Identification (3)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute

1. Introduction

- **Topics:**
  1. Spatial econometrics
     - Spatial data, autocorrelation, spatial regressions
  2. Discrete choice
     - Random utility framework, estimating binary and multinomial regression models
  3. **Identification**
     - **Research design, IV, OLS, RDD, quasi-experiments, standard errors**
  4. **Hedonic pricing**
     - **Theory and estimation**
  5. **Quantitative spatial economics**
     - **General equilibrium models in spatial economics**

1. Introduction

- *Monday:*
  1. Spatial econometrics (1+2+3)
- *Tuesday:*
  2. Discrete choice (1+2+3)
- *Wednesday:*
  3. Identification (1+2+3)
- *Thursday:*
  4. Hedonic pricing (1+2)
  5. Quantitative spatial economics (1+2)

2. Quasi-experiments

- **Possible identification strategies**
  1. Randomised experiments
  2. Exhaustive set of controls
  3. Instrumental variables (IV)
  4. **Quasi-experiments (QE)**
     - ↳ **Regression-discontinuity designs (RDD)**

2. Quasi-experiments

- Use exogenous shocks in the economy to identify causal effects
  - 'Quasi'-experiments / natural experiments

- National policy changes, (arbitrary) policy rules, earthquakes, bombings
  - These shocks cannot be influenced by the individual decision makers
  - Recall: if shock is really random, selection effect is equal to zero
  - The research context indicates whether shock is indeed random

- **Regression-discontinuity design (RDD)**
  - **Quasi-experimental method**

- **Assume that we have a treatment effect that is dependent on r_i:**

$$x_i = \begin{cases} 1 & \text{if } r_i \geq r_0 \\ 0 & \text{if } r_i < r_0 \end{cases}$$

  - **r_0 is some cutoff value**

- **This leads to a regression:**

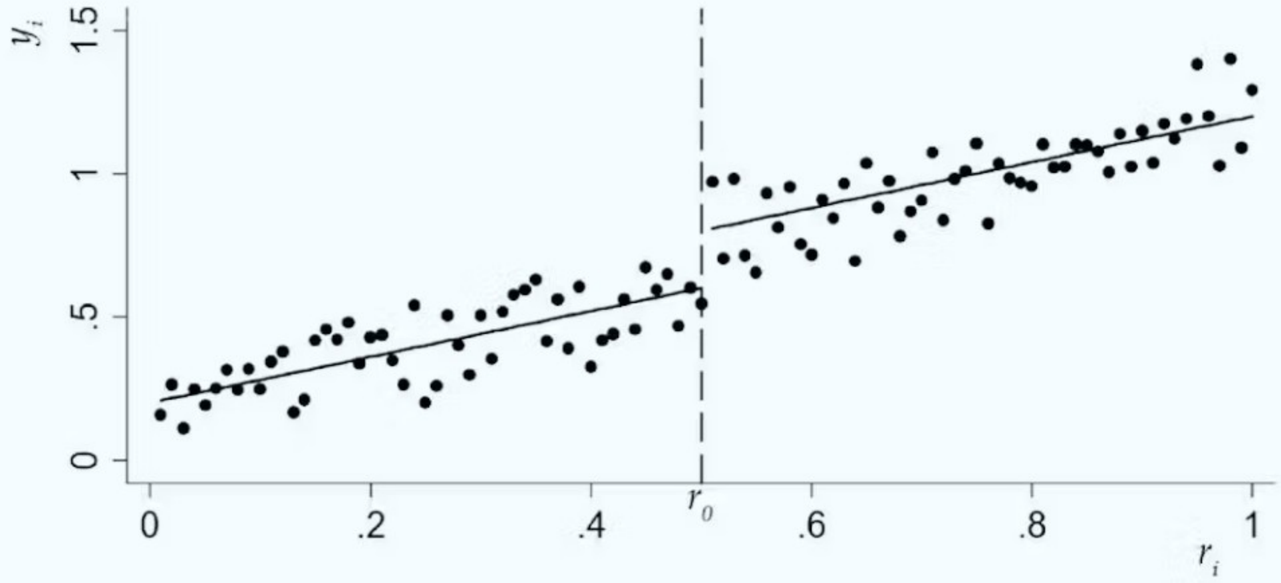
$$y_i = \alpha + \beta x_i + \gamma r_i + \epsilon_i$$

  - **Note that x_i is a fully deterministic function of r_i**
  - **Not perfectly collinear because r_i is continuous**

3. RDD

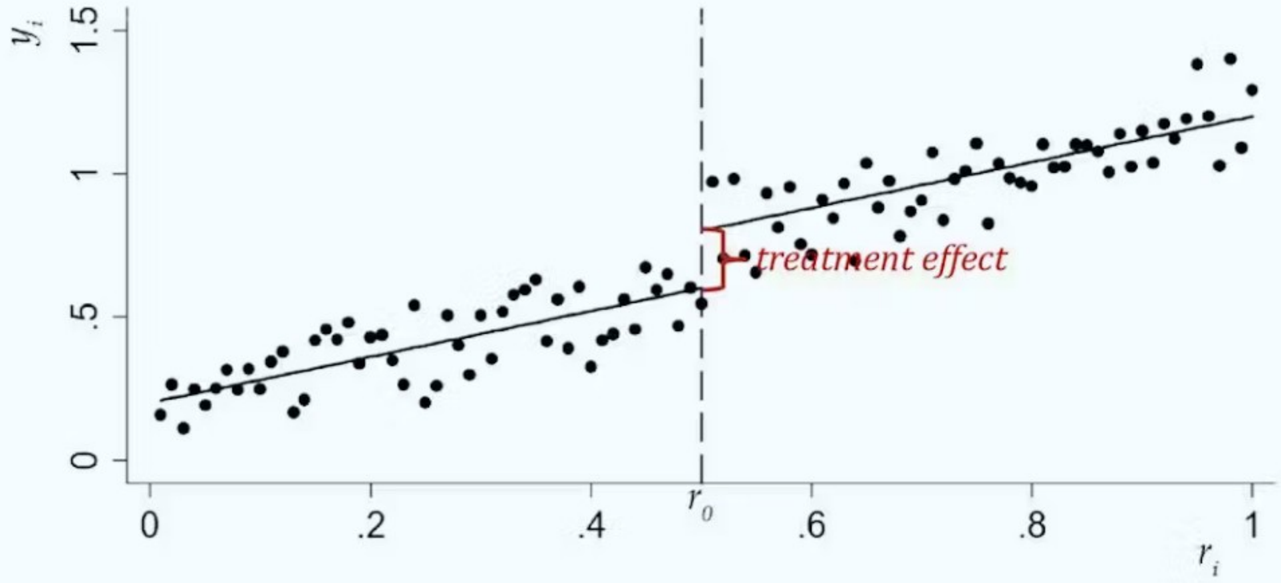
- **Example:**
  - **Students get a scholarship if they achieve a certain test-score**
  - **You aim to know the impact of the scholarship on job market outcomes**
    - » **e.g. wages**

- **Plot**



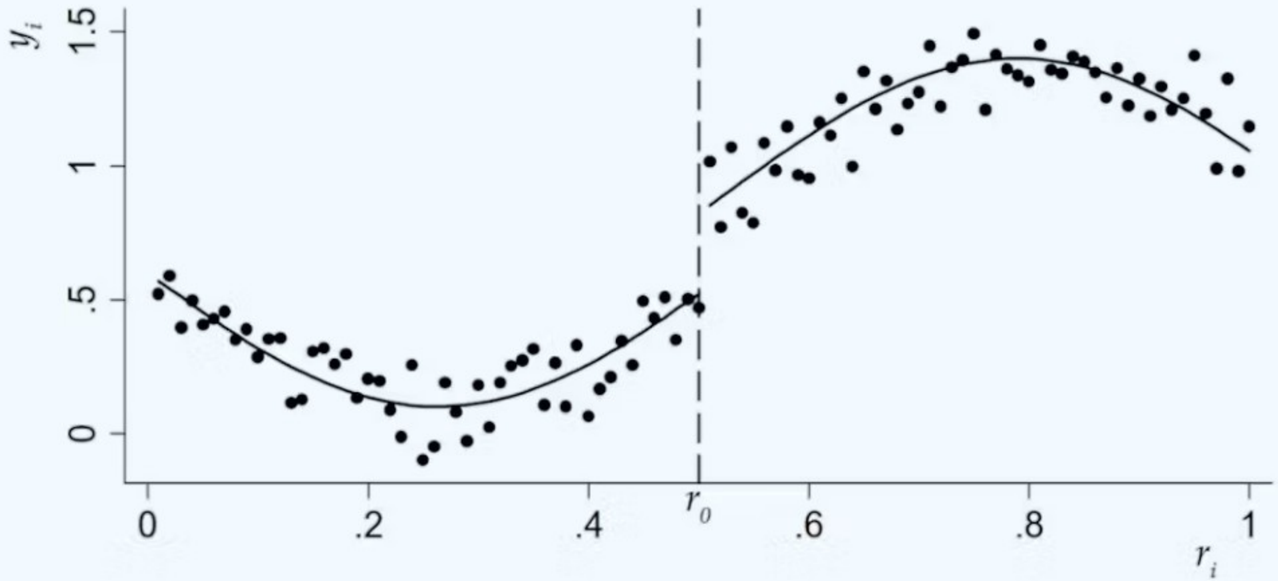
- **Control for test scores and investigate the jump in treatment at r_0**

- **Plot**

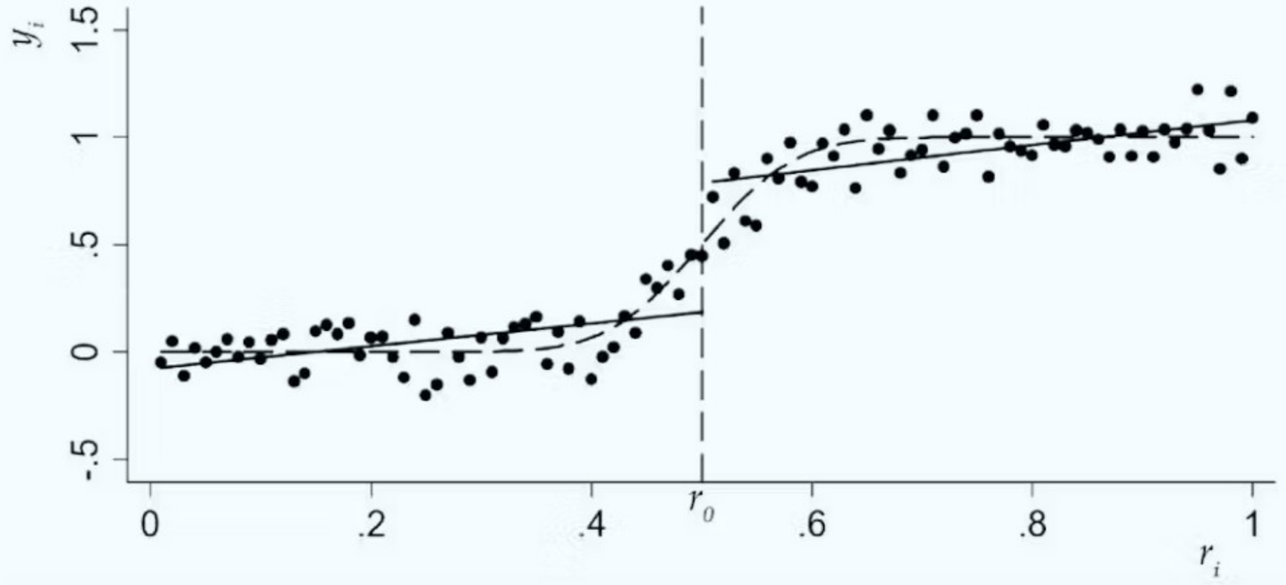


- **Control for test scores and investigate the jump in treatment at r_0**

- **What if x is non-linearly related to Y**
  - $y_i = \alpha + \beta x_i + f(r_i) + \epsilon_i$
  - **Specify $f(r_i)$:**
    - » $f(r_i) = \sum_{q=1}^{Q} \gamma_{q-} r_i^q (r_i < r_0) + \sum_{q=1}^{Q} \gamma_{q+} r_i^q (r_i \geq r_0)$
    - » **Q^{th}-order polynomial**
    - » **Can be estimated by OLS**
    - » **For example:** $f(r_i) = \gamma_1 r_i (r_i < r_0) + \gamma_2 r_i^2 (r_i < r_0) + \gamma_3 r_i (r_i \geq r_0) + \gamma_4 r_i^2 (r_i \geq r_0)$

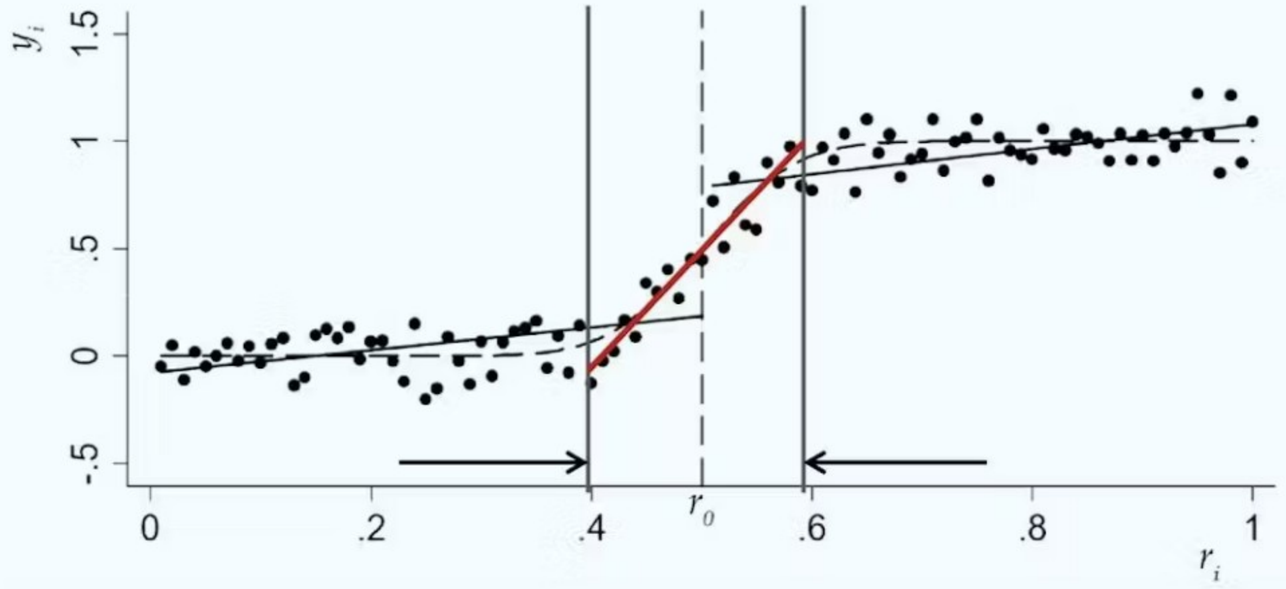


- To check for nonlinearities in a RDD is important



3. RDD

- To check for nonlinearities in a RDD is important
  - To reduce the possibility of mistakes, you may focus on observations 'close' to r_0
  - Reduces precision



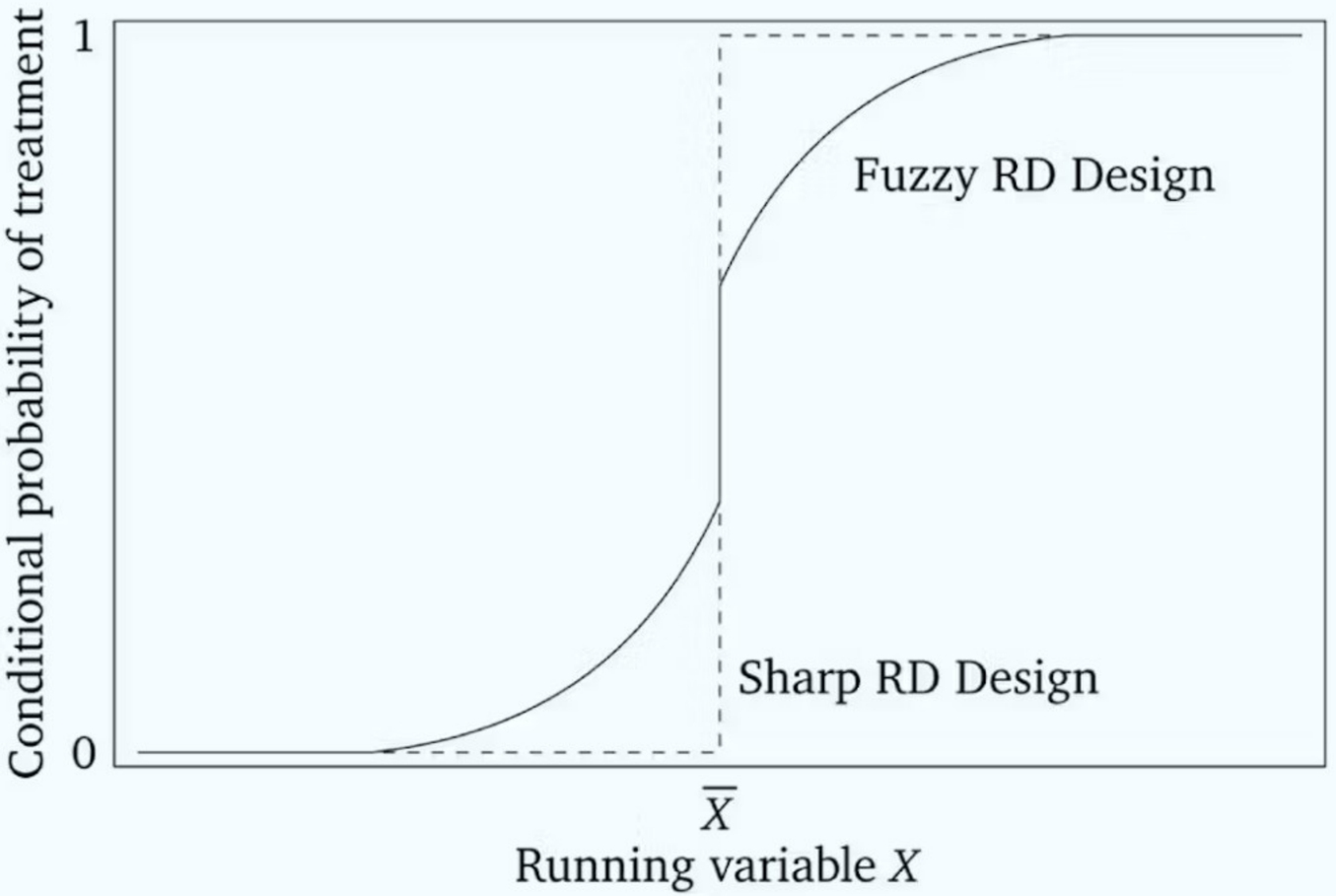
3. RDD

- Two different versions
  - *Sharp* RDD → Jump in treatment
  - *Fuzzy* RDD → Jump in probability of treatment

- Previous slides: sharp RDD

- Fuzzy RDDs are very common
  - Assignment is often 'fuzzy'

Illustration of a fuzzy RDD



- **Fuzzy RDD**
  - $\text{Prob}[x_i = 1 \mid r_i] = \begin{cases} g_1(r_i) & \text{if } r_i \geq r_0 \\ g_0(r_i) & \text{if } r_i < r_0 \end{cases}$
    where $g_1(r_i) \neq g_0(r_i)$

- $\text{Prob}[x_i = 1 \mid r_i] = g_0(r_i) + [g_1(r_i) - g_0(r_i)]z_i$
  - $z_i = \mathbb{I}(r_i \geq r_0)$

- **Looks complicated – it just means that treatment probability is discontinuous at some point**

- **This leads to a two-stage least squares estimator**
  - **First stage →** $x_i = \zeta + \eta z_i + g(r_i) + \xi_i$**, with** $z_i = \mathbb{I}(r_i \geq r_0)$
  - **Second stage →** $y_i = \alpha + \beta \hat{x}_i + f(r_i) + \epsilon_i$

- Koster and Van Ommeren (2019)

- What is the impact of urban renewal programmes on house prices?
  - € 216 million by national government
  - € 1 billion by public housing associations

- Investments mainly in restructuring of public housing stock

- **Example of targeted neighbourhoods in Amsterdam:**

- **Use first-differencing, denoted by Δ:**

$$\Delta y_{it} = \Delta\alpha + \beta\Delta x_{it} + \gamma\Delta c_{it} + \Delta\mu_t + \Delta\epsilon_{it}$$

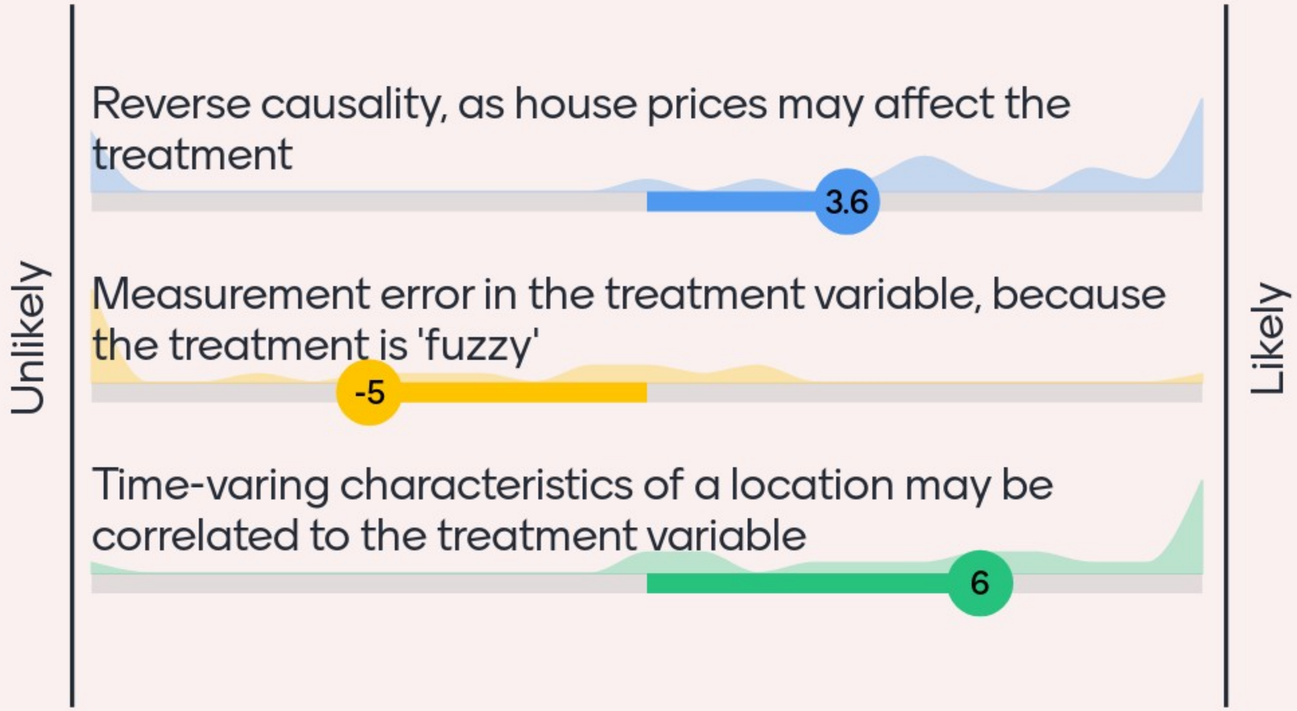
where i **property**
t **year**
y_{it} **log house price**
x_{it} **in a targeted neighbourhood**
c_{it} **control variables**
μ_t **time fixed effects**

→ **What are the benefits of using first-differences/panel data**

→ **What are potentially remaining endogeneity problems?**

What are the benefits of panel data/first-differencing in this setting?

What are potentially remaining endogeneity concerns?



3. RDD

- Endogeneity issue ➔ price trends of treated neighbourhoods may be different from other neighbourhoods
  - *e.g.* gentrification, trends in social interactions

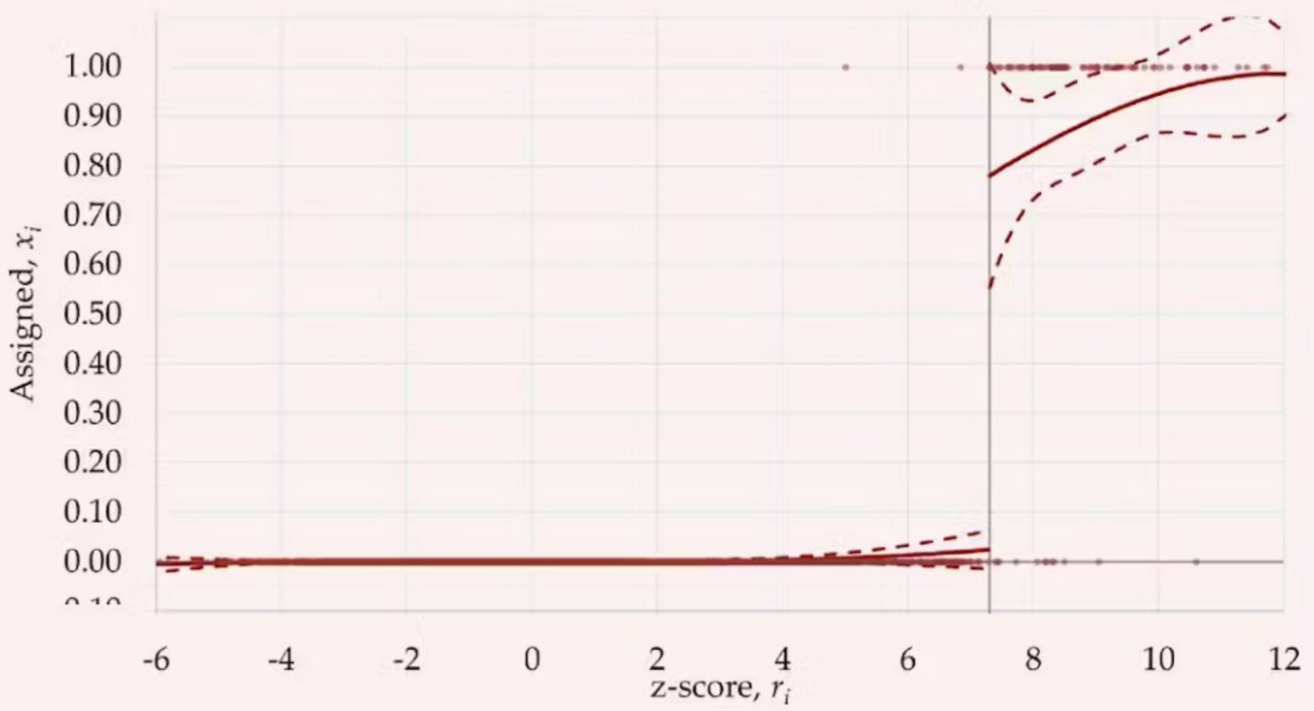
- Solution: use RDD

- **The neighbourhoods that are eligible were selected based on deprivation *z*-scores**

TABLE 1 — DEPRIVATION SCORES FOR NEIGHBOURHOODS

	All neighbourhoods		KW neighbourhoods	
	μ	σ	μ	σ
Social deprivation	0.000	0.654	1.167	0.322
Physical deprivation	0.000	0.611	2.070	0.660
Social problems	0.000	0.924	2.612	1.053
Physical problems	0.000	0.950	3.087	0.976
Overall	0.000	2.414	8.935	1.340
Number of neighbourhoods	4016		83	

- **Assignment of neighbourhoods based on z-scores**



→ **Is this a fuzzy or a sharp RDD?**

Is this a sharp or a fuzzy regression-discontinuity design?

5

Sharp RDD

✗

14

Fuzzy RDD

✓

0

This is not an RDD

✗

- **Only select observations close to the threshold** $z_{\ell t} = 7.3$

- **+'IV'-strategy**
  - $\Delta x_{it} = \zeta + \eta \Delta z_{it} + \theta \Delta c_{it} + \Delta \nu_t + \Delta \xi_{it}$ 1^{st} *stage*
    where $z_{it} = \mathbb{I}(r_i \geq r_0)$
    $z_{it} = 0$ **before the programme started**
  - **Use fitted value of** Δx_{it} **in second stage**

- **Results**

Table 4.4 – URBAN RENEWAL AND HOUSE PRICES

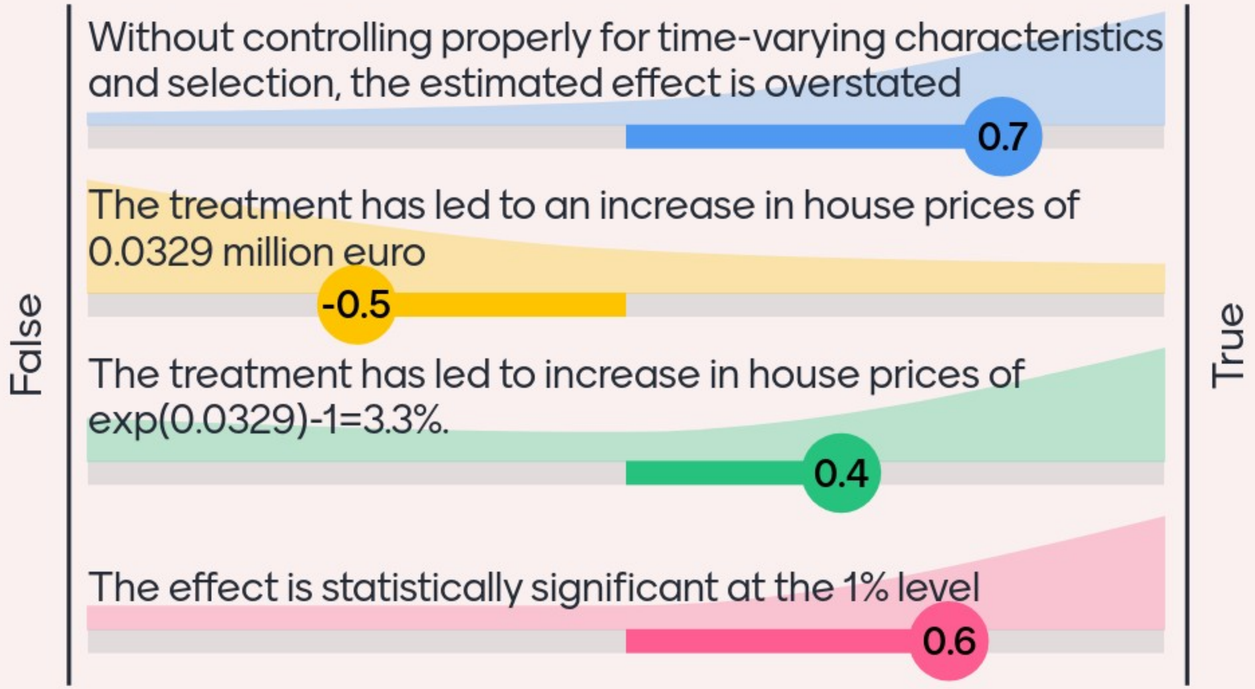
(Dependent variable: the change in the log of house prices)

	First-differences	+Fuzzy RDD
	(1)	(2)
ΔKW investment	0.0441***	0.0329***
	(0.0114)	(0.0122)
Number of observations	169,664	22,589
R^2-within	0.375	
Kleibergen-Paap F-statistic		5444
Bandwidth, δ		3.383

Notes: We exclude observations within 2.5km of targeted neighbourhoods to avoid picking up spillover effects beyond the neighbourhood boundaries. In column (3) the change in the KW investment is instrumented with the change in the eligibility based on the scoring rule. Standard errors are clustered at the neighbourhood level and in parentheses; *** $p < 0.01$, ** $p < 0.5$, * $p < 0.10$.

→ **Please interpret the coefficients**

What statements regarding the results are true?



3. RDD

- Urban renewal programmes have led to changes in prices of about 3-3.5%
  - Neighbourhoods have become more attractive
  - In that respect, the programme was effective
    - » Total house price increase is higher than the investments

- More information on this study in Urban Economic Challenges & Policies

4. Standard errors

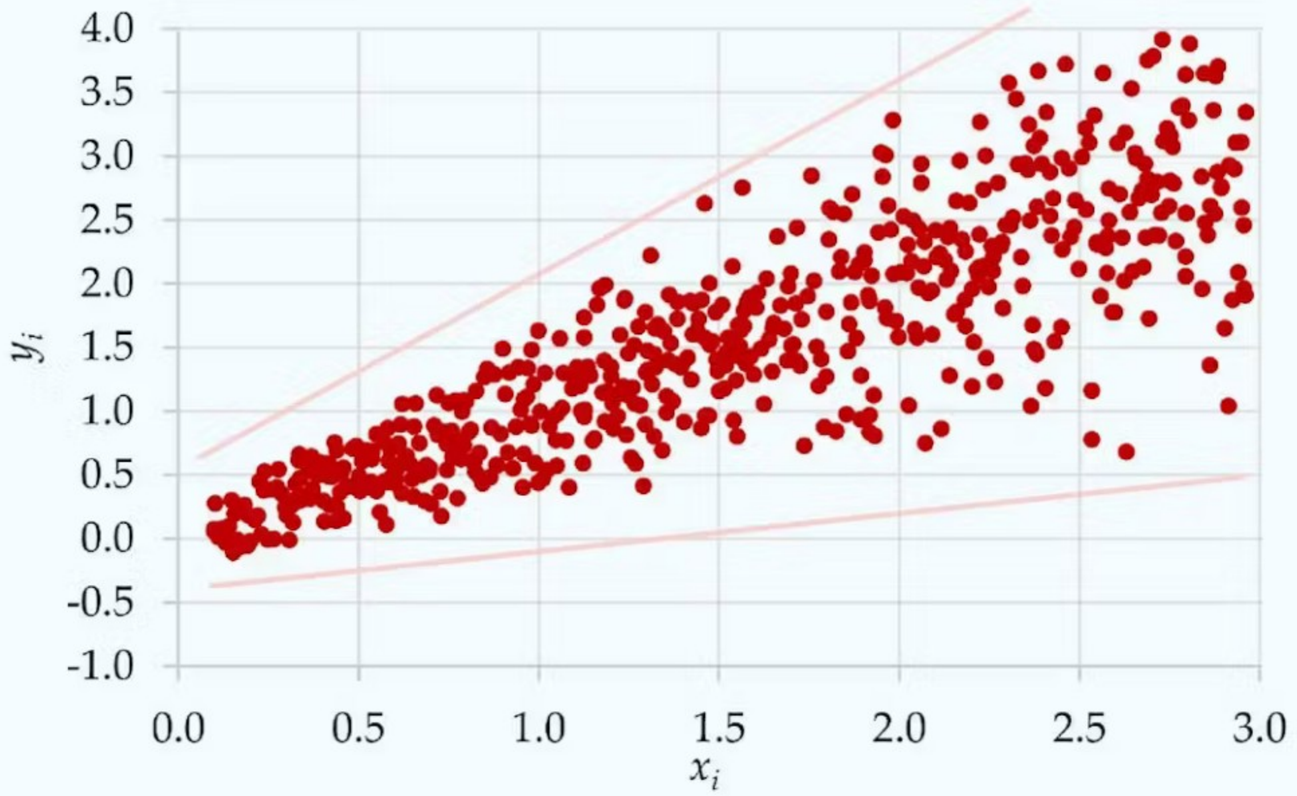
- This course has almost entirely focused on estimating average effects

- But how to assess statistical significance?
  - Use correctly estimated standard errors
  - May be very important!

- Some issues with standard errors
  1. Heteroscedasticity
  2. Clustering
  3. Serial correlation

1. Heteroscedasticity

- Conditional variance of y_i given x_i changes with i



In which of the figures below, residuals are homoscedastic?

A

✓

B

✗

1. Heteroscedasticity

- To estimate standard errors, we typically assume homoscedasticity

- Solution: use robust standard errors
  - In STATA, type *r* after REGRESS
  - This leads to consistent s.e.

- However, robust standard errors are biased
  - Only a problem in small samples

2. Clustering

- Issue $y_{ig} = \alpha + \beta x_g + \epsilon_{ig}$
  - You basically multiply the size of the dataset leading to artificially low standard errors
  - *The effective number of observation is much lower*
  - Can make a big difference!

- More generally, to obtain consistent standard errors, you assume that $\mathrm{E}[\epsilon_{ig}\epsilon_{jg}] = 0$
  - This is certainly not the case in the above example

2. Clustering

- This is the formula for the standard error when $\mathbb{E}[\epsilon_{ig}\epsilon_{jg}] = 0$:

$$SE(\hat{\beta}) = \frac{\sigma_\epsilon}{\sqrt{N}}\frac{1}{\sigma_x}$$

- Let's for simplicity assume that everyone in the municipality has the same well-being
  - Then the correct standard error is:

$$SE(\hat{\beta}) = \frac{\sigma_\epsilon}{\sqrt{G}}\frac{1}{\sigma_x}$$

- Say that you have 9,000 individuals but only 90 municipalities
  - Standard error is 10 times larger

2. Clustering

- Solution: cluster your standard errors at the appropriate level

- Not always clear at what level you should cluster
  - ...when different variables are aggregated at different levels
  - Pragmatic approach: choose standard errors that lead to the most conservative conclusions (→ highest standard errors)
  - Use multi-way clustering
    → In REGHDFE command in Stata

- Note: clustered standard errors are not correct for a few number of clusters

3. Serial correlation

- **Time series specifications:**
  - $\Delta y_{it} = \alpha_t + \rho \Delta x_{it} + \Delta X'_{it}\gamma + \Delta \eta_i$

- **Same problem as before:** $\mathbb{E}[\epsilon_{it}\epsilon_{it-1}] \neq 0$

- **Solution: cluster at individual level i?**
  - **But:** $\mathbb{E}[\epsilon_{it}\epsilon_{it-1}] \neq \mathbb{E}[\epsilon_{it}\epsilon_{it-2}]$

- **This issue is still under study!**
  - **Two-way clustering may be a solution**
  - **Use REGHDFE command in Stata**

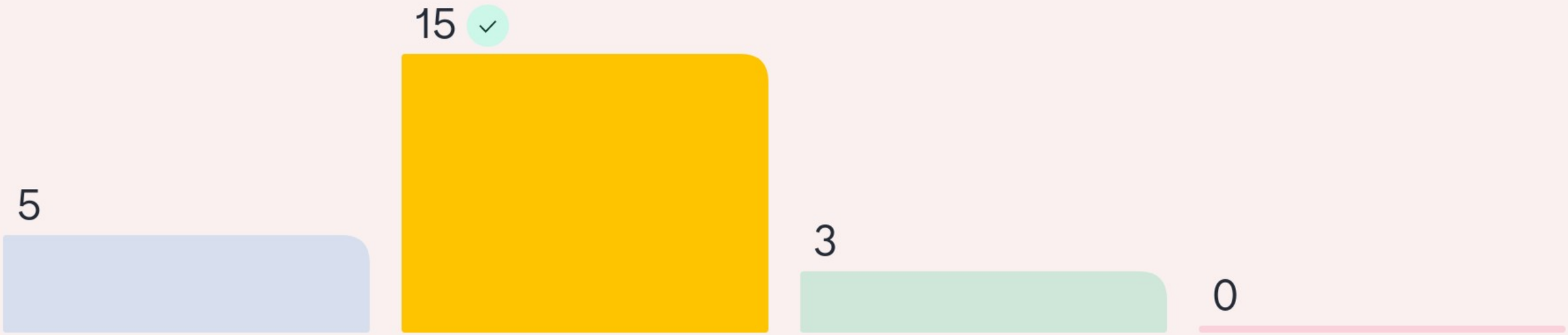
4. Standard errors

- Statistical hypothesis testing is dependent on *statistical significance*

- Recall that economic significance ≠ statistical significance
  - A large effect may be imprecise
  - A small, but stat. sign. effect may be irrelevant

- Always discuss both economic and statistical significance

Assume $\log p_i = \alpha + \beta park_i + \gamma x_i + \epsilon_i$. Given that $\beta = 0.01(0.003)$, is this economically/statistically significant?



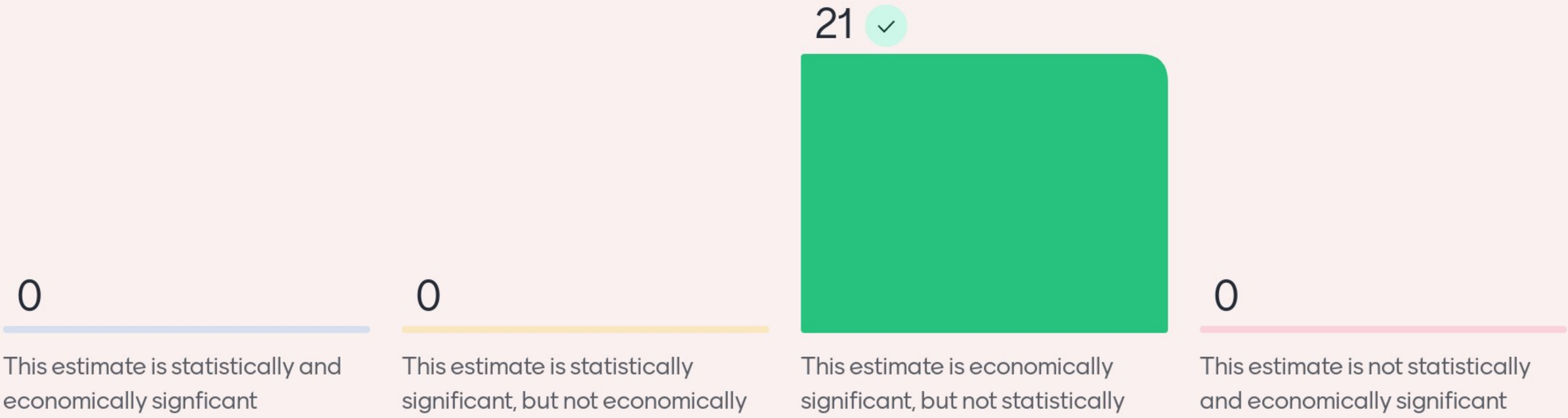
This estimate is statistically and economically signficant

This estimate is statistically significant, but not economically

This estimate is economically significant, but not statistically

This estimate is not statistically and economically significant

Assume $\log p_i = \alpha + \beta park_i + \gamma x_i + \epsilon_i$. Given that $\beta = 0.22(0.15)$, is this economically/statistically significant?



5. Summary

Today:

- **Several identification strategies have been discussed to measure causal effects**
  - **Quasi-experiments**
    - **→ RDD**

- **Some remarks on standard errors**
  - **Heteroscedasticity is easy to address**
  - **Clustering is important**
  - **Economic vs. statistical significance**

5. Summary

Identification:

- Setting up a research project
- Discuss why RCT measures an average causal effect of a treatment
- Alternatives to RCTs
  - OLS with controls
  - IV
  - Quasi-experimental methods
- More economic reasoning than pure econometrics!

Identification (3)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute