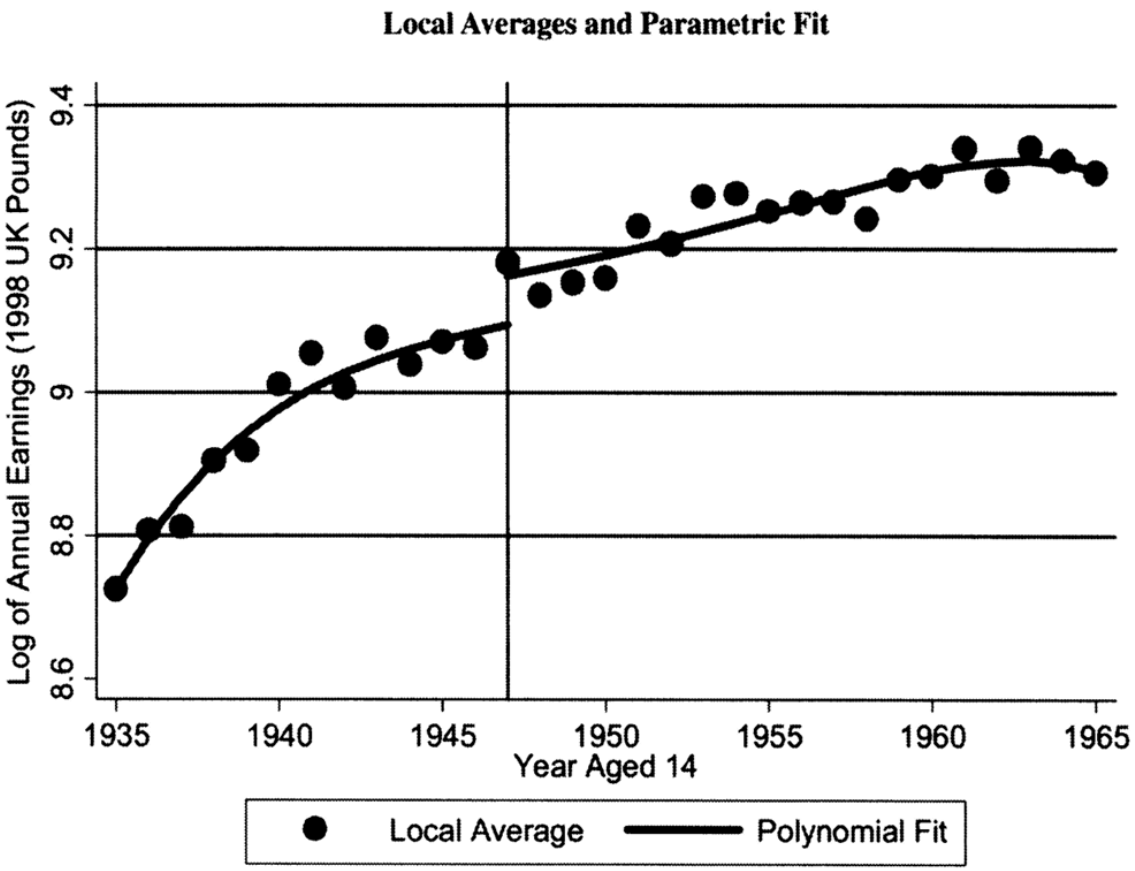


FIGURE 6. AVERAGE ANNUAL LOG EARNINGS BY YEAR AGED 14
(*Great Britain*)

Note: Local averages are plotted for British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys. The curved line shows the predicted fit from regressing average log annual earnings on a birth cohort quartic polynomial and an indicator for the school-leaving age faced at age 14. The school leaving age increased from 14 to 15 in 1947, indicated by the vertical line. Earnings are measured in 1998 U.K. pounds using the U.K. retail price index.

Data

- UK General Household Survey (1983-1998) and Northern Ireland Household Survey (1985-1998).

- It contains information on 30,487 individuals who were aged 14 between 1935 and 1965, and 32 to 64 years old at the time of the survey.

	sex	age	agelfted	nireland	yearat14	wght	learn	drop15	yearat14	yearat14_2	yearat1
0	male	63	14.0	0.0	35.0	35	8.425849	0.0	35.0	1225.0	42875.0
1	female	63	14.0	0.0	35.0	16	7.332621	0.0	35.0	1225.0	42875.0
2	male	63	15.0	0.0	35.0	3	9.024424	0.0	35.0	1225.0	42875.0
3	female	63	15.0	0.0	35.0	1	5.560682	0.0	35.0	1225.0	42875.0
4	male	63	16.0	0.0	35.0	2	8.811445	0.0	35.0	1225.0	42875.0
5	female	63	16.0	0.0	35.0	3	6.721386	0.0	35.0	1225.0	42875.0
6	male	63	17.0	0.0	35.0	2	8.880419	0.0	35.0	1225.0	42875.0
7	female	63	17.0	0.0	35.0	1	7.886788	0.0	35.0	1225.0	42875.0

Exercise 1 - tips

"Replicate" Figures 4 and Figure 6

- Restrict sample: **British-born**, **age 32–64**, **earnings > 0**
- Compute **weighted cohort averages**:

$$\bar{Y}_t = \frac{\sum w_i Y_i}{\sum w_i}$$

- Fit **separate quartic polynomials** on each side of the cutoff (z = 1947):

$$\bar{Y} = \begin{cases} \beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3 + \beta_4 z^4 + \varepsilon, & z < 1947 \\ \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \gamma_3 z^3 + \gamma_4 z^4 + \varepsilon, & z \geq 1947 \end{cases}$$

```
import statsmodels.api as sm
import numpy as np

coef_before_1947 = np.polyfit(df_before_1947['yearat14'], df_before_1947['weighted_avg_outcor
poly_before_1947 = np.poly1d(coef_before_1947)


x_before_1947 = np.linspace(df_before_1947['yearat14'].min(), df_before_1947['yearat14'].max
y_before_1947 = poly_before_1947(x_before_1947)
```

Exercise 2 - tips

Replicate Table 1 for GB.

