



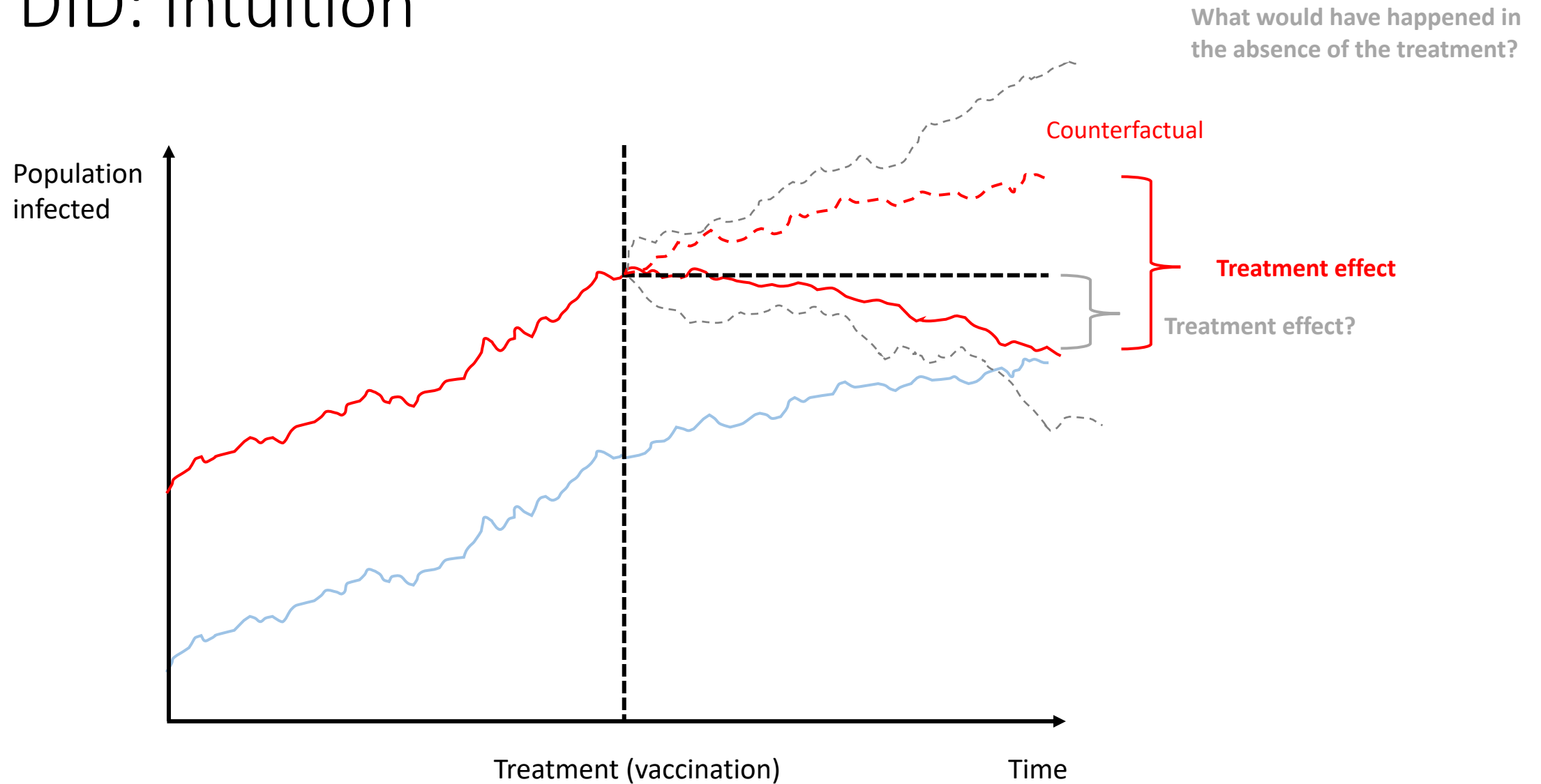
Introduction of empirical methods in modern econometrics

