Identification (1)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute

1. Introduction

- **Today:**
  1. Spatial econometrics
  2. Discrete choice
  3. **Identification**

- **Tomorrow:**
  4. **Hedonic pricing**
  5. **Quantitative spatial economics**

1. Introduction

- **Today:**
  1. Spatial econometrics
  2. Discrete choice
  3. **Identification**
     - **Research design, IV, OLS, RDD, Quasi-experiments**

- **Tomorrow:**
  4. **Hedonic pricing**
  5. **Quantitative spatial economics**

1. Introduction

- Academics usually aim to identify *causal* effects

- Causal effects: one process, *a cause*, contributes to the production of another process
  - the effect of a 'treatment' variable x on an outcome variable y

1. Introduction

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

Ill. Niklas Elmehed © Nobel Prize Outreach.

David Card

"for his empirical contributions to labour economics"

Ill. Niklas Elmehed © Nobel Prize Outreach.

Joshua D. Angrist

Ill. Niklas Elmehed © Nobel Prize Outreach.

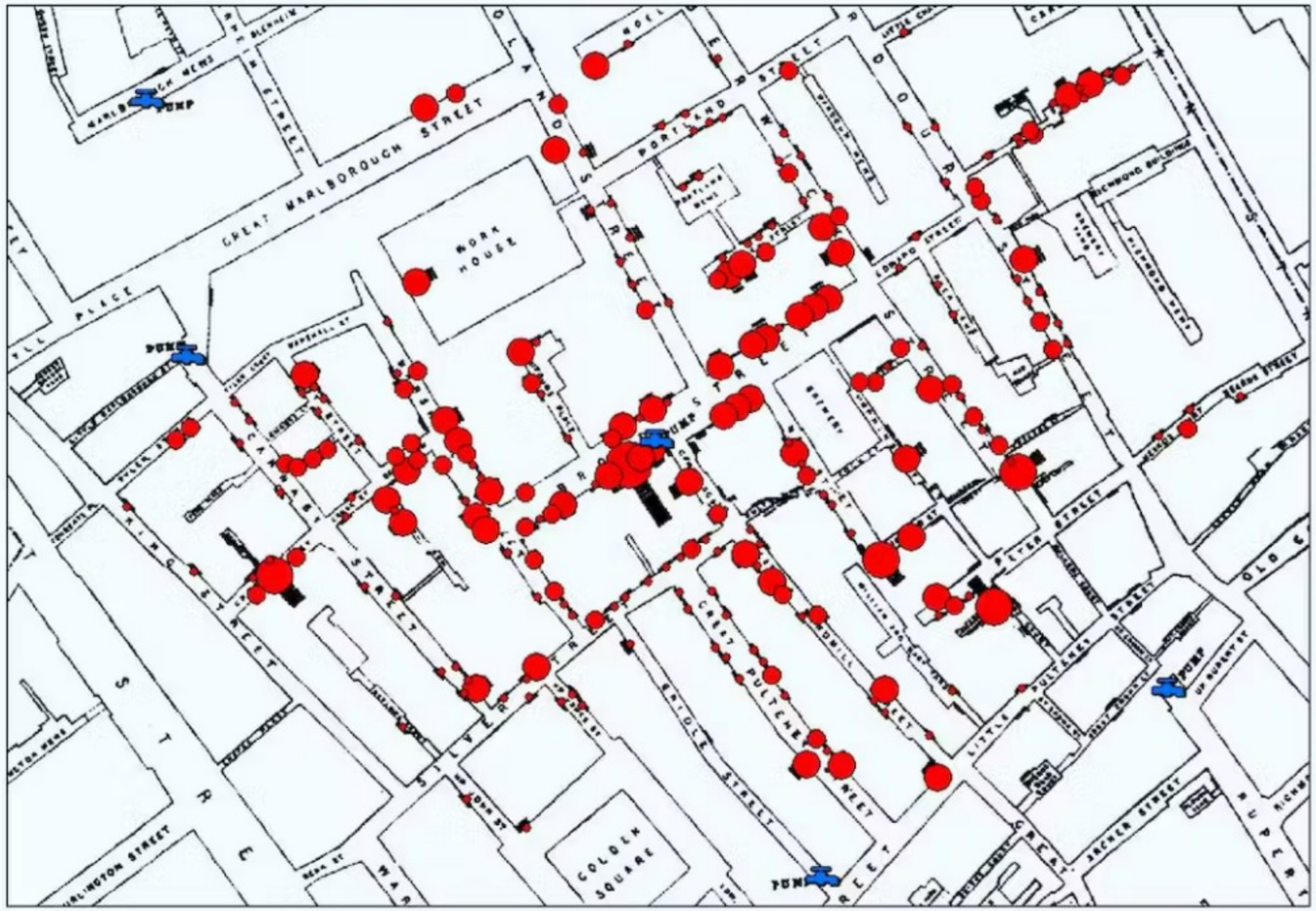
Guido W. Imbens

"for their methodological contributions to the analysis of causal relationships."

October 11, 2021

1. Introduction

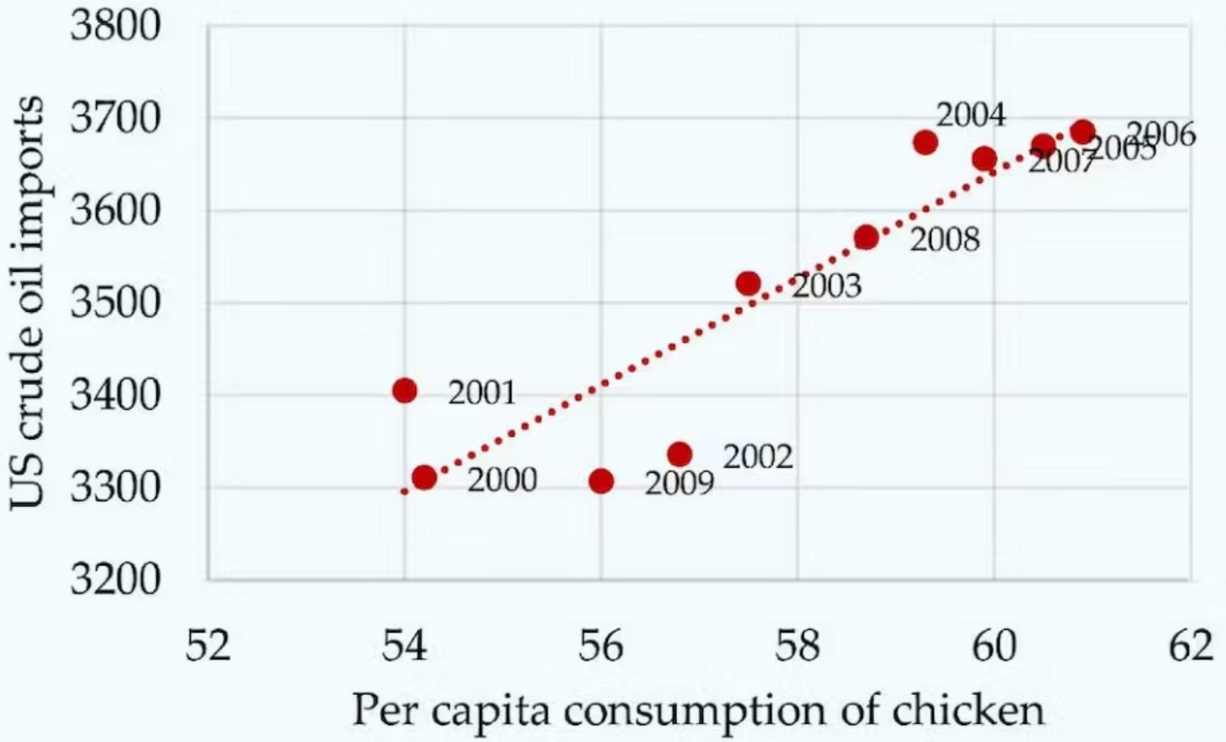
- Interesting correlations



- Spatial correlation of cholera deaths in 1854
  - John Snow
  - Contaminated water…

1. Introduction

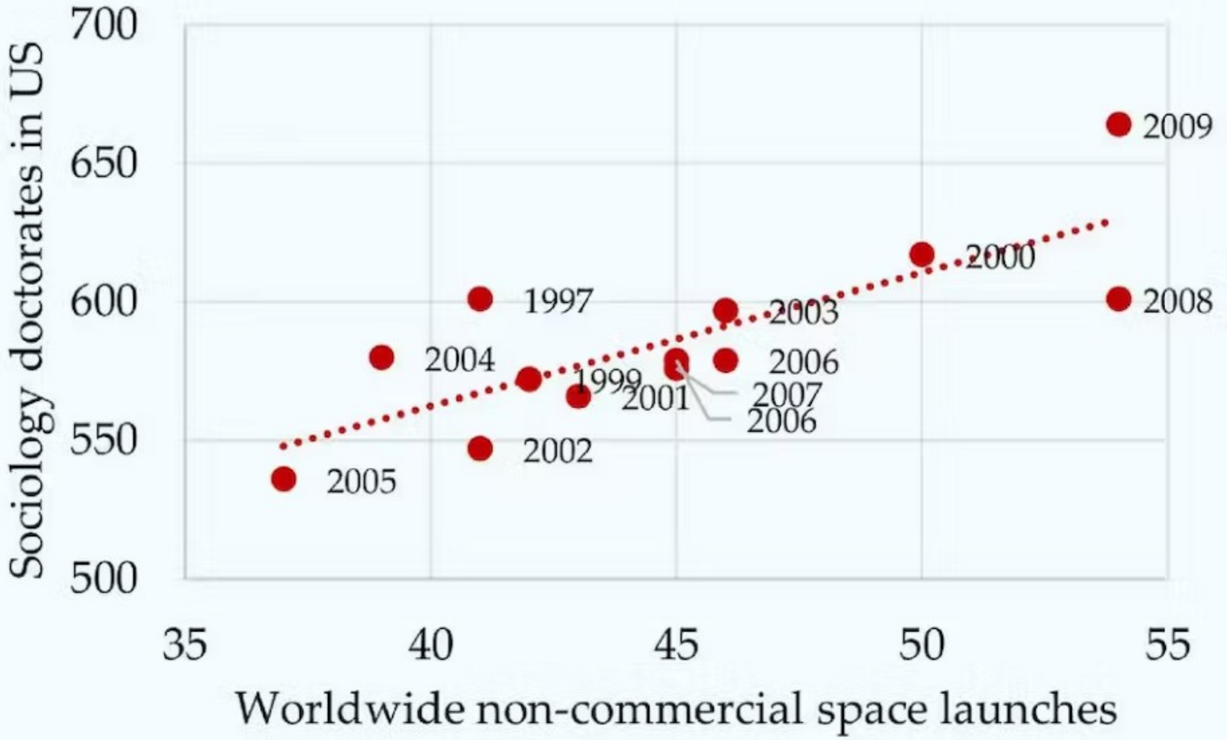
Spurious correlations



$\rho = 0.90$

1. Introduction

Spurious correlations



$\rho = 0.79$

Today

- Setting up a research project

- Alternatives to RCTs
  - OLS with controls
  - IV
  - Quasi-experimental methods

- More economic reasoning than pure econometrics!

1. Introduction

- Today
  - Learn to set up your own research project
  - … and think about identification issues

- Plan:

  Part #1: Research design
  Part #2: Randomised experiments, OLS, IV
  Part #3: Quasi-experiments, RDD

2. Research design

- 8 steps when undertaking research

1. Formulate your hypotheses
2. Determine the 'treatment' variable(s) and the 'outcome' variable(s)
3. Think of an identification strategy to identify causal effects
4. Select samples, discuss measurement error and provide descriptives
5. Determine functional form of variables of interest
6. Think of different issues in estimating standard errors
7. Estimate model and interpret the results
8. Provide robustness checks of the results

2. Research design

1. Formulate your hypotheses

- Economic hypotheses

- Based on economic theory

- Humans often use *reverse* causal reasoning
  - *"House prices have gone down the last years, but why?"*
  - *Forward* causal inference supplies answers
  - *Reverse* causal inference supplies questions

2. Research design

2. Determine the '<u>treatment</u>' variable(s) and the '<u>outcome</u>' variable(s)

- Define what variables are available in your data
- Focus on one (or a few) x variable(s) and one (or a few y variables
- Think about expected order of magnitude

3. Think of an identification strategy to identify causal effects

- What is your 'treatment' group and what is your 'control' group?

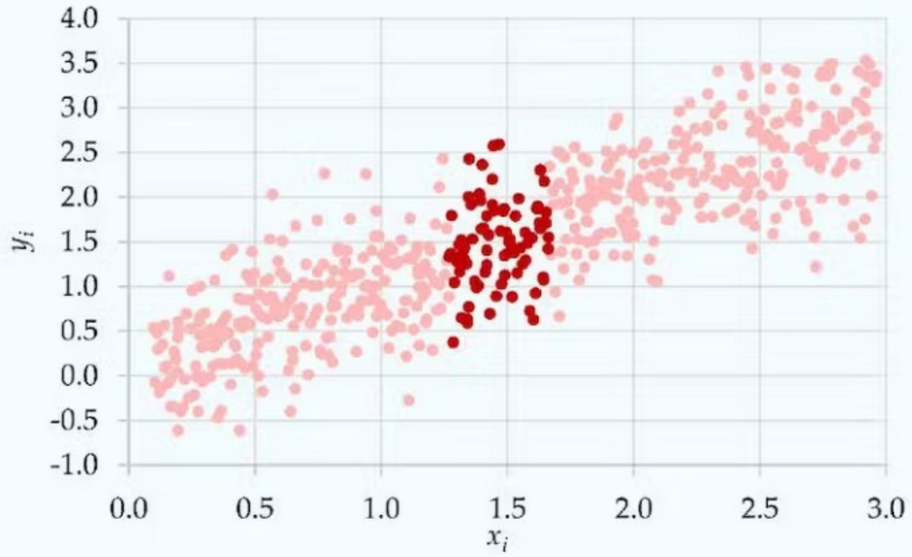
- Discuss endogeneity issues
  - Might there be a selection effect?
  - What are potential unobserved factors? Are these correlated with the treatment status?
  - Reverse causality?
  - *(Measurement error?)*

3. Think of an identification strategy to identify causal effects

- Define the appropriate econometric methods
  - Discuss the identifying assumptions at length!

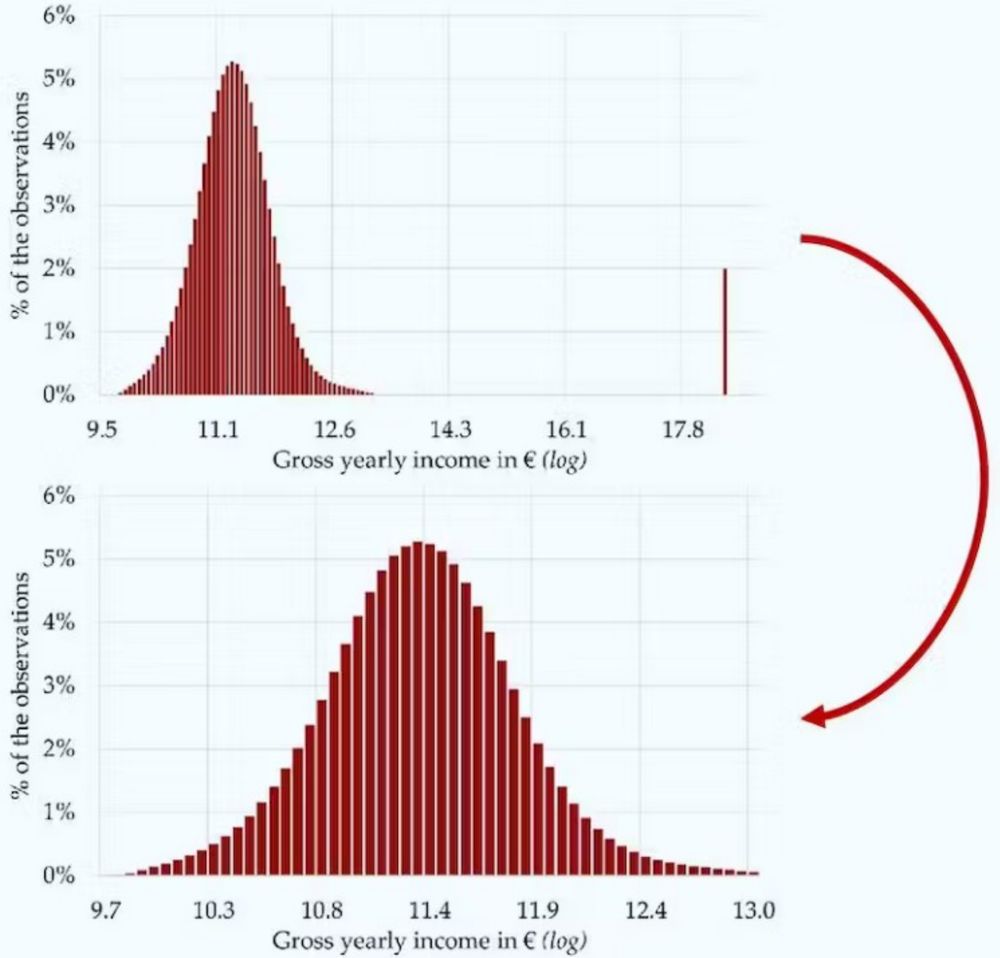
4. Select samples, discuss measurement error and provide descriptives

- Should you use the full dataset?
- Variance in x is necessary!



4. Select samples, discuss measurement error and provide descriptives

- **Data cleaning**

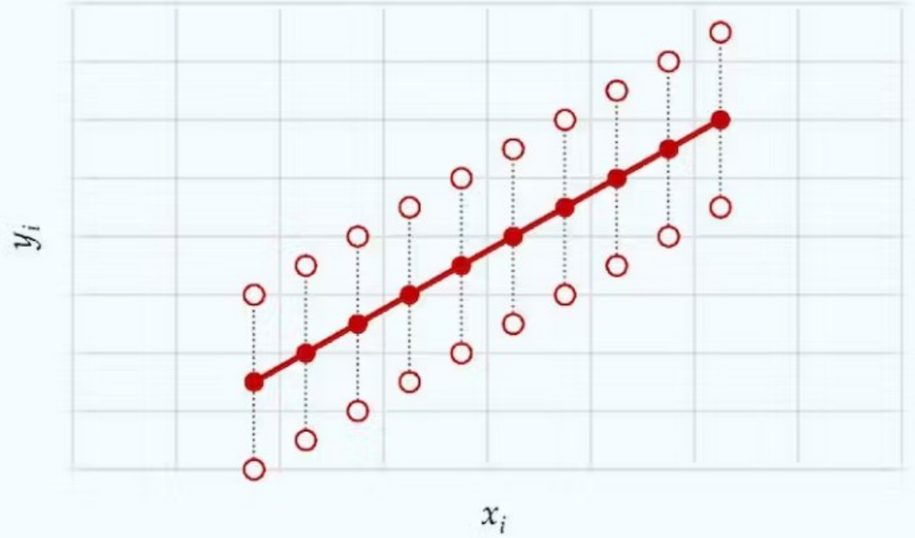


4. Select samples, discuss measurement error and provide descriptives

- Measurement error is present in many datasets

4. Select samples, discuss measurement error and provide descriptives

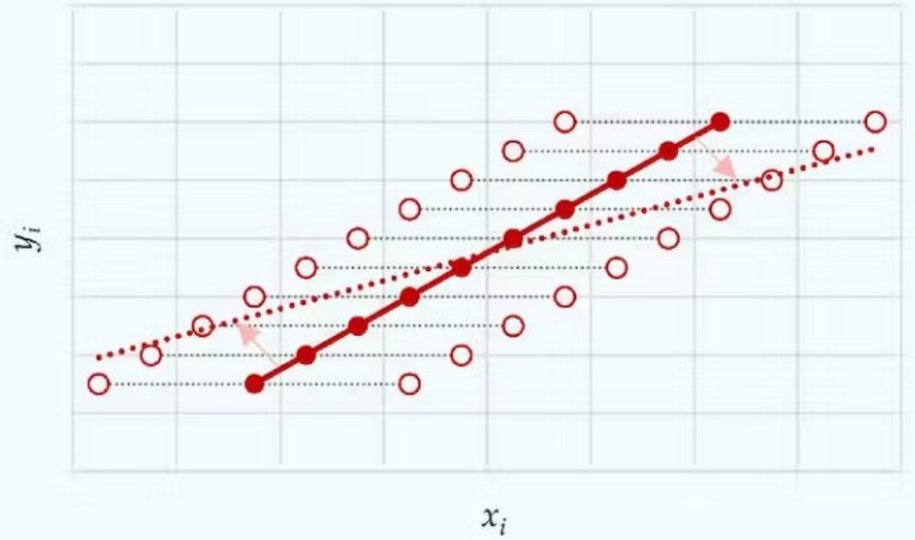
- Random measurement error in y is not so much of a problem



- $y_i^* - u_i = \beta x_i + \epsilon_i \rightarrow y_i^* = \beta x_i + (\epsilon_i + u_i)$

4. Select samples, discuss measurement error and provide descriptives

- Random measurement error in x biases the effect towards zero



- $y_i = \beta(x_i + u_i) + \epsilon_i$ ➔ $\beta \to 0$ if u_i is large

5. Determine functional form of variables of interest

- The specification of $f(\cdot)$ is referred to as the functional form

$$y_i = f(x_i, c_i, \epsilon_i)$$

- Often a linear functional form is assumed:

$$y_i = \beta x_i + \gamma c_i + \epsilon_i$$

5. Determine functional form of variables of interest

- Economists are often interested in elasticities
  - Elasticity is percentage change of y in response to a change in x
  - $\frac{\partial y}{\partial x}\frac{x}{y}$

- They therefore often estimate log-linear regressions:

$$\log y_i = \beta \log x_i + \gamma c_i + \epsilon_i$$

  - because $\beta = \frac{\partial \log y}{\partial \log x} = \frac{\partial y}{\partial x}\frac{x}{y}$

5. Determine functional form of variables of interest

- When use logs?
  - Economic theory
  - Residuals have a skewed distribution
  - Heteroscedasticity
  - Different unit sizes

6. Think of different issues in estimating standard errors

- Whether β is statistically significant depends on standard error
  - The smaller the standard error, the more precise your conclusions are

- Issues to bear in mind…
  - Should you cluster your standard errors?
  - Is heteroscedasticity a problem?
  - Is there serial/spatial autocorrelation?

7. Estimate model and interpret the results

- Use statistical software to estimate your model

- Usually we are interested in marginal effects
  - How much does y change (in units or %) when x change with one unit (or %)
  - $\frac{\partial y}{\partial x}$ (in levels) or $\frac{\partial y}{\partial x}\frac{x}{y}$ (in %)

7. Estimate model and interpret the results

- Properly interpret β and its statistical significance
  - "When x increases by 1 *(units)* y increases by .. *(units)*. This effect is statistically significant at the ...% level."
  - Specify the units!

7. Estimate model and interpret the results

- Statistical hypothesis testing is dependent on *statistical significance*

- 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

- See McCloskey and Ziliak (1996)

7. Estimate model and interpret the results

- **Make sure how your variables are measured**
  - logs, dummies, etc.

Specifications:	x	$\log x$
y	$\boldsymbol{y = \rho x + \eta}$ $\hat{\rho} = \frac{\partial y}{\partial x}$ $x \uparrow 1 \rightarrow y \uparrow \hat{\rho}$	$\boldsymbol{y = \rho \log x + \eta}$ $\hat{\rho} = \frac{\partial y}{\partial \log x}$ $x \uparrow 1\% \rightarrow y \uparrow \hat{\rho}/100$
$\log y$	$\boldsymbol{\log y = \rho x + \eta}$ $\hat{\rho} = \frac{\partial \log y}{\partial x}$ $x \uparrow 1 \rightsquigarrow y \uparrow (\hat{\rho} * 100)\%$ *(for marginal changes in x)*	$\boldsymbol{\log y = \rho \log x + \eta}$ $\hat{\rho} = \frac{\partial \log y}{\partial \log x}$ $x \uparrow 1\% \rightarrow y \uparrow \hat{\rho}\%$

7. Estimate model and interpret the results

- **Note on larger changes in x in log-linear regressions.**

- **Let's assume the model** $\log y = \rho x + \epsilon$**, with** $x \in 0.1$**.**
  - **Example: dummy variables**
  - **Halvorsen & Palmquist:** $x \uparrow 1 \rightarrow y \uparrow ((e^{\hat{\rho}} - 1) * 100)\%$

8. Provide robustness checks of the results

- You make many somewhat arbitrary choices

- Test for sensitivity of your results with respect to these choices
  - … sensitivity analysis

3. Summary

Today:

- Economists are generally interested in *causal* effects

- 8 steps when undertaking research
  1. Formulate your hypotheses
  2. Determine the 'treatment' variable(s) and the 'outcome' variable(s)
  3. Think of an identification strategy to identify causal effects
  4. Select samples, discuss measurement error and provide descriptives
  5. Determine functional form of variables of interest
  6. Think of different issues in estimating standard errors
  7. Estimate model and interpret the results
  8. Provide robustness checks of the results

Identification (1)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute

Identification (2)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute

- **8 steps when undertaking research**

1. Formulate your hypotheses
2. Determine the 'treatment' variable(s) and the 'outcome' variable(s)
3. **Think of an identification strategy to identify causal effects**
4. Select samples, discuss measurement error and provide descriptives
5. Determine functional form of variables of interest
6. Think of different issues in estimating standard errors
7. Estimate model and interpret the results
8. Provide robustness checks of the results

1. Introduction

- In economics, identification of causal effects is of key importance
  - Step 3 is key → *"think of an identification strategy to identify causal effects"*

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

2. Randomised experiments

- In most economic studies, RCTs are not applied
  - No experimental setting possible
  - Ethical reasons / fairness
  - Costly
  - Expected substantial heterogeneity in outcomes
  - Hard to measure long-run effects
  - Lab setting may bias outcomes
    - » Recall: biases in Stated Preference surveys

2. Randomised experiments

- A more philosophical critique on RCTs
  - We might find a causal effect of x on y, but do not know *why* there is an effect

- Theoretical models and reasoning are needed to explain *why* we would expect a causal effect
  - Deaton (2010)

3. OLS with controls

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

3. OLS with controls

- Use an exhaustive set of controls
  - In some applications, you might know all explanatory variables

- For example, computers?
  - You aim to know the willingness to pay for a new processor
  - $price_i = \rho(processor\ quality_i) + (characteristics)_i'\gamma + \eta_i$

- Not all characteristics are available in the data
  - Houses, cars, etc.
  - Omitted variable bias…

3. OLS with controls

- Use first-differencing or fixed effects to make this approach more convincing
  - Controls for all time-invariant factors
  - Requires '*within*' variation

3. OLS with controls

- **First-differencing**

  $\Delta y_{it} = \Delta\alpha_t + \beta\Delta x_{it} + \gamma\Delta c_{it} + \Delta\epsilon_i$

  where $\Delta y_{it} = y_{it} - y_{it-1}$**, etc.**

- **This controls for all time-invariant characteristics of** i
  - **Hence there should be variation in** x_{it} **over time**

3. OLS with controls

- **Fixed effects**

$$y_{ig} = \bar{y}_i + \beta(x_{ig} - \bar{x}_g) + \gamma(c_{ig} - \bar{c}_g) + (\epsilon_{gt} - \bar{\epsilon}_g)$$
$$= \beta x_{ig} + \gamma c_{ig} + \mu_g + \epsilon_{ig}$$

where μ_g is a fixed effect at the level of group g

- Fixed effects vs. first-differencing
  - Identical to first-differencing when having two observations per group
  - Fixed effects is more efficient

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

4. IV

- **Find valid instrumental variables**

- $x_i = \zeta + \eta z_i + \xi_i$ *(1st stage)*
  $y_i = \alpha + \beta \hat{x}_i + \epsilon_i$ *(2nd stage)*

→ **What are the two conditions for valid instruments?**

- **There are two conditions for valid instruments**

I. **Instrument relevance**: $\text{cov}[z_i, x_i] \neq 0$
   - μ **should be statistically significant and strong**
   - **Rule-of-thumb:** $F > 10$
   - **Use Kleibergen-Paap** F**-statistic with multiple endogenous variables**

II. **Instrument exogeneity**: $\text{cov}[z_i, \epsilon_i] = 0$
   - **Instrument should not be correlated to error term**
   - **Instrument should only influence** y **via** x
   - ***Based on economic reasoning***

4. IV

- Use exogenous sources of variation
  - Use economic models to find valid instruments
  - Use national policies or natural shocks, etc. to construct instrument
  - Use historical/long-lagged instruments

4. IV

- IV identifies local average treatment effect
  - Say that the instrument is only observed for a certain group, then IV identifies treatment effect for this group
  - Different instruments may lead to different β

- Example: gender of children and housing demand

Identification (2)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute

Identification (3)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute

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:**

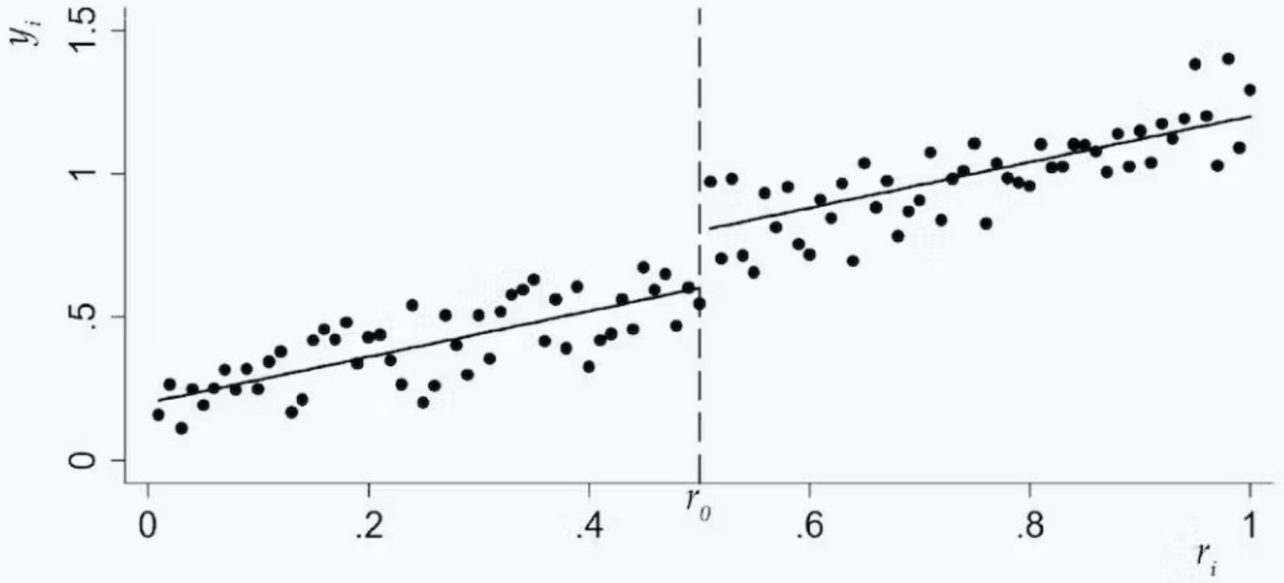
$$y_i = \alpha + \beta x_i + \gamma r_i + \epsilon_i$$

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

3. RDD

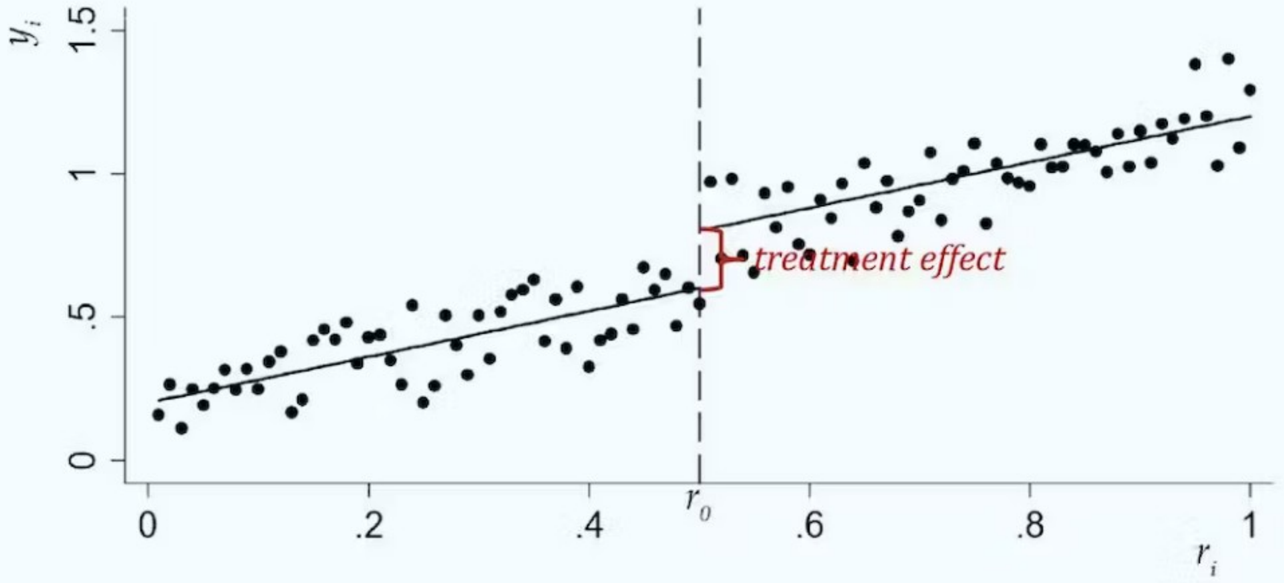
- **Example:**
  - **Students get a scholarship if they achieve a certain test-score**
  - **You aim to know the impact of the scholarship on job market outcomes**
    - » **e.g. wages**

- **Plot**



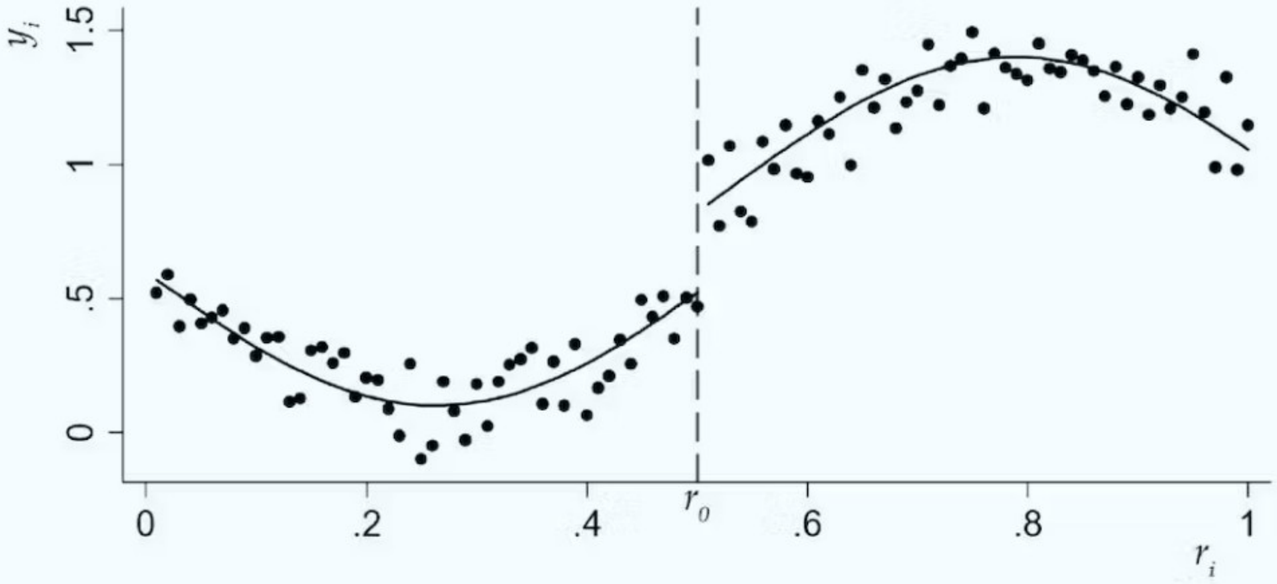
- **Control for test scores and investigate the jump in treatment at r_0**

- **Plot**

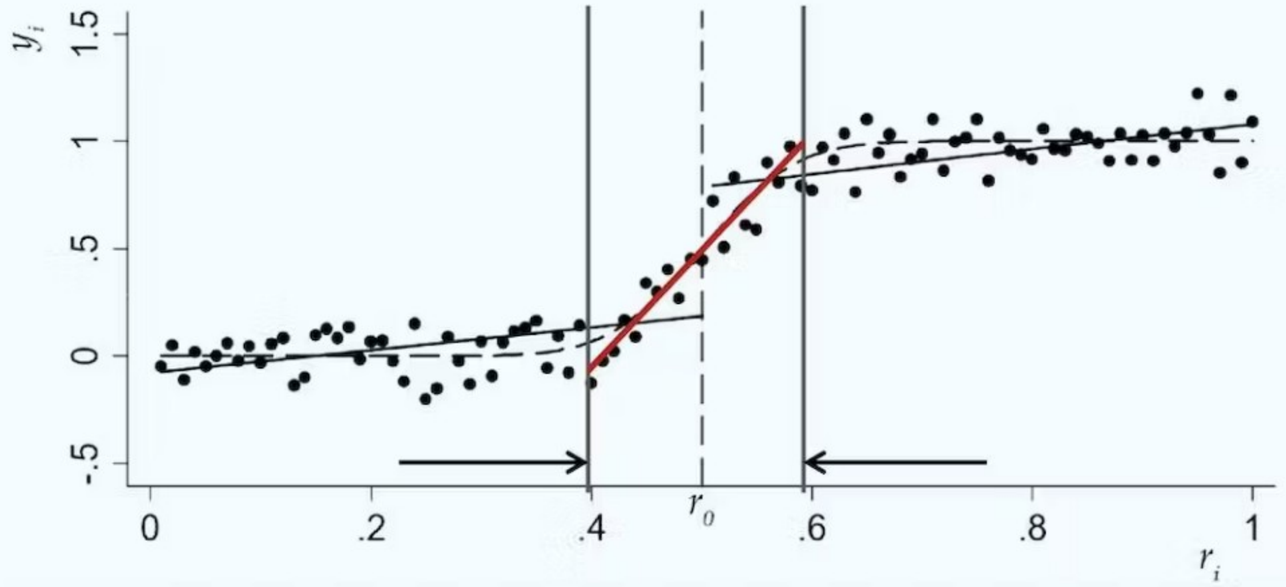


- **Control for test scores and investigate the jump in treatment at r_0**

- **What if x is non-linearly related to Y**
  - $y_i = \alpha + \beta x_i + f(r_i) + \epsilon_i$
  - **Specify $f(r_i) = \gamma_1 r_i + \gamma_2 r_i^2 + \cdots + \gamma_q r_i^q$**
    - » **q^{th}-order polynomial**
    - » **Can be estimated by OLS**



- 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



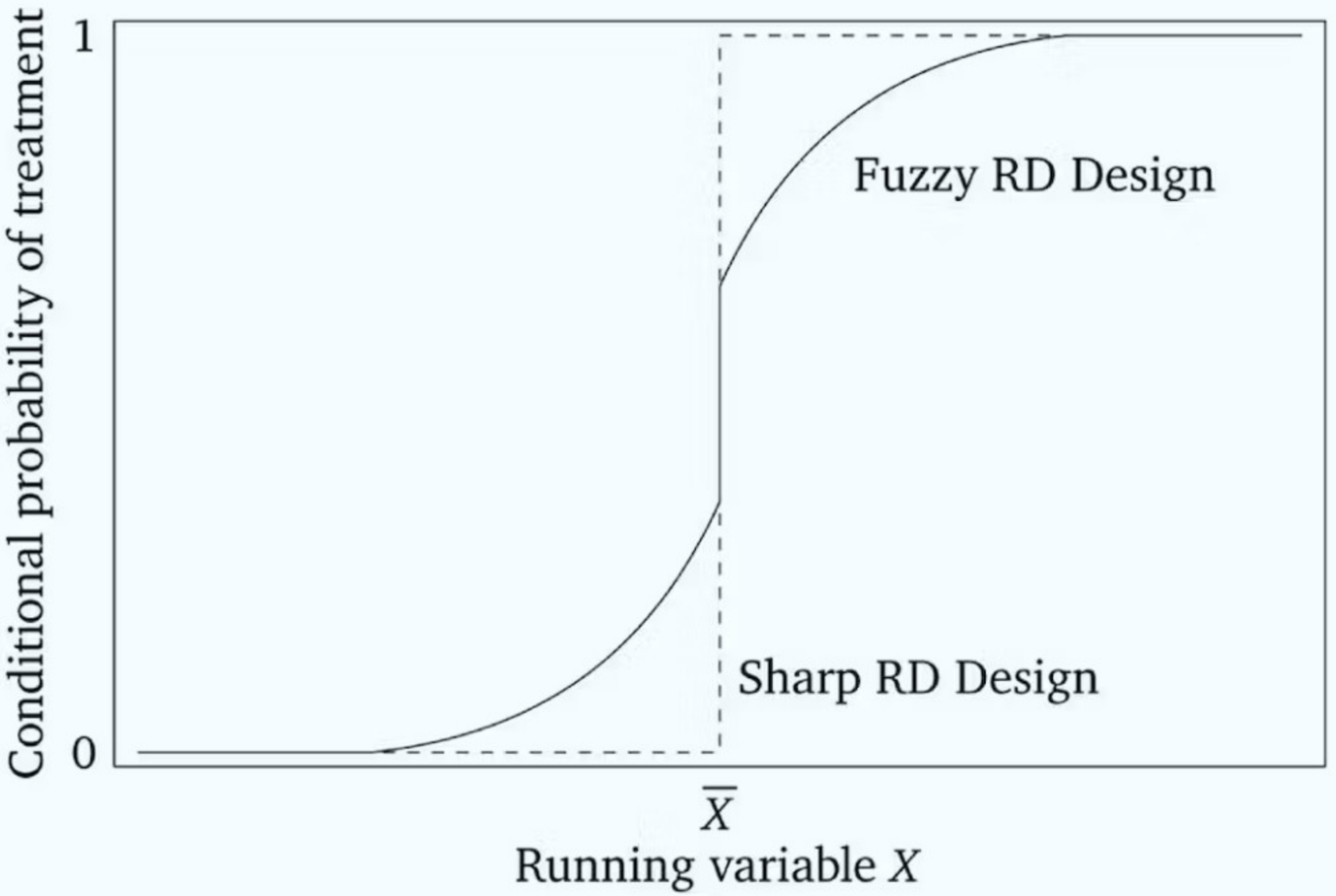
3. RDD

- 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 →** $y_i = \alpha + \beta \hat{x}_i + f(r_i) + \epsilon_i$

5. Summary

Today

- Setting up a research project

- Alternatives to RCTs
  - OLS with controls
  - IV
  - Quasi-experimental methods

5. Summary

- 8 steps when undertaking research

1. Formulate your hypotheses
2. Determine the 'treatment' variable(s) and the 'outcome' variable(s)
3. Think of an identification strategy to identify causal effects
4. Select samples, discuss measurement error and provide descriptives
5. Determine functional form of variables of interest
6. Think of different issues in estimating standard errors
7. Estimate model and interpret the results
8. Provide robustness checks of the results

Identification (3)

Applied Econometrics for Spatial Economics

Hans Koster

Professor of Urban Economics and Real Estate

VU VRIJE UNIVERSITEIT AMSTERDAM

tinbergen institute