TABLE 1—ESTIMATED EFFECT OF MINIMUM SCHOOL-LEAVING AGE ON AGE FINISHED FULL-TIME EDUCATION AND LOG ANNUAL EARNINGS
(*Great Britain and Northern Ireland, ages 25–64, 1935–1965*)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(First stage) dependent variable: Age finished full-time school			(Reduced form) dependent variable: log annual earnings			
Sample population							Initial sample size
Great Britain	0.440	0.436	0.453	0.065	0.064	0.042	57264
	[0.065]***	[0.071]***	[0.076]***	[0.025]**	[0.026]**	[0.043]	
Northern Ireland	0.397	0.391	0.353	0.054	0.074	0.074	8921
	[0.074]***	[0.073]***	[0.100]***	[0.27]*	[0.025]***	[0.045]	
G. Britain and N. Ireland with	0.418	0.397	0.401	0.073	0.058	0.059	66185
N. Ireland Fixed Effect	[0.040]***	[0.043]***	[0.045]***	[0.016]***	[0.016]***	[0.018]***	
Birth Cohort Polynomial Controls	Quartic	Quartic	Quartic	Quartic	Quartic	Quartic	
Age Polynomial Controls	None	Quartic	None	None	Quartic	None	
Age Dummies	No	No	Yes	No	No	Yes	

Notes: The dependent variables are age left full-time education and log annual earnings. Each coefficient is from a separate regression. Each regression includes controls for a birth cohort quartic polynomial and indicator whether a cohort faced a school leaving age of 15 at age 14. Columns 2, 3, 5, and 6 also include age controls: a quartic polynomial and fixed effects where indicated. Each regression includes the sample of 25- to 64-year-olds from the 1983 through 1998 General Household Surveys, who were aged 14 between 1935 and 1965. Data are first aggregated into cell means and weighted by cell size. Regressions are clustered by birth cohort and region (Britain or N. Ireland).

Hint: use the person weight and cluster standard errors by birth cohort and region.

```
import statsmodels.api as sm
X = df[regression_vars]
y = df[outcome_var]
model = sm.WLS(y, sm.add_constant(X), weights=weights)
results = model.fit(cov_type='cluster', cov_kwds={'groups': df[yobirthvar]})
```

Exercise 3 - tips

Estimate RD-IV as in Table 2 for Great Britain

TABLE 2—OLS AND IV RETURNS TO (COMPULSORY) SCHOOLING ESTIMATES FOR LOG ANNUAL EARNINGS
(*Great Britain and Northern Ireland, ages 25–64, 1935–1965*)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Returns to schooling: OLS			Returns to compulsory schooling: IV			
							Initial sample size
Great Britain	0.078	0.079	0.079	0.147	0.145	0.149	57264
	[0.002]***	[0.002]***	[0.002]***	[0.061]**	[0.063]**	[0.064]**	
Northern Ireland	0.111	0.113	0.113	0.135	0.187	0.21	8921
	[0.004]***	[0.004]***	[0.004]***	[0.071]*	[0.070]**	[0.135]	
G. Britain and N. Ireland with	0.082	0.082	0.083	0.174	0.149	0.148	66185
N. Ireland fixed effect	[0.001]***	[0.001]***	[0.001]***	[0.042]***	[0.044]***	[0.046]***	
Birth cohort polynomial controls	Quartic	Quartic	Quartic	Quartic	Quartic	Quartic	
Age polynomial controls	None	Quartic	None	None	Quartic	None	
Age dummies	No	No	Yes	No	No	Yes	

Notes: The dependent variable is log annual earnings. Each regressions includes controls for a birth cohort quartic polynomial and age left full-time education (instrumented by an indicator whether a cohort faced a school leaving age of 15 at age 14 in columns 4 through 6). Columns 2, 3, 5, and 6 also include age controls: a quartic polynomial and fixed effects where indicated. Each regression includes the sample of 25- to 64-year-olds from the 1983 through 1998 General Household Surveys who were aged 14 between 1935 and 1965. Data are first aggregated into cell means and weighted by cell size. Regressions are clustered by birth cohort and region (Britain or N. Ireland).

```
from linearmodels.iv import IV2SLS
iv_model = "outcome ~ yearat14 + yearat14_2 + yearat14_3 + yearat14_4 + [agelfted ~ instrumer
iv_result = IV2SLS.from_formula(iv_model, df, weights=df['wght']).fit(cov_type='kernel')
```

References

Books - Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Alcohol Policy & Mortality RD - Carpenter, C., & Dobkin, C. (2009). *The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age*. American Economic Journal: Applied Economics, 1(1), 164–182.

Surge Pricing & Uber - Chen, M. K., & Sheldon, M. (2016). *Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform*. American Economic Review Papers & Proceedings, 106(5), 177–182. - Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). *Using Big Data to Estimate Consumer Surplus: The Case of Uber*. NBER Working Paper No. 22627.

Web Resources (for background on surge pricing) - "How does Uber do surge pricing using location data?"
https://anubpattnaik.medium.com/how-does-uber-do-price-surge-using-location-data-cfee03415022 - "Surge Pricing." Uber Marketplace
https://www.uber.com/us/en/marketplace/pricing/surge-pricing/