Session 3: Differences-in-differences,
Regression discontinuity and examples

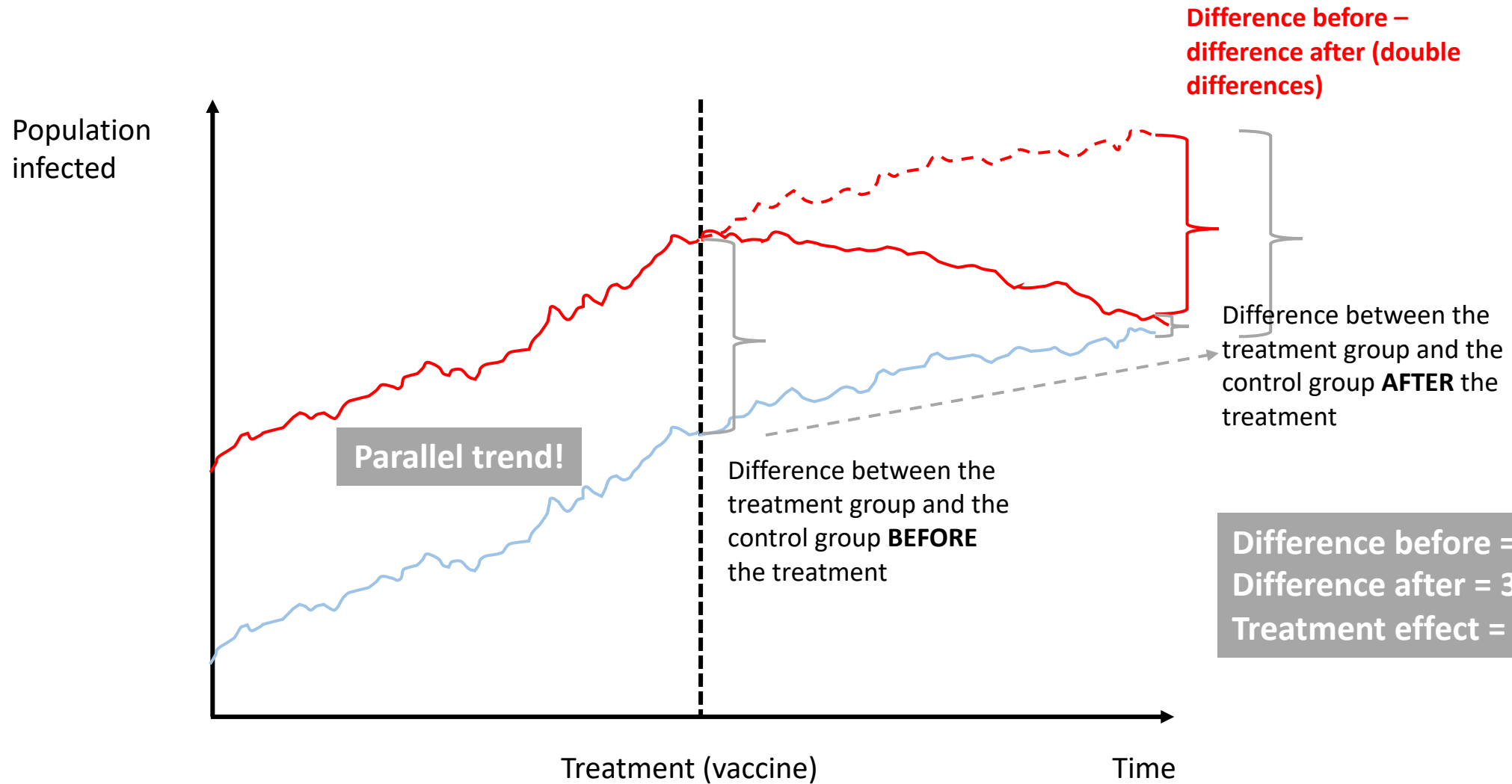
Dianzhuo ZHU

2021-10-01

DID: Intuition



DID: Intuition



DID: The estimation equation

$$Y_{it} = \alpha + \beta(TREAT_i * POST_t) + \delta TREAT_i + \gamma POST_t + \eta X_{it} + \epsilon_{it}$$

- $TREAT_i$: treatment dummy, 1 if observation i is in the treatment group; 0 otherwise
- $POST_t$: period dummy, 1 if observation i is in the treatment period; 0 otherwise
- X_{it} are the covariates (other independent variables) that we want to control in this equation
- **β indicates the differences-in-differences treatment effect!**

| Expected outcome Y_{it} (not including X_{it}) | $TREAT_i = 1$ | $TREAT_i = 0$ | $TREAT_i = 1 -$ $TREAT_i = 0$ |
|--|------------------------------------|-------------------|----------------------------------|
| $POST_t = 1$ | $\alpha + \beta + \delta + \gamma$ | $\alpha + \gamma$ | $\beta + \delta$ |
| $POST_t = 0$ | $\alpha + \delta$ | α | δ |
| $POST_t = 1 - POST_t = 0$ | $\beta + \gamma$ | γ | β |

DID: The parallel trend assumption

- **Some treatment decisions are based on the outcome (endogenous)**

- For example, police force is strengthened in areas where the crime level is higher (**ABSOLUTE VALUE**)

- But DID is still validate as long as the policy is not based on the crime increase rate (**TREND**)

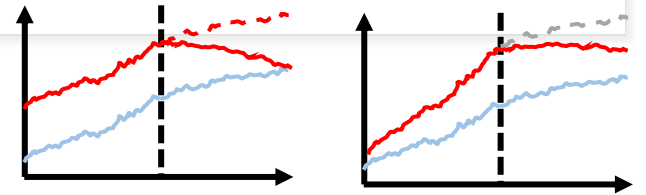
- Machin and Marie (2011): Street Crime Initiative in 10 out of 43 police forces of England and Wales in 2002/03

- The 10 treated services have higher crime levels, but they shall the parallel trend as the remaining 33

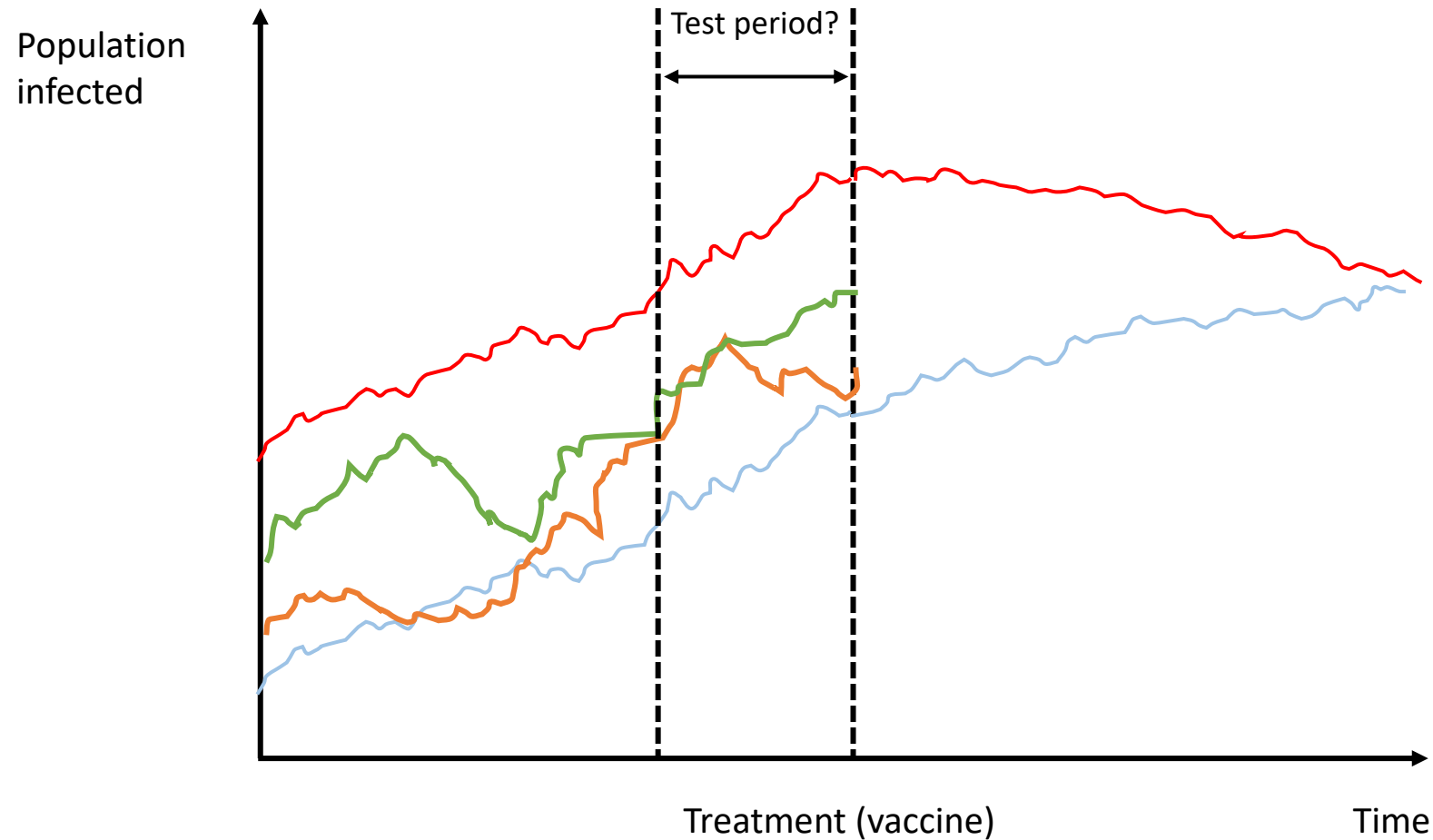
- But the treatment may as well due to **an exogenous “shock”**

- Unpredictable natural and social events, policy implementations...

- Di Tella and Schargrodsky (2004): intensified police presence that occurred around religious buildings in Buenos Aires following a terrorist attack



How to test the parallel trend assumption?



How to test the parallel trend assumption?

$$Y_{it} = \gamma_i + \lambda_t + \sum_{\tau=-k}^{\tau=-1} \delta_{\tau} D_{i\tau} + \sum_{\tau=1}^{\tau=m} \delta_{\tau} D_{i\tau} + X_{it}\beta + \epsilon_{it}$$

- $\tau = 0$ is the treatment period
- $D_{i\tau}$ is the treatment variable of i at period τ
- k defines how many pre-treatment periods we would like to test. Say we take $k = 2$
- If the trends are parallel, δ_{-2} and δ_{-1} should not predict the outcome

DID: Strengths and Limitations

- Strengths
 - Intuitive interpretation
 - Can obtain causal effect using observational data if assumptions are met
 - Can use either individual and group level data
 - Comparison groups can start at different levels of the outcome. (DID focuses on change rather than absolute levels)
- Limitations
 - Requires to identify treatment and control groups
 - Cannot be used if the intervention allocation is determined by the baseline trend
 - Cannot be used if comparison groups have non parallel trends
 - Abadie (2005) has proposed solution
 - Cannot use if the compositions of groups before and after the treatment are not stable

DID: Best practices

- Be sure that the outcome **trend** does not influence **allocation** of the treatment/intervention
- Acquire more data points before and after to test **parallel trend** assumption
 - Cleverly determine the pre-treatment length for the test
 - Always show the treatment/control, before/after graph
 - Pre-trend test may not be enough, justify with economic insights
- Be sure to examine the **composition of population** in treatment/intervention and control groups before and after the intervention
- It's also important to add other covariates in the model

DID: Paper example

- Di Tella, R., & Schargrodsky, E. (2004). Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review*, 94(1), 115-133.


DID: Paper example

- Difficulty in evaluating the impact of police force on crime
 - Endogeneity (reverse causality): the government of a city in which the crime rate increases will hire more police officers
 - How to break this endogeneity in order to identify CAUSAL effects of police on crime?

DID: Paper example

- Exogeneous shock: A terrorist attack
 - On July 18, 1994, terrorists exploded a bomb that destroyed the Asociación Mutual Israelita Argentina (AMIA), the main Jewish center in Argentina
 - One week later, the federal government assigned police protection to every **Jewish and Muslim building** in the country
 - The geographical distribution of these institutions can be presumed to be **EXOGENOUS** (uncorrelated) to crime (intuition + proof)
- Differences-in-differences approach
 - Before & after the attack
 - Treated (protected areas) & control groups: distance of each block in the sample to the nearest Jewish institution (one block away, two blocks away...)
 - Outcome variable: number of car thefts (why?)


DID: Paper example

- Data: Number of car thefts per block
 - In three neighborhoods in Buenos Aires
 - April 1 – December 31, 1994
 - Attack: July 18, 1994, Protection: since July 25, 1994
 - Previous levels of police presence maintained in the rest of the neighborhoods
- Results: **75% fall** of motor vehicle thefts in protected areas compared to control group, **large and significant** effect
 - The effect remains local: police presence in a given block does not reduce car thefts in neighboring blocks
 - Mechanism (how did the policy achieve this?)
 - Deterrence: police presence makes criminal activity less attractive 
 - Incapacitation: police officers arrest criminals, leaving fewer of them around to commit crimes

| Demographic characteristics | Census tracts without Jewish institutions (A) | Census tracts with Jewish institutions (B) | Difference (C) = (A) - (B) |
|-----------------------------|--|---|-------------------------------|
| Home ownership rate | 0.696 (0.008) | 0.663 (0.017) | 0.032 (0.019) |
| Overcrowding rate | 0.014 (0.001) | 0.017 (0.002) | -0.002 (0.003) |
| Poverty rate | 0.042 (0.003) | 0.052 (0.008) | -0.010 (0.009) |
| Education of household head | 11.653 (0.147) | 11.052 (0.300) | 0.600 (0.335) |
| Number of household members | 2.719 (0.023) | 2.685 (0.054) | 0.034 (0.059) |
| Female population | 0.556 (0.001) | 0.552 (0.003) | 0.003 (0.003) |
| Unemployment rate | 0.053 (0.001) | 0.059 (0.003) | -0.005 (0.003) |
| Age | 38.005 (0.128) | 37.690 (0.223) | 0.315 (0.258) |
| Number of census tracts | 53 | 14 | |

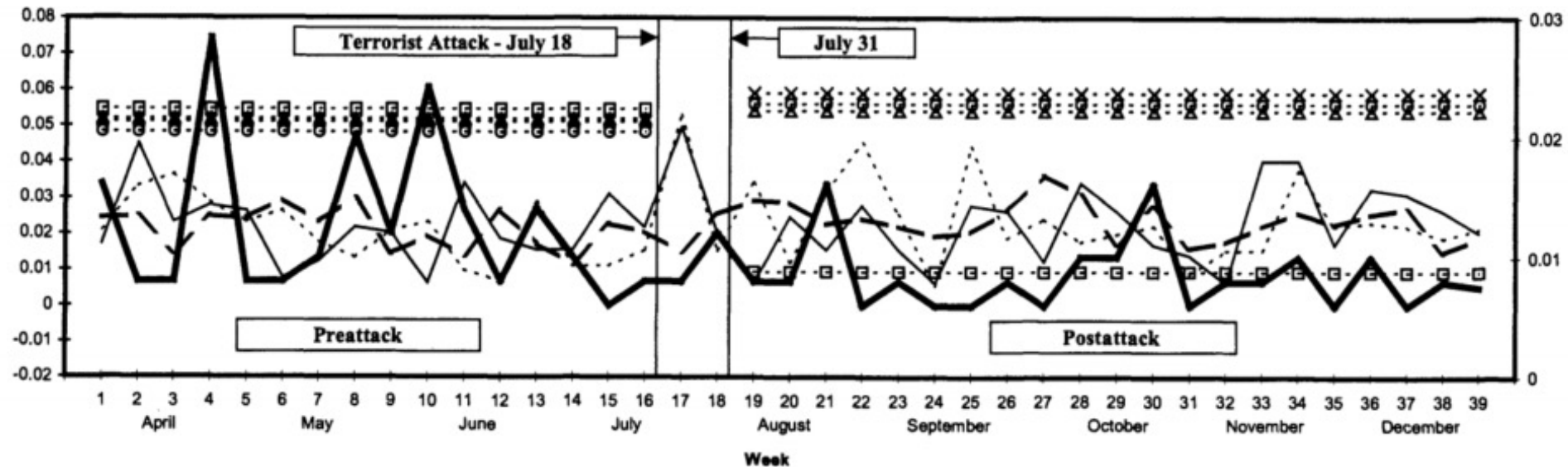
Notes: Columns (A) and (B) present the mean of each variable for census tracts without and with Jewish institutions in our sample. Column (C) presents the differences of means. Standard deviations are in parentheses. Home ownership rate is the percentage of owner-occupied houses. Overcrowding rate is the percentage of households with more than three people per room. Poverty rate is the percentage of households with at least one unmet basic need (overcrowding; four or more members per working member and household head with low educational attainment; poor quality housing; school-age children not attending school; or no fecal evacuation system). Education of the household head is the average educational attainment of the household head in number of years. Female population is the percentage of women in the total population. Unemployment rate is the rate of unemployment for the population of age 14 or higher. Age is the average age of the population.

Source: 1991 Population Census.



Protected and unprotected areas look similar

Weekly average car thefts of protected and unprotected areas were similar before the police reinforcement, but differ afterwards



By Week (Left Axis)

— Jewish Institution in the Block
 — One Block from Nearest Jewish Institution
 Two Blocks from Nearest Jewish Institution
 - - - More than Two Blocks from Nearest Jewish Institution

Means (Right Axis)

. . □ . . Pre and Post Means for Jewish Institution in the Block
 . . ☆ . . Pre and Post Means for One Block from Nearest Jewish Institution
 . . ○ . . Pre and Post Means for Two Blocks from Nearest Jewish Institution
 . . × . . Pre and Post Means for More than Two Blocks from Nearest Jewish Institution

DID: The regression

$$\begin{aligned}
 \text{Car Theft}_{it} = & \alpha_0 \text{Same Block Police}_{it} \\
 & + \alpha_1 \text{One Block Police}_{it} \\
 & + \alpha_2 \text{Two Blocks Police}_{it} \\
 & + M_t + F_i + \varepsilon_{it},
 \end{aligned}$$

$$Y_{it} = \alpha + \beta(TREAT_i * POST_t) + \delta TREAT_i + \gamma POST_t + \eta X_{it} + \epsilon_{it}$$

Car Theft_{it} is the number of car thefts in block *i* for month *t*;

Same-Block Police_{it} is a dummy variable that equals 1 for the months after the terrorist attack (August, September, October, November, and December) if there is a protected institution in the block, 0 otherwise;

One-Block Police_{it} is a dummy variable that equals 1 after the terrorist attack (August, September, October, November, and December) if the block is one block away from the nearest protected institution, 0 otherwise;

Two-Blocks Police_{it} is a dummy variable that equals 1 after the terrorist attack (August, September, October, November, and December) if the block is two blocks away from the nearest protected institution, 0 otherwise;

M_t is a month fixed effect;

F_i is a block fixed effect;

ε_{it} is the error term.¹⁹

DID: The results

TABLE 3—THE EFFECT OF POLICE PRESENCE ON CAR THEFT

| | Difference-in-difference | | | Cross section | Time series |
|--------------------------|--------------------------|------------------------|------------------------|------------------------|------------------------|
| | (A) | (B) | (C) | (D) | (E) |
| <i>Same-Block Police</i> | −0.07752*** (0.022) | −0.08007*** (0.022) | −0.08080*** (0.022) | −0.07271*** (0.011) | −0.05843*** (0.022) |
| <i>One-Block Police</i> | | −0.01325 (0.013) | −0.01398 (0.014) | −0.01158 (0.010) | −0.00004 (0.013) |
| <i>Two-Blocks Police</i> | | | −0.00218 (0.012) | −0.00342 (0.009) | 0.01701 (0.010) |
| Block fixed effect | Yes | Yes | Yes | No | Yes |
| Month fixed effect | Yes | Yes | Yes | Yes | No |
| Number of observations | 7,884 | 7,884 | 7,884 | 4,380 | 3,816 |
| R^2 | 0.1983 | 0.1984 | 0.1984 | 0.0036 | 0.1891 |

Notes: Dependent variable: number of car thefts per month per block. Least-squares dummy variables (LSDV) regressions. Car thefts that occurred between July 18 and July 31 are excluded. Column (D) excludes observations for the preattack period (April through July). Column (E) excludes observations for the blocks that are more than two blocks away from the nearest protected institution. Huber-White standard errors are in parentheses.

*** Significant at the 1-percent level.

DID or IV?

| Variable | Feature | Block 1 (Jewish) | Block 2 (non Jewish) | Block 3 (non Jewish) |
|---|-------------|---------------------|----------------------------|----------------------------|
| Existing police force (number of officers) | Endogeneous | ? | ? | ? |
| New police assigned after the attack? | Exogeneous | 1 | 0 | 0 |
| Newly assigned police force (number of officers) | Exogeneous | ? | ? | ? |

- In this example, the exogeneous feature of the police force assignment could also be used as an instrumental variable to tackle the endogeneity challenge of the existing police force on crime
 - If the police force assigned after the attack is highly correlated with the existing police force
- Why didn't the authors use the IV approach?
 - They can only observe whether or not new police forces were introduced to each block
 - They don't have data on the distribution of police forces (confidential)

DID: To learn more

- **To learn more about the method:**

- Lechner, M. (2011). *The estimation of causal effects by difference-in-difference methods*. Now.
- Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *The Quarterly journal of economics*, 119(1), 249-275. **(On the problem of serial correlation of error terms)**
- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1), 1-19. **(Provides a solution when pre-test periods do not share a common trend)**
- [This blog](#) provides nice discussions of some advanced problems of DiD and best practices

- **To see other examples:**

- Card, D., & Krueger, A. B. (2000). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply. *American Economic Review*, 90(5), 1397-1420.
- Duflo, E. (2001). Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review*, 91(4), 795-813.
- Hanushek, E. A., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal*, 116(510), C63-C76.
- Jensen, R. (2007). The digital divide: Information (technology), market performance, and welfare in the South Indian fisheries sector. *The Quarterly Journal of Economics*, 122(3), 879-924.
- Machin, S., & Marie, O. (2011). Crime and police resources: The street crime initiative. *Journal of the European Economic Association*, 9(4), 678-701.
- Schmidt, P. (2020). The effect of car sharing on car sales. *International Journal of Industrial Organization*, 102622.

RD: Intuition

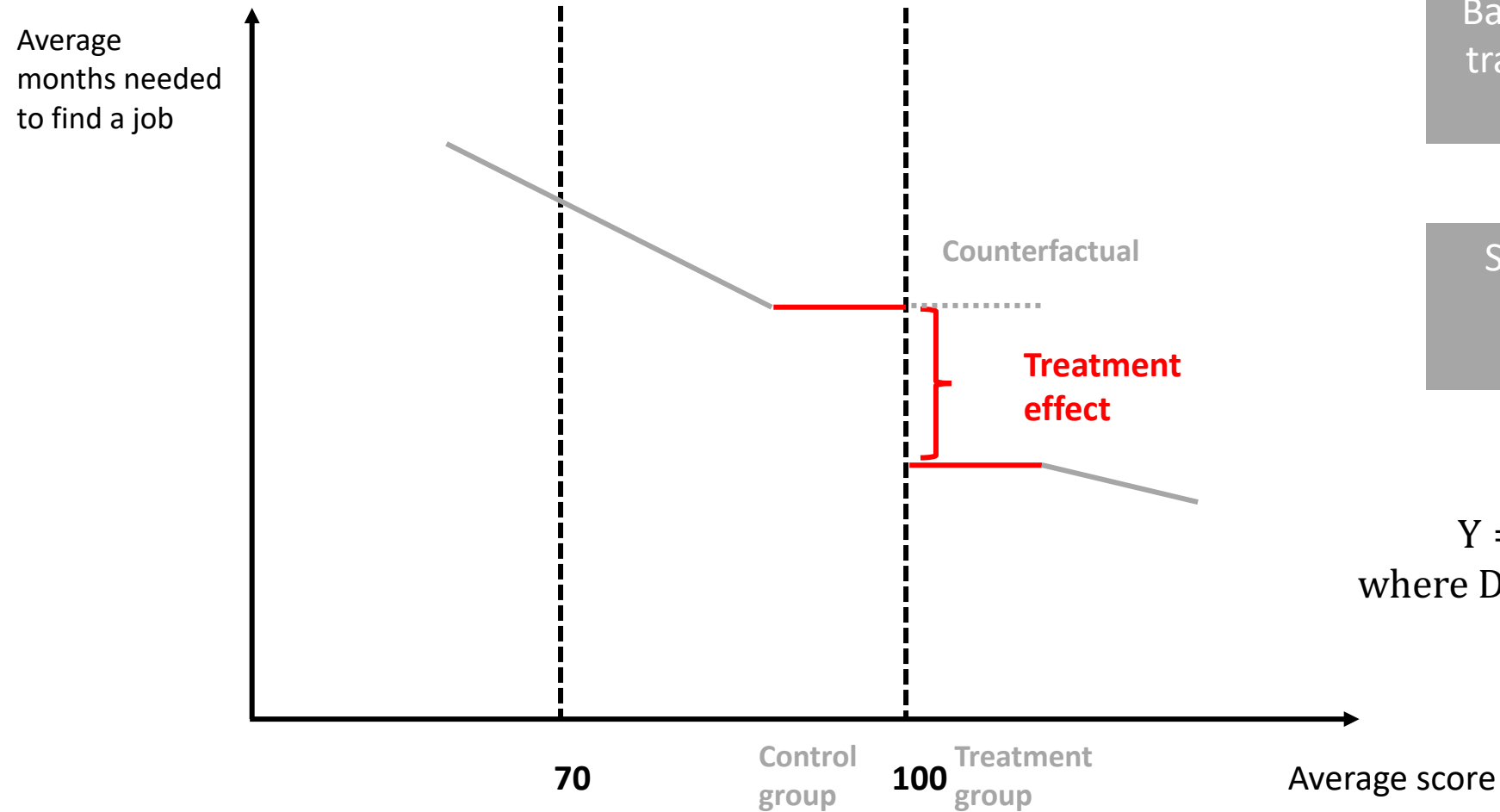
- The key step of generating a quasi-experiment situation using observational data is to find comparable control and treatment groups
- What do the following policies have in common?
 - To apply for University A, the minimum TOEFL score is 100
 - Children above 6 must wear masks at school
 - The final bill is rounded. For example, if your calculated bill is 9.51€, you will pay 10€; if it is 9.49€, you will pay 9€
- Do you see the opportunities of generating comparable treatment and control groups?
 - **Observations around the threshold are similar to each other**, but the only difference is whether they receive the treatment or not

RD: Intuition

$$\text{University application} = \begin{cases} 1 \text{ (accepted) if TOEFL score} \geq 100 \\ 0 \text{ (rejected) if TOEFL score} < 100 \end{cases}$$

- The test score is continuous, but the treatment is **discontinuous**
- Treatment status is a **deterministic function** of average score (once we know the average score, we know the treatment, sharp RD)

RD: Intuition



Bandwidth choice:
trade off between
noise and bias

Subjects cannot
manipulate
treatment

$$Y = \alpha + \beta X + \rho D + \varepsilon,$$

where D is the treatment dummy



RD: Paper example

- Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, 122(563), 957-989.

RD: Paper example

- Uncertainty of the quality of a product before purchasing
 - Expert opinion
 - Social learning (peers and family)
 - Online large-scale consumer evaluations (digital “word-of-mouth”)
- Challenge: causal impact of online review on sales
 - Endogeneity: products that receive higher notes are of better quality
 - RD design: Similar product, different notes => sales

RD: Paper example

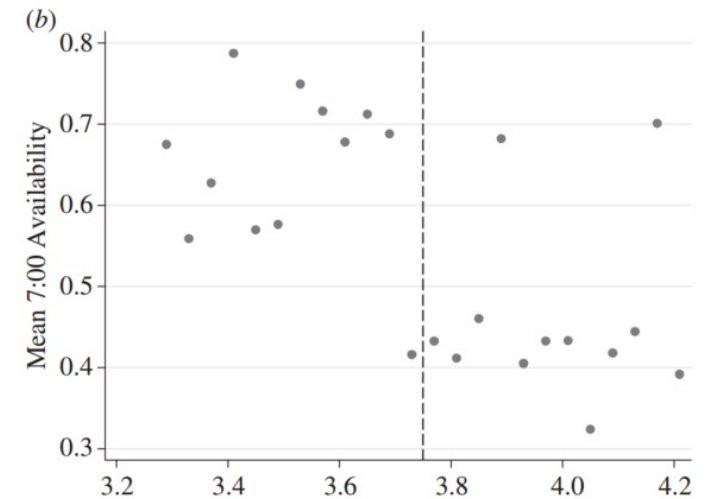
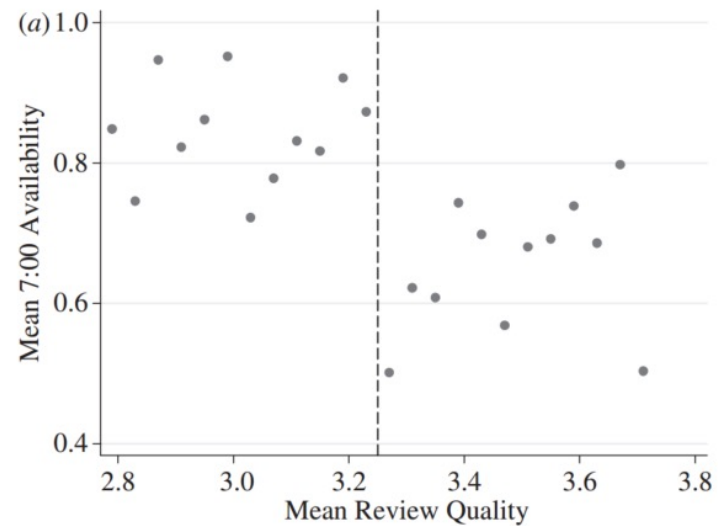
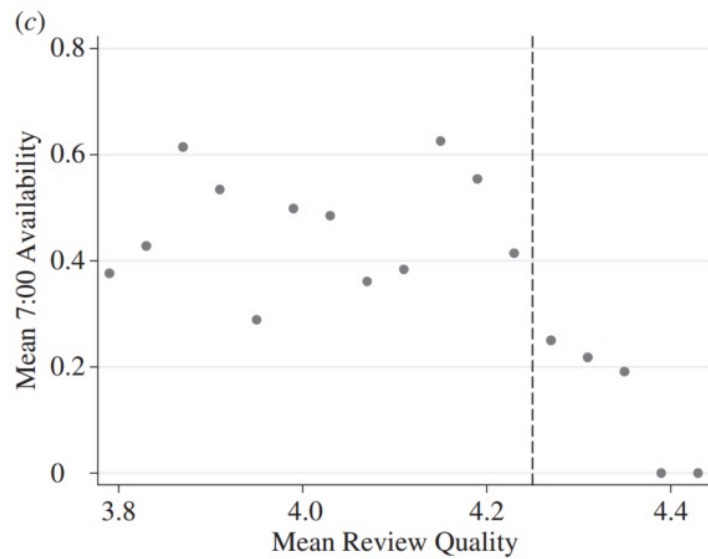
- Yelp.com: restaurant rating
- From 1 to 5 stars
- Take the average of all ratings received by the business
- Round off to the nearest half-star
 - 3.24 \rightarrow 3; 3.26 \rightarrow 3.5
- Would the discontinuity design be invalid?
 - Yes, if the restaurants can **manipulate the average ratings** to select which side they want to be



RD: Paper example

- Sans Francisco (high Yelp usage)
- Yelp for the review data:
 - Assigned stars
 - All the historical reviews
 - Number of reviews
 - Average rating at different points in time
- A large online reservation website for reservation data (sales)

RD: Paper example



RD: The regression

$$y_{it} = \alpha + \beta \times DR_{it} + \gamma f(R_{it}) + \varepsilon_{it}, \quad (1)$$

where y_{it} is an indicator for the availability of a reservation for a party of four at a particular hour in restaurant i on date t , DR_{it} is the rating that Yelp displays next to the restaurant's name and R_{it} is the actual average rating of reviews left for that restaurant.

RD: The results

Table 2
Regression Discontinuity Results at Individual Thresholds

| Yelp display rating | 6:00 PM availability | | | 7:00 PM availability | | | 8:00 PM availability | | |
|----------------------------|----------------------|-------------------|---------------------|----------------------|---------------------|----------------------|----------------------|-------------------|-------------------|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 3.5 Yelp stars | −0.079 (0.086) | | | −0.213** (0.096) | | | −0.150* (0.080) | | |
| 4 Yelp stars | | −0.101 (0.075) | | | −0.192** (0.093) | | | −0.095 (0.086) | |
| 4.5 Yelp stars | | | 0.004 (0.185) | | | −0.113 (0.127) | | | −0.119 (0.149) |
| Yelp rating | −0.228 (0.201) | 0.145 (0.203) | −0.131 (0.230) | 0.082 (0.216) | 0.024 (0.255) | −0.022 (0.271) | 0.088 (0.180) | 0.008 (0.218) | −0.321 (0.276) |
| Yelp rating × Yelp star | 0.372 (0.287) | −0.275 (0.309) | −2.934** (1.342) | −0.057 (0.335) | −0.048 (0.375) | −1.817*** (0.674) | −0.080 (0.282) | −0.329 (0.352) | −1.324 (0.869) |
| Observations | 8,705 | 11,858 | 5,597 | 8,705 | 11,858 | 5,597 | 8,705 | 11,858 | 5,597 |

Notes. Contains RD estimates of the effects of an additional Yelp half-star on availability. Availability measures indicate whether the reservations were available at that time on Thursday, Friday or Saturday when queried 36 hours in advance. Standard errors are clustered at the restaurant level. Asterisks denote significance levels at: *10%, **5% and ***1%.

RD: To learn more

- **To learn more about the method:**

- Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2), 615-635.
- Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2), 281-355.
- Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3), 933-959.

- **To see other examples:**

- Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of economics and statistics*, 86(1), 226-244.
- Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, 99(1), 179-215.
- Montizaan, R., Cörvers, F., & De Grip, A. (2010). The effects of pension rights and retirement age on training participation: Evidence from a natural experiment. *Labour Economics*, 17(1), 240-247.
- Battistin, E., Brugiavini, A., Rettore, E., & Weber, G. (2009). The retirement consumption puzzle: evidence from a regression discontinuity approach. *American Economic Review*, 99(5), 2209-26.
- Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). *Using big data to estimate consumer surplus: The case of uber* (No. w22627). National Bureau of Economic Research.



IV: Paper example

- Czernich, N., Falck, O., Kretschmer, T., & Woessmann, L. (2011). Broadband infrastructure and economic growth. *The Economic Journal*, 121(552), 505-532.

Paper example – Broadband and economic growth

- Can broadband infrastructure **lead to** (causal relationship) economic growth, and how big is the effect (coefficient estimation)?
- Why should we care about this question?
 - Evaluate the impact of existing installations
 - Argument for installing new broadband infrastructure
- Two challenges:
 - How to measure?
 - Broadband infrastructure: broadband penetration rate (the share of the population that has subscribed to broadband in country i)
 - Economic growth: growth of GDP per capita
 - Endogeneity

Paper example – Broadband and economic growth

- Sample
 - OECD countries
 - 1996-2007
- Choice of other independent variables: Follow the previous literature:
 - Endogenous growth theory in macro economics(Lucas, 1988; Romer, 1990)
 - Physical capital (capital formation/GDP)
 - Human capital (average number of years of schooling of 15-64)
 - State of the technology (a function of the broadband infrastructure)

Paper example – Broadband and economic growth - Endogeneity

- **Reverse causality**
 - Individuals in high-income countries may also have a high ability to pay for broadband services, resulting in more rapid broadband penetration
- **Omitted variables** (which both correlates with broadband infrastructure and with economic growth)
 - State intervention in telecommunications might depend on the level of economic development in a country, thus confounding the effect of regulation and sectoral policies with the effect of broadband diffusion
 - Given the rapid technological change in recent decades, broadband diffusion took place at the same time as the diffusion of other technologies such as mobile telephony and computers, making it difficult to isolate the specific effect of broadband

Paper example – Broadband and economic growth – OLS result

Table 3
Broadband Diffusion and Growth of Gross Domestic Product (GDP) per Capita

| Dependent variable: growth of GDP per capita | Model 1 | Model 2 | Model 3 |
|--|---------------------|---------------------|---------------------|
| Broadband penetration rate | 0.065*** (3.08) | 0.091*** (4.31) | 0.083*** (3.03) |
| Years since broadband introduction | −0.003*** (2.95) | −0.004*** (3.89) | −0.003** (2.59) |
| Growth of capital formation/GDP | 0.069*** (6.76) | | |
| Growth of years of education | −0.007 (0.30) | | |
| ΔGrowth of working-age population | −0.231 (1.06) | | |
| GDP per capita in 1996 | −0.001*** (3.90) | −0.001*** (3.97) | −0.001*** (3.68) |
| Constant | 0.049*** (6.98) | 0.053*** (7.68) | 0.046*** (8.60) |
| R ² | 0.30 | 0.19 | 0.11 |
| Observations | 240 | 240 | 300 |
| Countries | 20 | 20 | 25 |

Notes. Ordinary least squares estimation for 1996–2007. Sample of OECD countries. Five countries drop from the full model because of lack of data on control variables. Broadband penetration rate measured as broadband subscribers per 100 inhabitants; broadband line refers download speeds of at least 256 kbit/s. Robust t statistics in parentheses. Significance at * 10%, ** 5% and *** 1% levels.

Flash back: 2SLS estimation

Given the linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad (1)$$

where X_1 is an endogenous variable

First stage of 2SLS: Regress the endogenous variable on the instrument and other explanatory variables and obtain the estimated value of the endogenous variable

$$\hat{X}_1 = \gamma_0 + \gamma_1 Z_1 + \gamma_2 X_2 + \gamma_3 X_3 + v \quad (2)$$

where Z_1 is the instrumental variable. We call the above equation the **”reduced form”** of X_1

γ_1 needs to be significant (why?)

\hat{X}_1 is no longer related to ϵ (why?)

Second stage of 2SLS: Plug in the fitted values of \hat{X}_1 derived from equation (2) into the original linear regression equation (1):

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad (3)$$

now ϵ is no longer correlated with \hat{X}_1 , and β_1 is an unbiased estimator

Paper example – Broadband and economic growth – IV & 2SLS

- Two instrumental variables
 - Voice telephony penetration rate
 - Cable TV penetration rate
- Why?
 - “Most commonly used broadband standards (e.g. ADSL, VDSL) rely on the copper wire of the voice telephony network or the coaxial cable of the cable TV network between households and the main distribution frame or the street cabinet.”
- Needs to be strongly correlated with the endogenous variable

Table 4
The Diffusion Curve: First Stage of the Instrumental Variable Model

| Dependent variable: broadband penetration rate | Model 1 | Model 2 | Model 3 |
|---|-----------------------------|-----------------------------|----------------------------|
| Voice telephony penetration rate | 0.585*** (6.16) | 0.487*** (4.59) | 0.653*** (7.24) |
| Cable TV penetration rate | 0.279** (2.11) | 0.301** (2.54) | 0.334** (2.47) |
| Diffusion speed (β) | 0.647*** (15.60) | 0.623*** (13.36) | 0.613*** (9.15) |
| Inflexion point (τ) | 2,004.531*** (10,423.78) | 2,004.532*** (10,508.86) | 2,004.485*** (7,939.33) |
| Log of capital formation/GDP | | | 0.035** (2.24) |
| Log of years of education | | | 0.034* (2.08) |
| Growth of working-age population | | | 0.347 (0.86) |
| Constant (γ_0) | -0.057 (1.25) | 0.002 (0.04) | -0.104** (2.25) |
| R ² | 0.96 | 0.93 | 0.95 |
| F-test (cable TV penetration rate = voice telephony penetration rate = 0) | 26.69 | 32.18 | 31.92 |
| Observations | 260 | 325 | 240 |
| Countries | 20 | 25 | 20 |

Notes. Non-linear least squares estimation for 1996–2008. Diffusion speed and inflexion point do not vary across countries. In contrast, the saturation level is country-specific and is a linear function of the voice telephony penetration rate and the cable TV penetration rate in the year before the first emergence of broadband. Model 3 contains additive control variables and is for 1996–2007. Sample of OECD countries. Five countries drop from the full model because of lack of data on control variables. Robust t-statistics in parentheses. Significance at * 10%, ** 5% and *** 1% levels. GDP, gross domestic product.

1st stage of 2SLS

Paper example – Broadband and economic growth – IV & 2SLS

| Dependent variable: growth of GDP per capita | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|--|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| Predicted broadband penetration rate | 0.092*** (3.57) | 0.135*** (5.27) | 0.145*** (4.10) | | | |
| Lagged predicted broadband penetration rate | | | | 0.116*** (4.05) | 0.156*** (5.20) | 0.159*** (4.49) |
| Years since predicted broadband introduction | -0.003*** (3.92) | -0.005*** (5.40) | -0.005*** (4.03) | -0.004*** (4.66) | -0.005*** (5.53) | -0.004*** (4.24) |
| Growth of capital formation/GDP | 0.069*** (5.73) | | | 0.071*** (5.28) | | |
| Growth of years of education | -0.004 (0.19) | | | -0.018 (1.18) | | |
| ΔGrowth of working-age population | -0.18 (0.77) | | | -0.227 (0.98) | | |
| GDP per capita in 1996 | -0.001*** (6.06) | -0.001*** (6.57) | -0.001*** (4.38) | -0.001*** (7.01) | -0.001*** (7.24) | -0.001*** (4.18) |
| Constant | 0.052*** (10.29) | 0.056*** (11.59) | 0.049*** (8.36) | 0.057*** (11.72) | 0.060*** (12.33) | 0.050*** (7.74) |
| R ² | 0.29 | 0.19 | 0.12 | 0.35 | 0.25 | 0.14 |
| Observations | 240 | 240 | 300 | 220 | 220 | 275 |
| Countries | 20 | 20 | 25 | 20 | 20 | 25 |

Notes. Second-stage estimation for 1996–2007. Broadband penetration rates and year of broadband introduction are predicted from the first-stage diffusion curves reported in Table 4. Sample of OECD countries. Five countries drop from the full model because of lack of data on control variables. Bootstrapped standard errors in parentheses. Significance at * 10%, ** 5% and *** 1% levels.

2nd stage of 2SLS

Paper example – Broadband and economic growth – Defend your IV

- Test for the **exclusion restriction**: An instrument impacts the outcome only through the instrumented variable
 - Telephone and cable network penetration ~~+~~ mobile phones and computers diffusion → economic growth
 - Open economy → High telephone and cable network penetration...
 - & ... Open economy → Economic growth
 - Controlled for trade openness, model result does not change
 - Also controlled for: average working population education level, average household size

Paper example – Broadband and economic growth – Defend your IV

Table 10
Additional Robustness Specifications

| Dependent variable: growth of GDP per capita | Model 1 | Model 2† | Model 3 | Model 4 |
|--|---------------------|---------------------|---------------------|---------------------|
| Predicted broadband penetration rate | 0.090*** (3.26) | 0.104*** (3.57) | 0.084*** (3.29) | 0.097*** (3.64) |
| Years since predicted broadband introduction | -0.003*** (3.75) | -0.004*** (3.78) | -0.003*** (3.99) | -0.004*** (3.98) |
| Voice telephony penetration rate | -0.009 (0.95) | | | |
| Cable TV penetration rate | -0.003 (0.37) | | | |
| Trade openness | | | 0.007*** (2.83) | |
| Log of years of education | | | | -0.009 (1.22) |
| Growth of capital formation/GDP | 0.069*** (5.47) | 0.070*** (5.07) | 0.073*** (6.14) | 0.069*** (5.54) |
| Growth of years of education | -0.002 (0.09) | -0.001 (0.03) | -0.008 (0.35) | 0.0004 (0.02) |
| ΔGrowth of working-age population | -0.174 (0.75) | -0.196 (0.78) | -0.195 (0.87) | -0.179 (0.76) |
| GDP per capita in 1996 | -0.001*** (4.94) | -0.001*** (5.44) | -0.001*** (4.59) | -0.001*** (5.84) |
| Constant | 0.054*** (8.80) | 0.053*** (9.58) | 0.042*** (6.96) | 0.067*** (4.56) |
| R ² | 0.23 | 0.29 | 0.32 | 0.30 |
| Observations | 240 | 216 | 240 | 240 |
| Countries | 20 | 18 | 20 | 20 |

Notes. Second-stage estimation for 1996–2007. Broadband penetration rates and year of broadband introduction are predicted from the first-stage diffusion curve reported in Table 4. Sample of OECD countries. Five countries drop from the full model because of lack of data on control variables. Bootstrapped z-statistics in parentheses. Significance at * 10%, ** 5% and *** 1% levels. †Model 2 excludes Canada and the US. GDP, gross domestic product.

Paper example – Broadband and economic growth – Defend your IV

Table 11
Measuring Penetration per Household

| | Broadband penetration per capita | | Broadband penetration per household | | Broadband penetration per household | |
|---|----------------------------------|-----------------------------|-------------------------------------|----------------------------|-------------------------------------|----------------------------|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
| <i>First stage; dependent variable</i> | | | | | | |
| Voice telephony penetration per capita | 0.470*** (3.09) | 0.643*** (4.04) | | | 1.281*** (4.96) | 0.882** (2.09) |
| Cable TV penetration per capita | 0.251* (2.02) | 0.371** (2.48) | | | 0.637* (1.75) | 0.724** (32.08) |
| Voice telephony penetration per household | | | 0.551*** (3.47) | 0.685*** (3.89) | | |
| Cable TV penetration rate per household | | | 0.357** (2.40) | 0.388*** (3.02) | | |
| Average household size | -0.139 (1.50) | 0.099 (0.92) | | | | |
| Diffusion speed (β) | 0.659*** (14.87) | 0.620*** (11.59) | 0.646*** (13.80) | 0.594*** (9.21) | 0.653*** (14.49) | 0.615*** (10.70) |
| Inflexion point (τ) | 2,004.391*** (10,495.73) | 2,004.511*** (10,347.74) | 2,004.442*** (8,522.16) | 2,004.266*** (7,531.25) | 2,004.430*** (9,450.50) | 2,004.406*** (9,122.98) |
| Constant (γ_0) | 0.333 (1.18) | -0.34 (1.05) | -0.187 (0.99) | -0.363 (1.66) | -0.097 (0.83) | 0.142 (0.69) |
| R ² | 0.95 | 0.92 | 0.93 | 0.90 | 0.94 | 0.86 |
| F-test (cable TV penetration rate = voice telephony penetration rate = 0) | 11.03 | 12.93 | 14.41 | 12.60 | 18.79 | 12.03 |
| <i>Second stage; dependent variable: growth of GDP per capita</i> | | | | | | |
| Predicted broadband penetration rate | 0.100*** (3.79) | 0.095*** (2.62) | 0.021* (1.89) | 0.024** (2.35) | 0.023** (2.11) | 0.024** (2.27) |
| Observations | 240 | 300 | 240 | 300 | 240 | 300 |
| Countries | 20 | 25 | 20 | 25 | 20 | 25 |

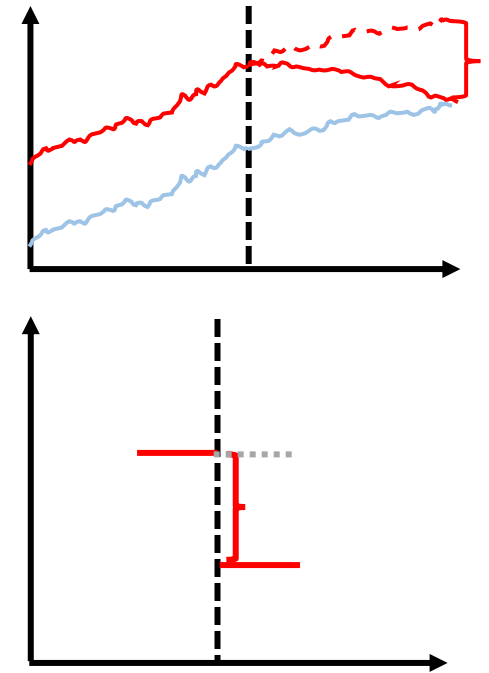
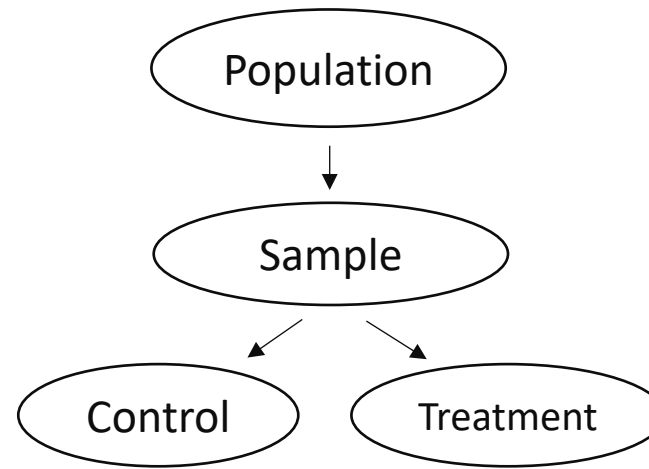
Notes. *Top panel:* non-linear least squares estimation for 1996–2007. See notes to Table 4 for additional details. Robust t statistics in parentheses. *Bottom panel:* second-stage estimation for 1996–2007. See notes to Table 6 for the additional control variables contained and for additional details. Bootstrapped z-statistics in parentheses. Significance at * 10%, ** 5% and *** 1% levels. GDP, gross domestic product.

Wrap up of the structure of a paper using IV

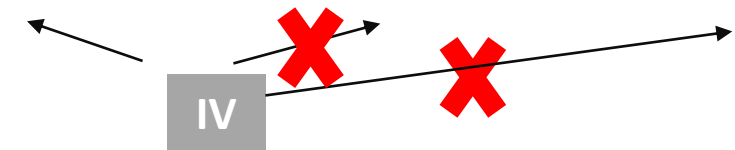
1. **Raise the research question:** we want to estimate the effect of variable A to variable B
2. **Explain the potential sources of endogeneity** of variable A
 - There may be several endogenous variables, each needs at least one IV
3. Propose one or several **instrumental variables**
4. **Report the OLS result** as a reference
5. **Report 2SLS results**
 - First stage: the IV significantly explains the endogenous variable
 - Second stage: show the true effect of variable A to variable B
6. **Robustness checks of the IV**
 - IV is not correlated with other independent variables or with the error term

A quick summary of the course

- Correlation \neq causation!
- Instrumental variable
- Differences-in-differences
- Regression discontinuity
- Econometrics vs. machine learning



Outcome = Endogeneous variable + other variables + error term



Next courses:

