Causal Inference: Methods and Applications

Dr. Murat M. Tunc

Data-driven Thesis

- 2^{nd} Part of **Secondary Data Methods**
  - Building up on last week's session by Dr. Poonacha Medappa
- **1) Hypothesis development**
  - Behavioral / cognitive / economic understanding
  - Why / how / when / **mechanisms of an effect**
- **2) Data collection**
  - APIs, web scraping, public databases
  - **Panel data:** Same individuals over multiple periods
- **3) Hypothesis testing**
  - Linear regression with **fixed effects**
    - Control for unobserved time-invariant factors
  - **Causal inference with quasi-experimental methods**

Causal Inference: Methods and Applications

Dr. Murat M. Tunc

Do trees make our cities safer?

- In city areas with nearby trees and natural landscapes
  - **Less domestic violence**
- On tree-lined streets
  - People drive more slowly, **reducing accident risk**
- Trees contribute to stronger ties among neighbors
  - Closer supervision of children in outdoor places
  - **Fewer** property and violent crimes
- Adolescents live in neighborhoods with more greenery
  - Display **less aggressive behavior**

Classes of Variables

- 1) The **outcome** variable
  - Dependent variable
- 2) **Principal** question **predictor**
  - Variable of interest
- 3) Covariates or **control predictors**
  - Independent control variables

Do trees make our cities safer?

City	Crime Rate	Tree Density	Population
Dallas, Texas	0.4	0.15	7,000,000
Tilburg, Netherlands	0.01	0.5	200,000
Albuquerque, New Mexico	0.9999	0.05	1,500,000
Antwerp, Belgium	0.05	0.25	500,000

- Regression
  - The **outcome** variable: Crime Rate
  - **Variable of interest**: Tree Density
  - Control variable: Population
- Estimation:
  - **Significant** (p-value < 0.01) **and negative**
- Conclusion:
  - **Trees make our cities safer**

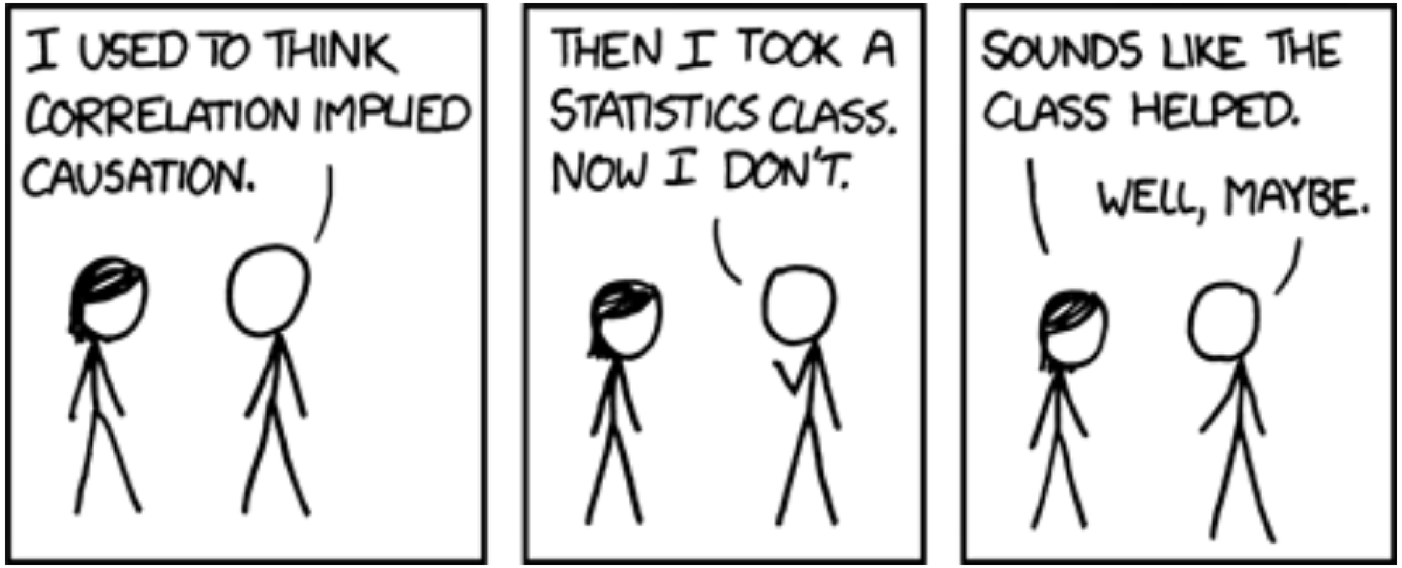
TILBURG UNIVERSITY

Correlation vs. Causation

- Is this effect a causal effect?
  - Are trees **the reason why** crime is lower in cities?
- If it's a causal effect
  - Police should **plant lots of trees** to Albuquerque, New Mexico and **crime rate will plummet**
- No, it's a **correlation** between variables
- **Selection bias**
  - People **choose** where to live
  - Suppose high-income people tend to **commit fewer crimes**
  - High income people also **like living in neighbors** with lots of trees

Do trees make our cities safer? Well, maybe

- Can we conclude that trees do **NOT** make our cities safer?
  - No
  - Trees **may**, in fact, make our cities safer
  - But, **given this dataset**, it is not possible to know whether they do
- How can we estimate causality?



Fundamental problem of causal inference

- How to test whether trees make our cities safer?
  - **Plant 1 million trees in a city** vs. **Don't plant any trees in a city**
    - **Treatment** vs. **Control**
    - **Compare the crime rates**
- Unit level **causal effect**
  - **Difference in outcome**, holding all other variables fixed

City	Crime Rate with Treatment	Crime Rate without Treatment	Causal Effect
A	0.16		?
B		0.04	?
C		0.01	?
D	0.23		?

TILBURG

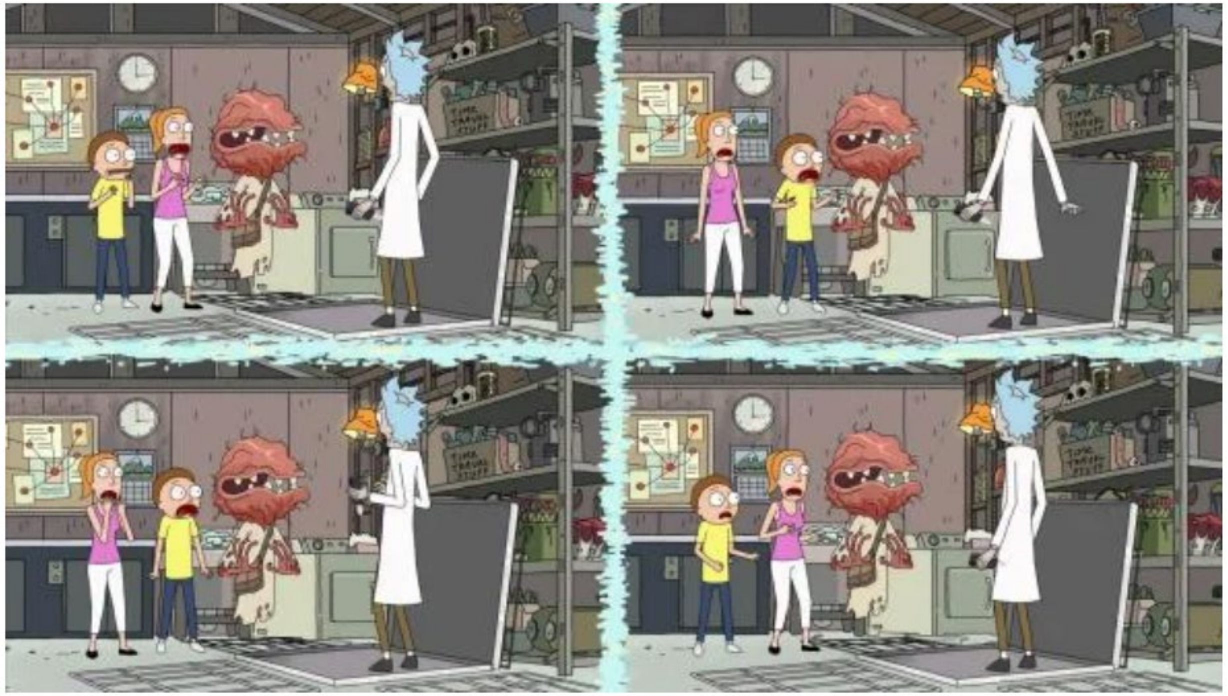
Fundamental problem of causal inference

- We can only **observe one outcome**
  - **Factual**
- We **never observe counterfactual**
  - What would have happened if
    - Germany won WW2
  - What would have happened if
    - Harry Potter and Draco Malfoy became friends
- **Causal inference is a missing data problem**

	… the value of the outcome in the ***Treatment Group*** is …	… the value of the outcome in the ***Control Group*** is …
For members of the ***Treatment Group*** …	Known	Missing
For members of the ***Control Group*** …	Missing	Known

Ideal Experiment

- Parallel worlds
  - **World 1:** Albuquerque
    - Plant 1 million trees
  - **World 2:** Albuquerque
    - Do not plant any trees
  - **Compare the worlds**



Tilburg University

How to approximate the ideal experiment?

- **Mice** and **Dice**
- **Mice:**
  - Control group
  - Treatment group
  - Both control and treatment group
    - **Equal in expectation**
- **Dice:**
  - **Random assignment** into control and experiment group
  - Exogenous variation

Experiment: Do trees make our cities safer?

- **Mice:**
  - Albuquerque, Dallas, Tilburg, Antwerp, New York, …
- **Dice:**
  - **Randomly assign** cities to the treatment group
  - **Treatment group**: Dallas, Antwerp, New York, Baltimore
    - Plant 1 million trees
  - **Control group**: Tilburg, Albuquerque, London, Hong Kong
    - Don't plant any trees

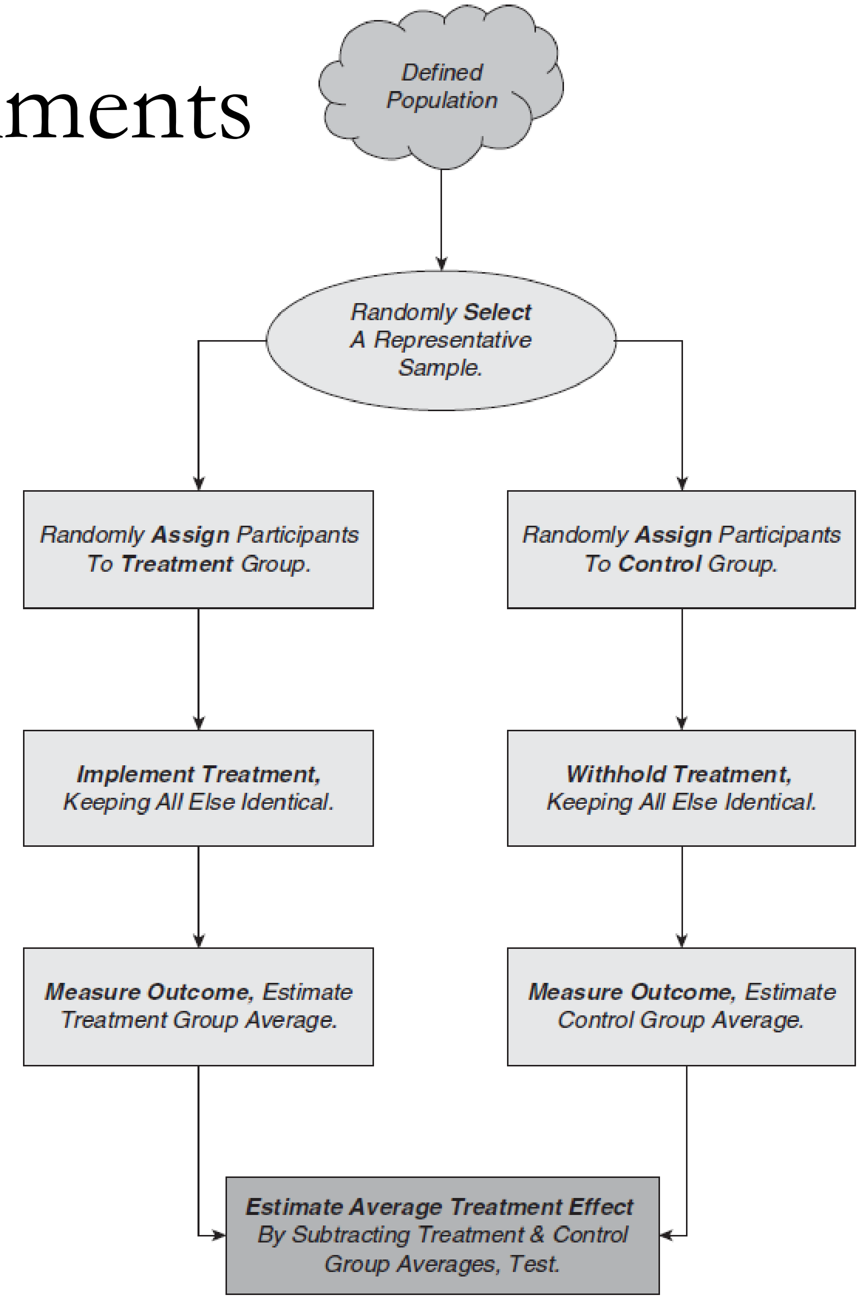
Mice and Dice

- Medical researchers **can do it in a lab**
- Economists **cannot do it in the real world**
  - Due to unethical reasons
- But, if there exists **an exogenous source of randomness**
  - **Assigns** people to control and treatment group
  - We **can establish causality**

Empirical Identification Strategies

1. Randomized Experiments
2. Natural Experiments / Difference-in-Differences
3. Regression Discontinuity
4. Instrumental Variables

Randomized Experiments



Tilburg University

Randomized Experiment: Example

- **Research Question:**
  - What is the causal effect of scholarship on academic success?
- **Mice:**
  - In 1997, a **scholarship** of $1,400 will be given to 1,300 children from low-income families in New York City
  - More than **10,000 applications**
- **Dice:**
  - **Lottery determined who gets** the tuition "voucher"
    - Random assignment

Randomized Experiment: Dataset

- The **outcome** variable:
  - Academic success after the 3rd year of the experiment
- **Variable of interest**:
  - Voucher receipt vs. no voucher
- **Covariates**:
  - Academic success before the experiment

Randomized Experiment: Dataset

	s_id	voucher	pre_ach	post_ach
1	42	0	74	83
2	194	0	7.5	4
3	218	1	2.5	3.5
4	261	1	0	26.5
5	304	1	11	2
6	323	1	8.5	15
7	339	1	0	23.5
8	348	1	37	52
9	349	1	71	60
10	386	0	24	13

Randomized Experiments: Methods

- The **better** your **research design**, the **simpler** your data **analysis**

1. Two-group t-test
2. Linear Regression
3. Linear Regression with covariates

Two-group t-test

```
ttest post_ach, by(voucher)
```

Strategy #1: Two-Group t-Test	Number of Observations	Sample Mean	Sample Standard Deviation	Standard Error
VOUCHER = 1	291	26.029	19.754	1.158
VOUCHER = 0	230	21.130	18.172	1.198
Difference		$\boxed{4.899}$		1.683
t-statistic		2.911		
df		519		
p-value		$\boxed{0.004}$		

Tilburg University

Linear Regression

reg post_ach voucher

Strategy #2: Linear Regression Analysis of POST_ACH on VOUCHER

Predictor	Parameter	Parameter Estimate	Standard Error	*t*-Statistic	*p*-value
INTERCEPT	β_0	21.130	1.258	16.80	0.000
VOUCHER	β_1	$\boxed{4.899}$	1.683	2.911	$\boxed{0.004}$
R^2 Statistic		0.016			
Residual Variance		19.072			

Linear Regression with Covariates

reg post_ach voucher pre_ach

Strategy #3: Linear Regression Analysis of POST_ACH on VOUCHER, with PRE_ACH as Covariate

Predictor	Parameter	Parameter Estimate	Standard Error	*t*-Statistic	*p*-value
INTERCEPT	β_0	7.719	1.163	6.64	0.000
VOUCHER	β_1	$\boxed{4.098}$	1.269	3.23	$\boxed{0.001}$
PRE_ACH	γ	0.687	0.035	19.90	0.000
R^2 Statistic		0.442			
Residual Variance		14.373			

TILBURG UNIVERSITY

Empirical Identification Strategies

1. Randomized Experiments
2. Natural Experiments / Difference-in-Differences
3. Regression Discontinuity
4. Instrumental Variables

Natural Experiments

- **Exogenous assignment**
  - Natural disaster
  - Policy change
- **Similar individuals** exposed to different treatments
  - Individuals **do not self-select** into treatment
  - Treatment and control group
    - Equal in expectation

Natural Experiments: Example

- **Research Question:**
  - What is the effect of minimum wage on employment?
- **Mice:**
  - Fast food restaurants in New Jersey and Pennsylvania
- **Dice:**
  - In April 1992, New Jersey increased the minimum wage from $4.25 to $5.05
    - Treatment group
  - Pennsylvania's minimum wage stayed at $4.25
    - Control group

Natural Experiments: Variables

- The **outcome** variable:
  - Employment in fast-food restaurants
- **Variable of interest**:
  - Treatment effect in NJ
  - New Jersey dummy variable * After policy change
- **Covariates**:
  - Average wage, number of open hours

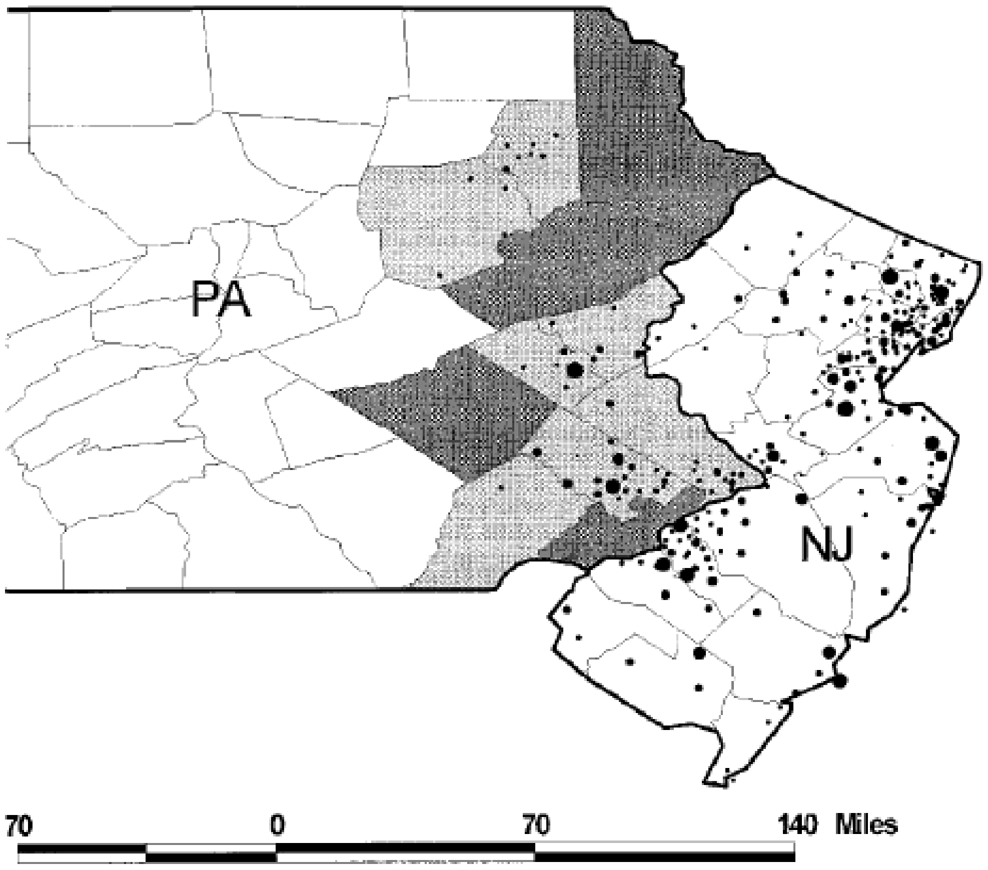
Natural Experiments: Dataset

	store_id	y_ft_emplo~t	d_nj	time
25	13	85	1	0
26	13	59	1	1
27	14	70.5	0	0
28	14	29	0	1
29	15	58	0	0
30	15	29	0	1
31	16	53	1	0
32	16	19	1	1
33	17	52.5	0	0
34	17	34	0	1
35	18	50	1	0
36	18	30	1	1
37	19	48.5	0	0
38	19	27	0	1
39	20	48	1	0
40	20	46.5	1	1
41	21	46.5	1	0
42	21	23.75	1	1

Tilburg University

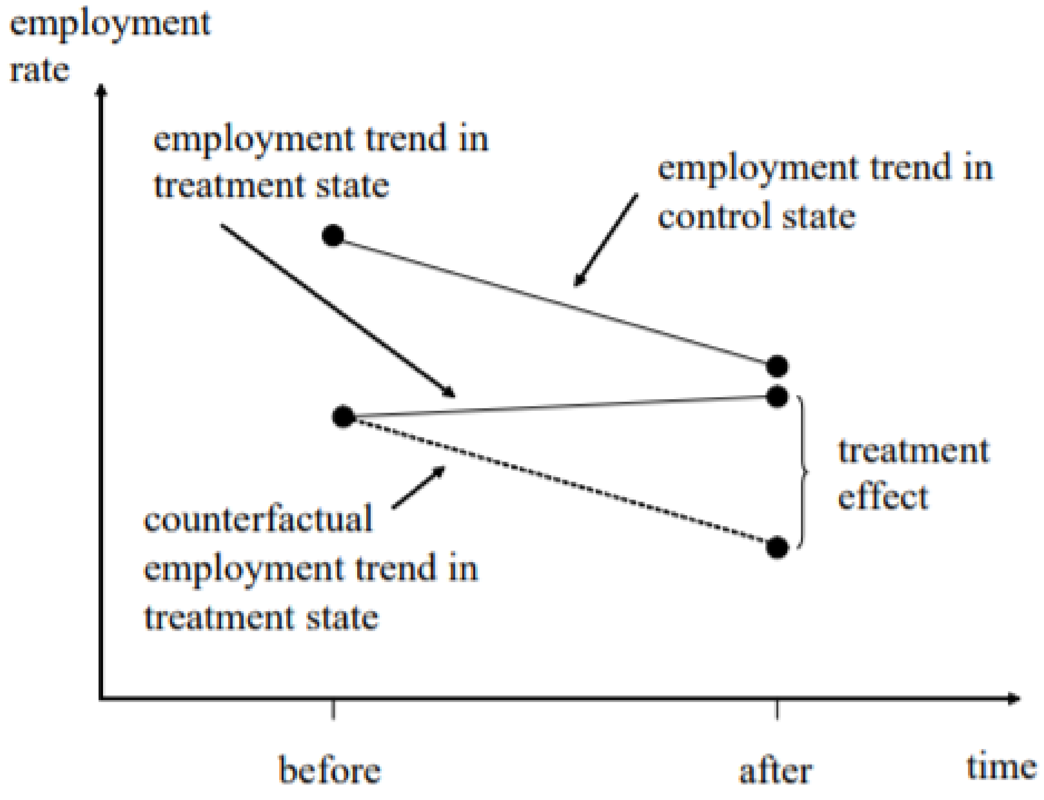
Natural Experiments: Difference-in-Differences

- Panel data
  - **Same** individuals **over multiple times**
- Difference 1:
  - Difference **within individual**
  - **After** the treatment **minus before**
    - NJ in Nov 92 - NJ in Feb 92
    - PA in Nov 92 – PA in Feb 92
- Difference 2:
  - Difference **across individuals**
    - Difference in NJ – Difference in PA



Natural Experiments: Counterfactual

- What would have happened in NJ if
  - The minimum wage **did not increase**
- Assume NJ and PA are
  - **Equal in expectation**
  - **Parallel trends** assumption



TILBURG UNIVERSITY

Difference-in-difference: Estimation

$$Y_{it} = \beta_1 + \beta_2 \text{Treat}_i + \beta_3 \text{Post}_t + \beta_4 (\text{Treat} \times \text{Post})_{it} + \varepsilon_{it}$$

```
xtset store_id time
xtreg y_ft_employment c.d_nj##c.time, fe cluster(store_id)
```

Variable	Stores by state PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	–2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	–0.14 (1.07)
3. Change in mean FTE employment	–2.16 (1.25)	0.59 (0.54)	2.76 (1.36)
4. Change in mean FTE employment, balanced sample of stores[c]	–2.28 (1.25)	0.47 (0.48)	$\boxed{2.75}$ (1.34)

Tilburg University

Difference-in-differences: Robustness

- **Parallel trends**
  - Before the treatment, the dependent variable must be parallel
    - Treatment and control group
- **Matching on observables**
  - Similar individuals between treatment and control group
  - Propensity score matching, IPTW, Coarsened exact matching

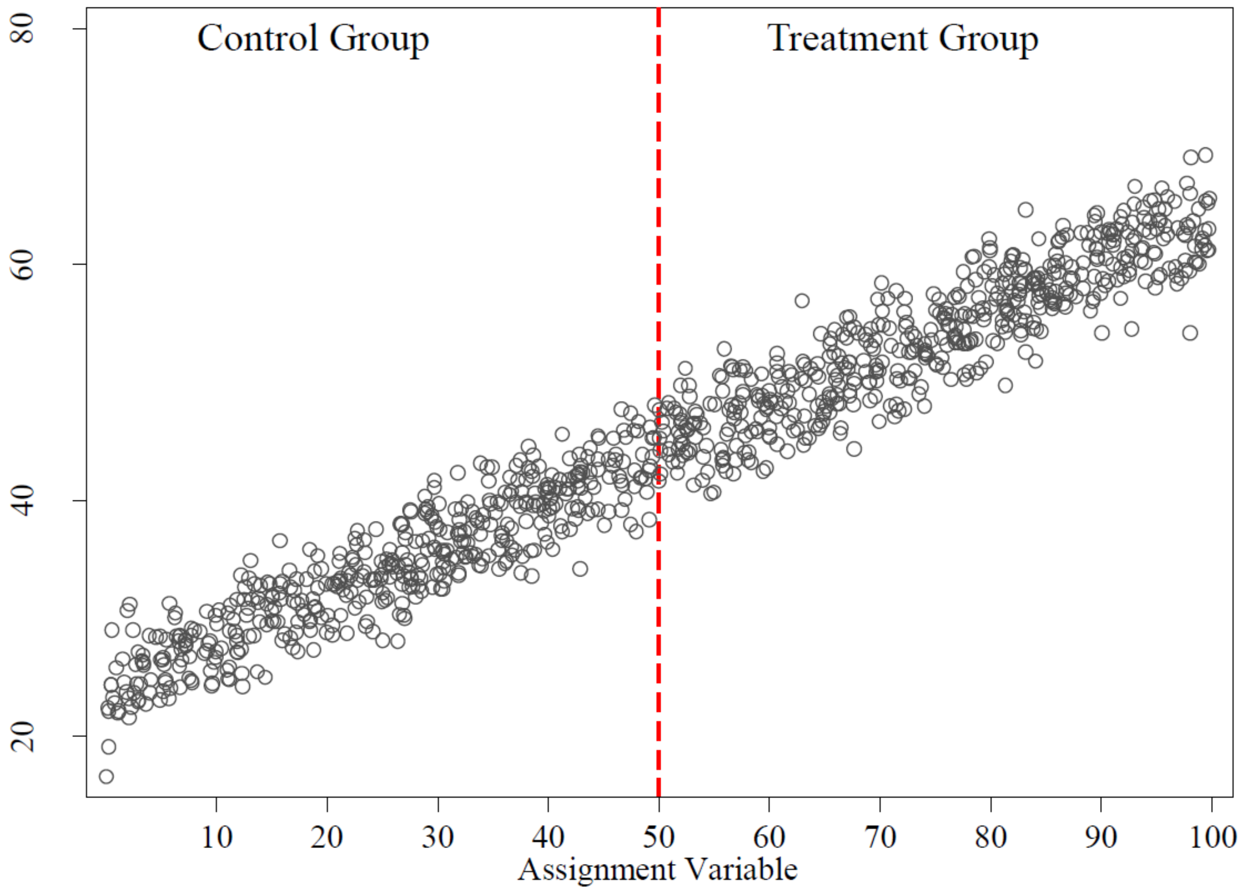
Empirical Identification Strategies

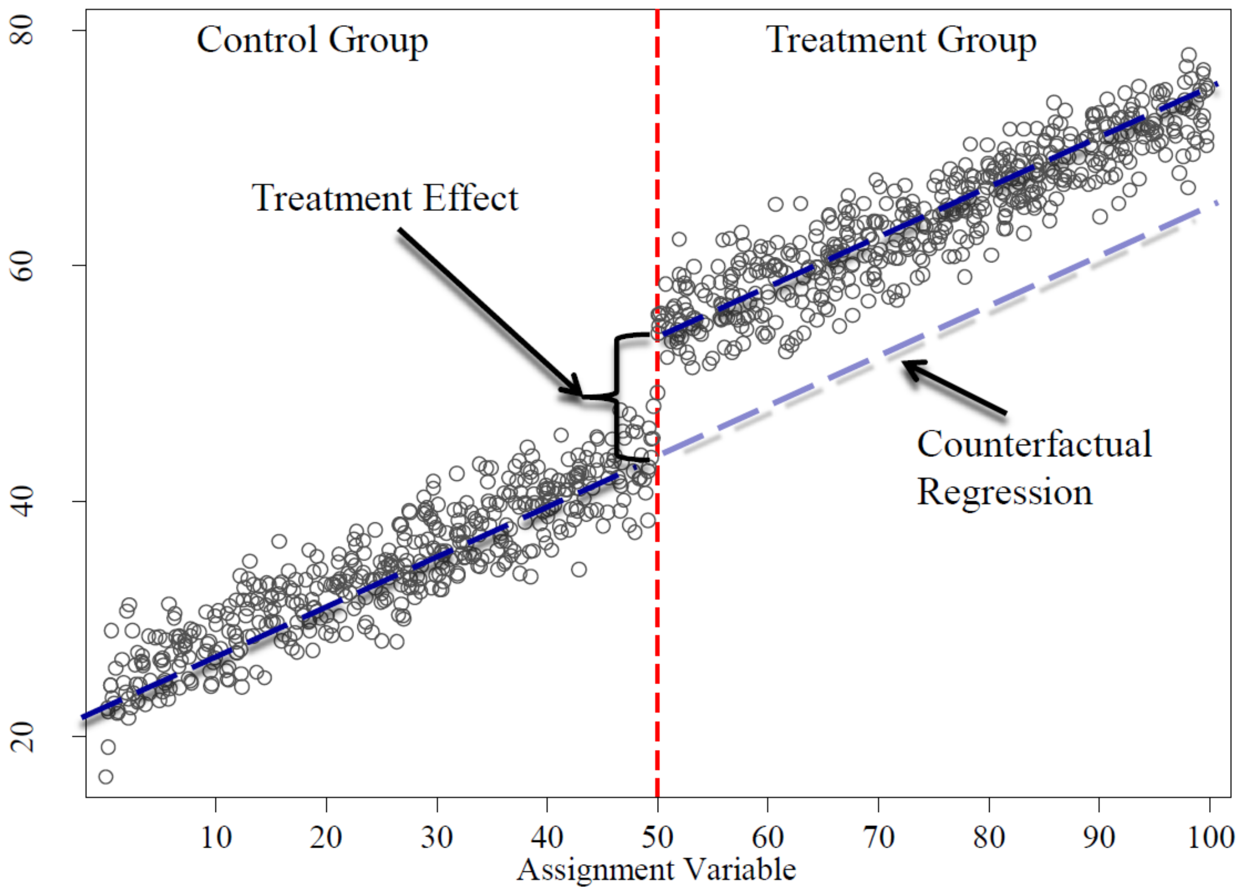
1. Randomized Experiments
2. Natural Experiments / Difference-in-Differences
3. Regression Discontinuity
4. Instrumental Variables

Regression Discontinuity

- **Units above** some sharp (arbitrary) threshold
  - **Treatment** group
- **Units below** the threshold
  - **Control** group
- Treated units **above but close** to threshold
  - Similar to control units **below but close**
    - On observable and unobservable variables
- **(Almost)** "as good as random" assignment to treatment

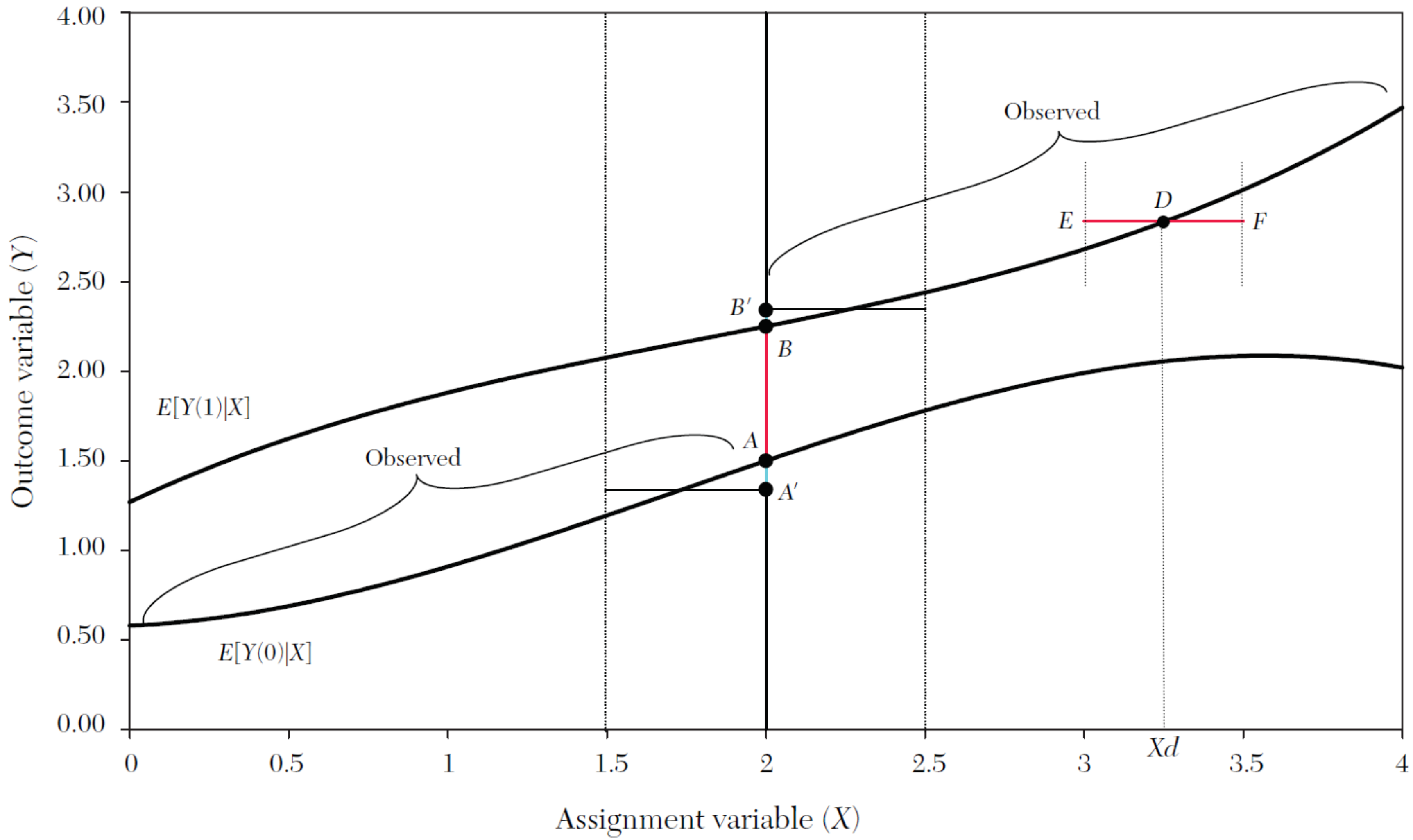
Regression Discontinuity





Tilburg University

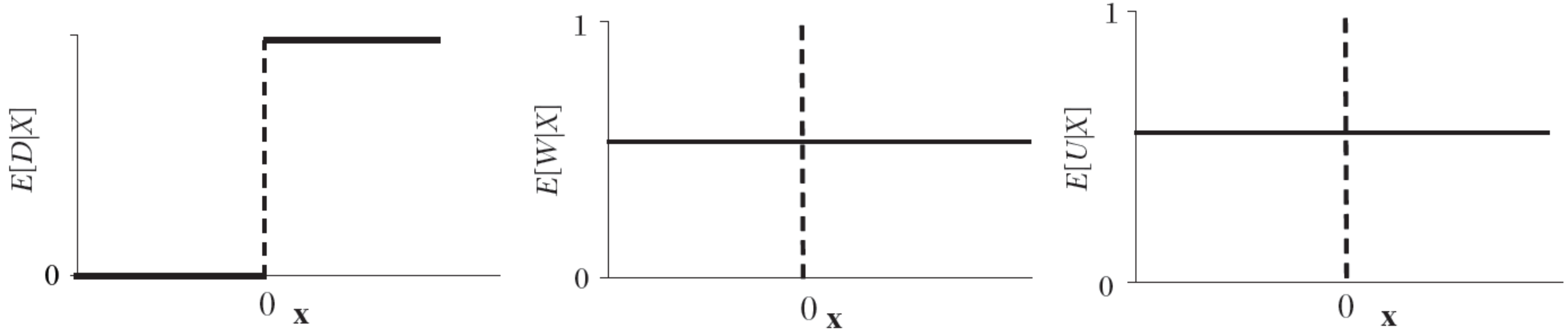
Potential Outcomes in Regression Discontinuity



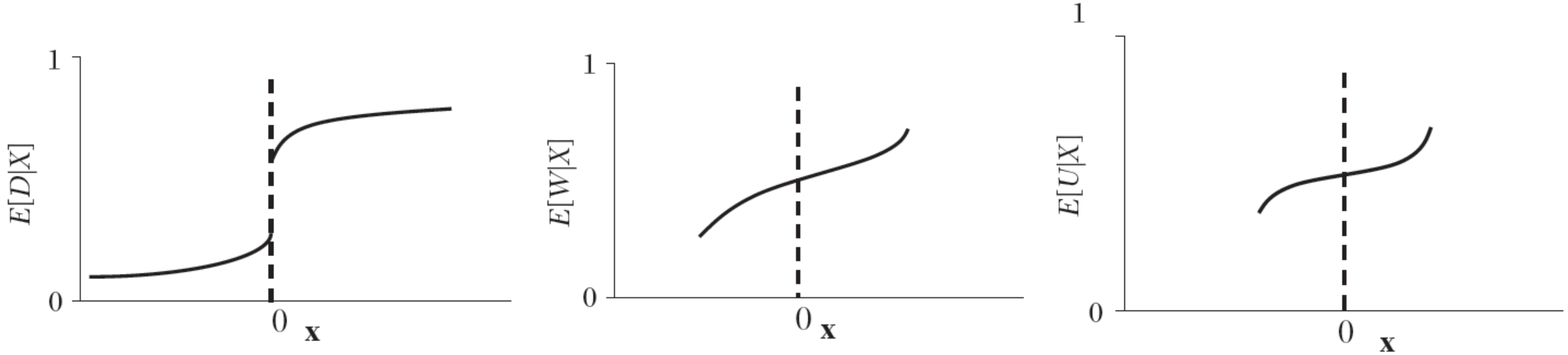
Tilburg University

Regression Discontinuity vs Randomized Experiment

A. Randomized Experiment



B. Regression Discontinuity Design



Regression Discontinuity: Example

- **Research Question:**
  - What is the causal effect of minimum legal drinking age (MLDA) on mortality rates?
- **Mice:**
  - Americans aged 20-22 between 1997 and 2003
  - Death rates (deaths per 100,000 people per year)
- **Dice:**
  - **Age 21 = MLDA in the US**
    - Arbitrary threshold, could be 18 / 16 / 23

Regression Discontinuity: Variables

- The **outcome** variable:
  - Motor vehicle accidents per 100,000 habitants
- **Variable of interest**:
  - Age over 21
- **Covariates**:
  - Age

Regression Discontinuity: Dataset

	mva	agecell	over21
20	30.23012	20.63014	0
21	30.12258	20.71233	0
22	29.74465	20.79452	0
23	30.71792	20.87671	0
24	30.41714	20.9589	0
25	.	20.99999	0
26	.	21	1
27	36.31681	21.0411	1
28	32.5758	21.12329	1
29	33.02229	21.20548	1
30	35.10687	21.28767	1
31	32.3587	21.36986	1
32	32.45526	21.45205	1

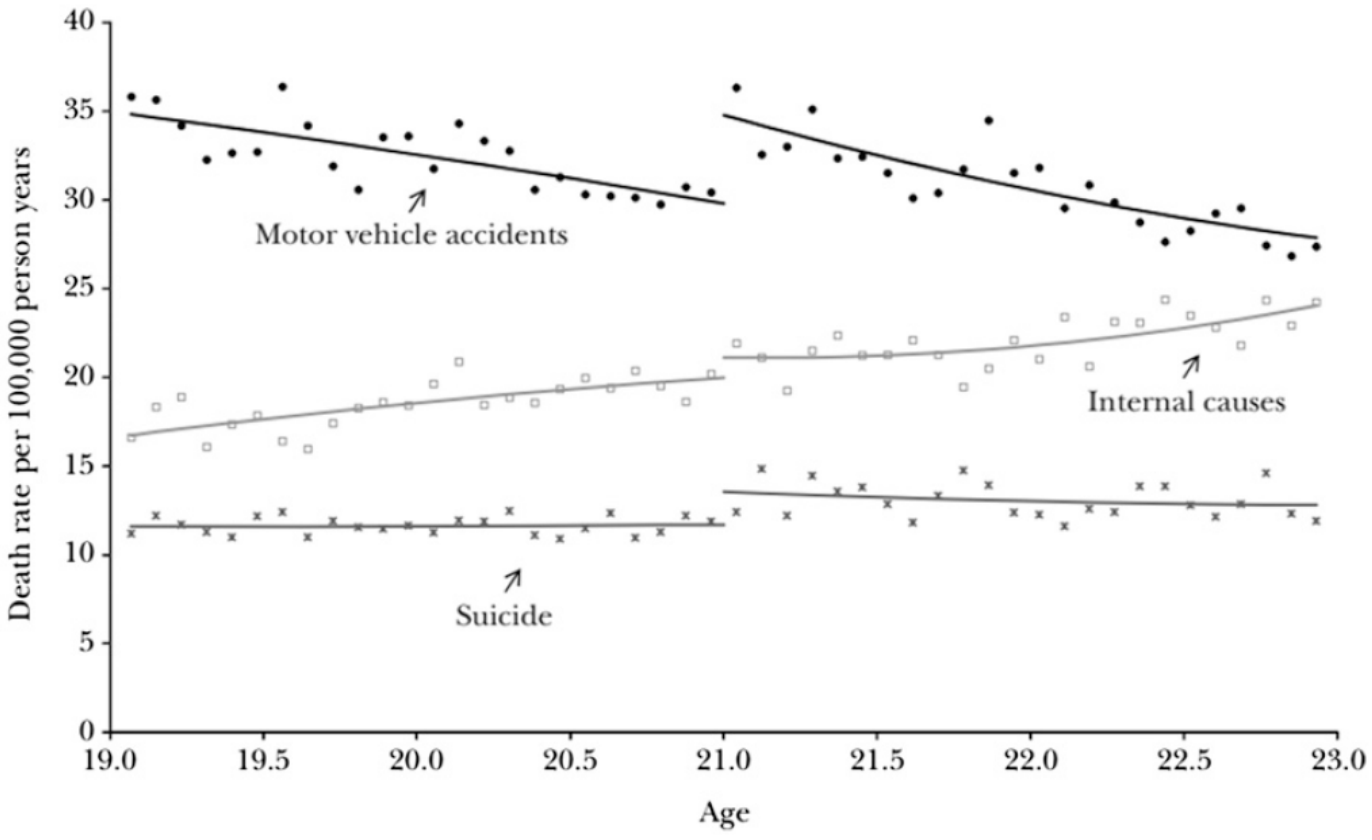
Regression Discontinuity: Counterfactual

- **People aged 21.1** are not so different than
  - **People aged 20.9**
- **Similar individuals** exposed to different treatments
  - Individuals **do not self-select** into treatment
  - Treatment and control group
    - Equal in expectation

Regression Discontinuity: Estimation

```
reg mva over21 agecell, robust
```

Age Profiles for Death Rates in the United States



Dependent variable	Ages 19–22		Ages 20–21	
	(1)	(2)	(3)	(4)
All deaths	7.66 (1.51)	9.55 (1.83)	9.75 (2.06)	9.61 (2.29)
Motor vehicle accidents	4.53 (.72)	4.66 (1.09)	4.76 (1.08)	5.89 (1.33)
Suicide	1.79 (.50)	1.81 (.78)	1.72 (.73)	1.30 (1.14)
Homicide	.10 (.45)	.20 (.50)	.16 (.59)	–.45 (.93)
Other external causes	.84 (.42)	1.80 (.56)	1.41 (.59)	1.63 (.75)
All internal causes	.39 (.54)	1.07 (.80)	1.69 (.74)	1.25 (1.01)
Alcohol-related causes	.44 (.21)	.80 (.32)	.74 (.33)	1.03 (.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Regression Discontinuity: Robustness

- Careful check for covariate **balance**
  - Below vs. above threshold
- **Placebo tests**:
  - Placebo discontinuity at different thresholds
- **Placebo outcomes**:
  - Regress on other covariates
- Bandwidth selection

Empirical Identification Strategies

1. Randomized Experiments
2. Natural Experiments / Difference-in-Differences
3. Regression Discontinuity
4. Instrumental Variables

Instrumental Variables

- What is the causal effect of **education** on **earnings**?
  - Can we estimate the effect with **OLS regression**?
- **Selection bias**
  - Smart people can get **more education**
    - Better exam scores, colleges admit smart people
  - Smart people tend to **earn more money**
    - They can easily learn the professional skills

Instrumental Variables

- How to overcome the **selection bias** in observational studies?
  - 1) Find an exogenous treatment
  - 2) Find an exogenous **instrument**
- What is an **instrumental variable**?
  - **Exogenously** assigned
  - Affects the outcome variable **only through treatment**
    - No direct effect

Instrumental Variables: Example

- **Research Question:**
  - What is the causal effect of education on earnings?
- **Mice:**
  - Americans born in 1930s-1940s
  - Weekly earnings
- **Dice:**
  - **Instrument variable:** Quarter of birth
    - Born in December vs born in January

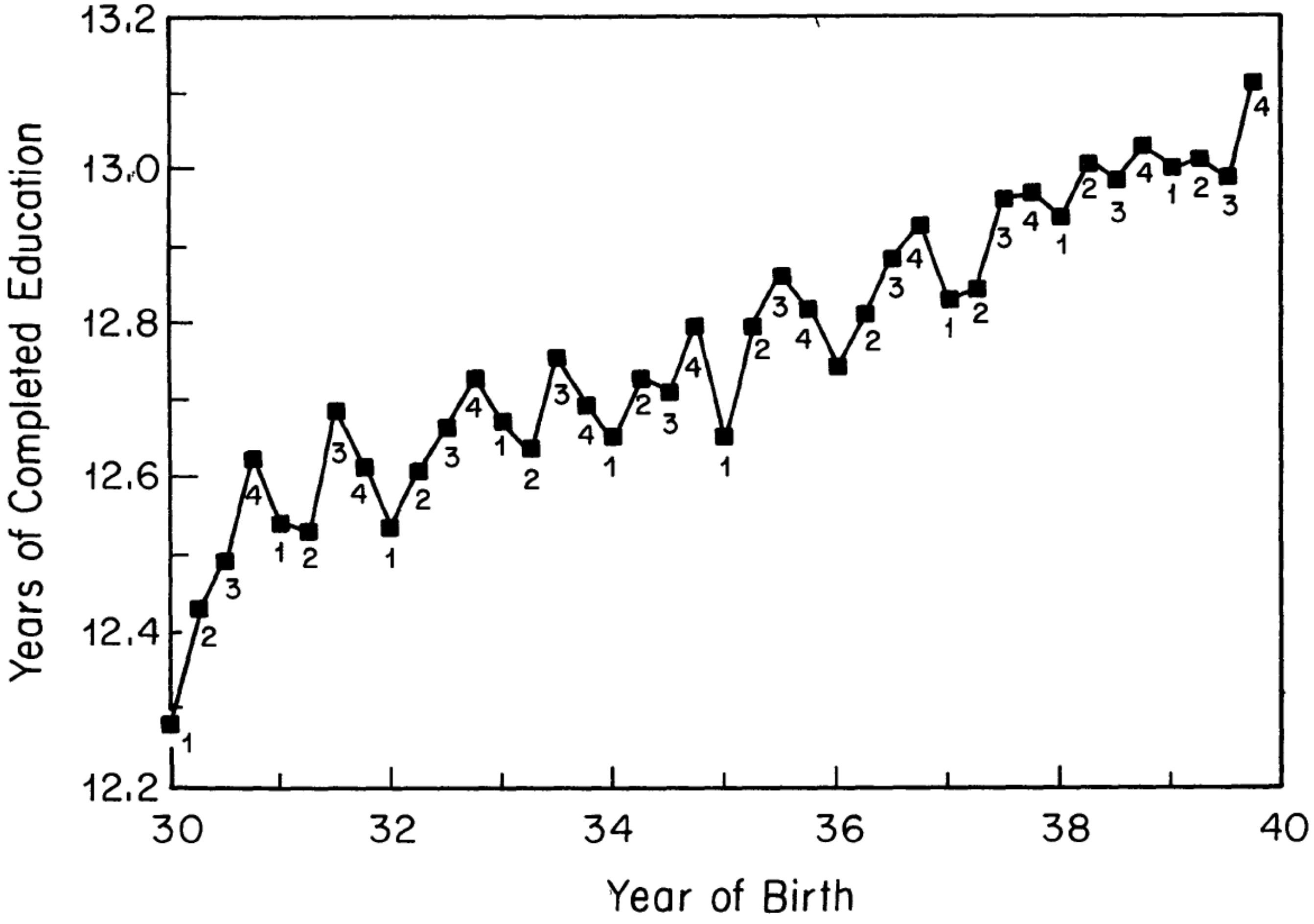
Instrumental Variables: Variables

- The **outcome** variable:
  - Weekly earnings
- **Variable of interest**:
  - Education
- **Instrument**:
  - Quarter of birth

Instrumental Variables: Instrument

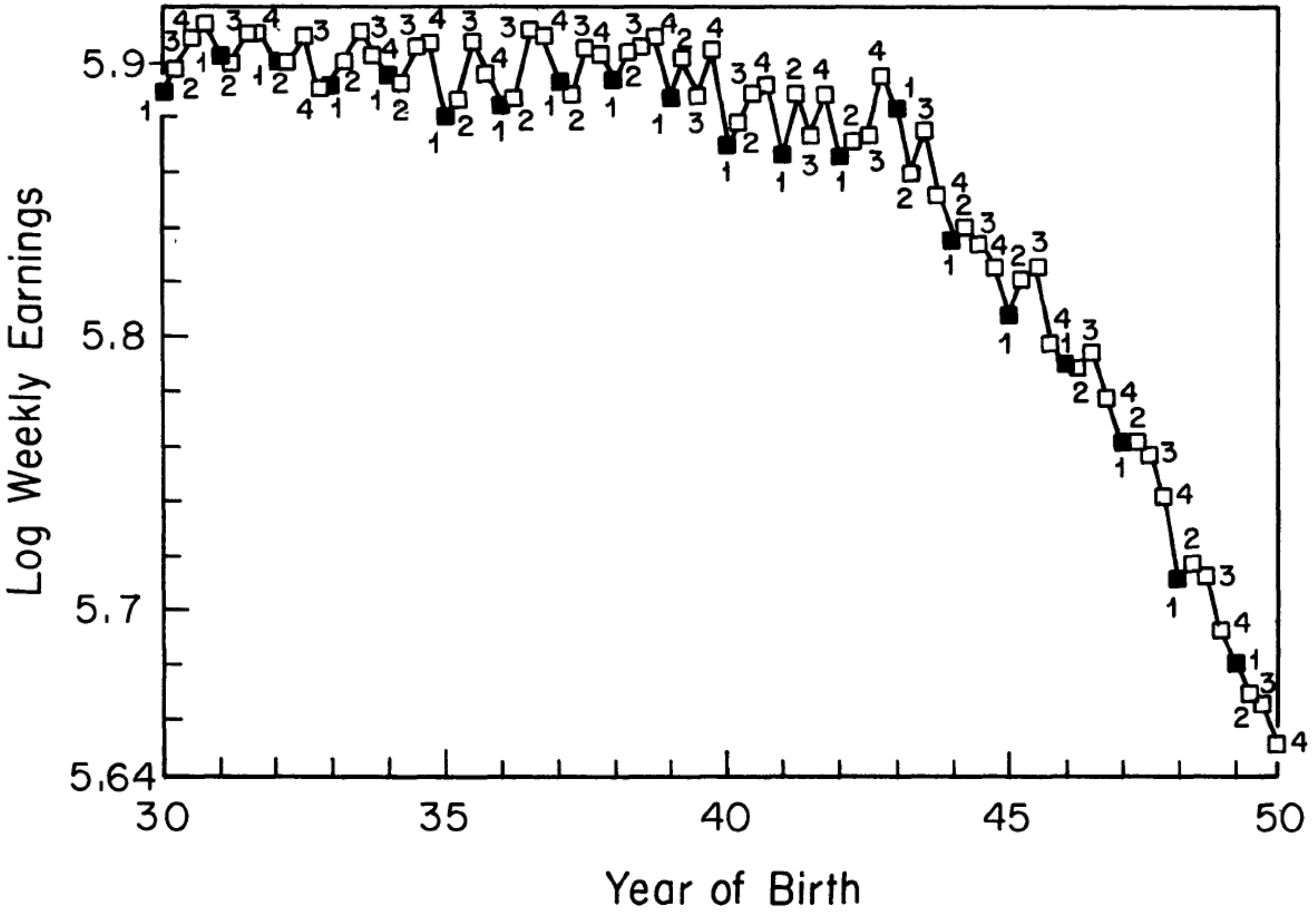
- Why **quarter of birth**?
  - Children start kindergarten education in the year they turn 5
  - Rick (born in Dec 1st, 1930) and Morty (born in Jan 1st 1930)
    - Both start kindergarten in September 1935
    - Rick (4 years and 9 months old) vs Morty (5 years and 8 months old)
- **Compulsory schooling** is until the age of 16
  - Assume Rick & Morty drop school when they turn 16
    - Rick has 12~ years of education
    - Morty has 11~ years of education

Instrumental Variables: Exogenously Assigned



Tilburg University

Instrumental Variables: No Direct Effect



Tilburg University

Instrumental Variables: Dataset

	lnw	s	yob	qob
1	5.790019	12	30	1
2	5.952494	11	30	1
3	5.315949	12	30	1
4	5.595926	12	30	1
5	6.068915	12	30	1
6	5.793871	11	30	1
7	6.389161	11	30	1
8	6.047781	12	30	1
9	5.354861	11	30	1
10	5.259597	7	30	1
11	5.239404	10	30	1
12	5.874553	12	30	1
13	6.001272	14	30	1
14	5.508173	12	30	1
15	5.866414	16	30	1
16	5.729413	12	30	1
17	5.729413	16	30	1
18	5.809437	8	30	1
19	6.657937	16	30	1

Instrumental Variables: Estimation

$\boxed{\textbf{ivregress 2sls } \text{lnw } (s = q4), \textbf{ robust}}$

	Born in quarters 1–3	Born in quarter 4	Difference
Log weekly wage	5.8983	5.9051	.0068
			(.0027)
Years of education	12.7473	12.8394	.0921
			(.0132)
IV estimate of the returns to schooling			.074
			(.028)

Instrumental Variables: Robustness

- First stage **F-statistic**:
  - Must be higher than 10
  - Strong instrument
- Finding a **good instrument** is difficult

Empirical Identification Strategies

1. Randomized Experiments
2. Natural Experiments / Difference-in-Differences
3. Regression Discontinuity
4. Instrumental Variables

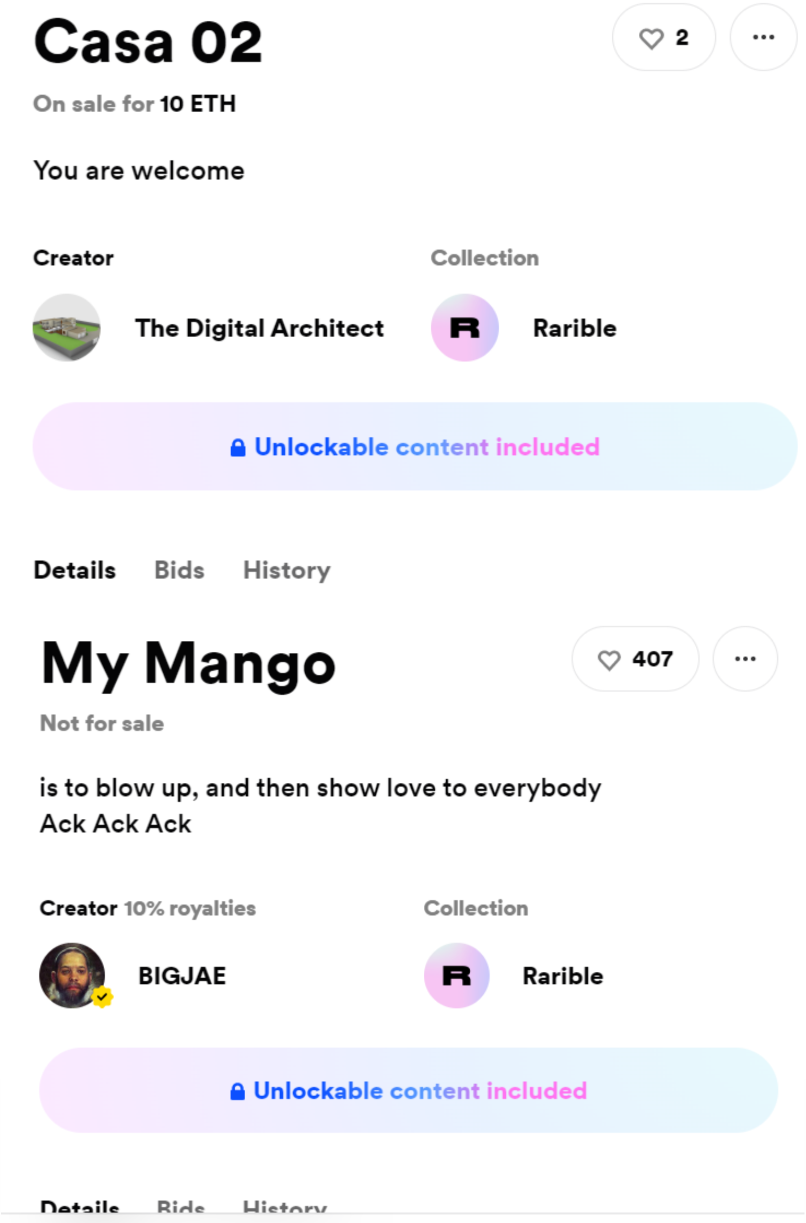
Sample Theses, Suggestions & Data Sources

Sample Thesis – M. Abdelkaui (Spring 2021)

- Panel data from Vinted
  - 8,789 sellers * 4 months = 35,156 observations
- Impact of exposing location on **star ratings**
- **Difference-in-differences**
  - **Treatment:** Users hide their location
  - **Control:** Users expose their location

Sample Thesis – T. v. d. Heuvel (Spring 2021)

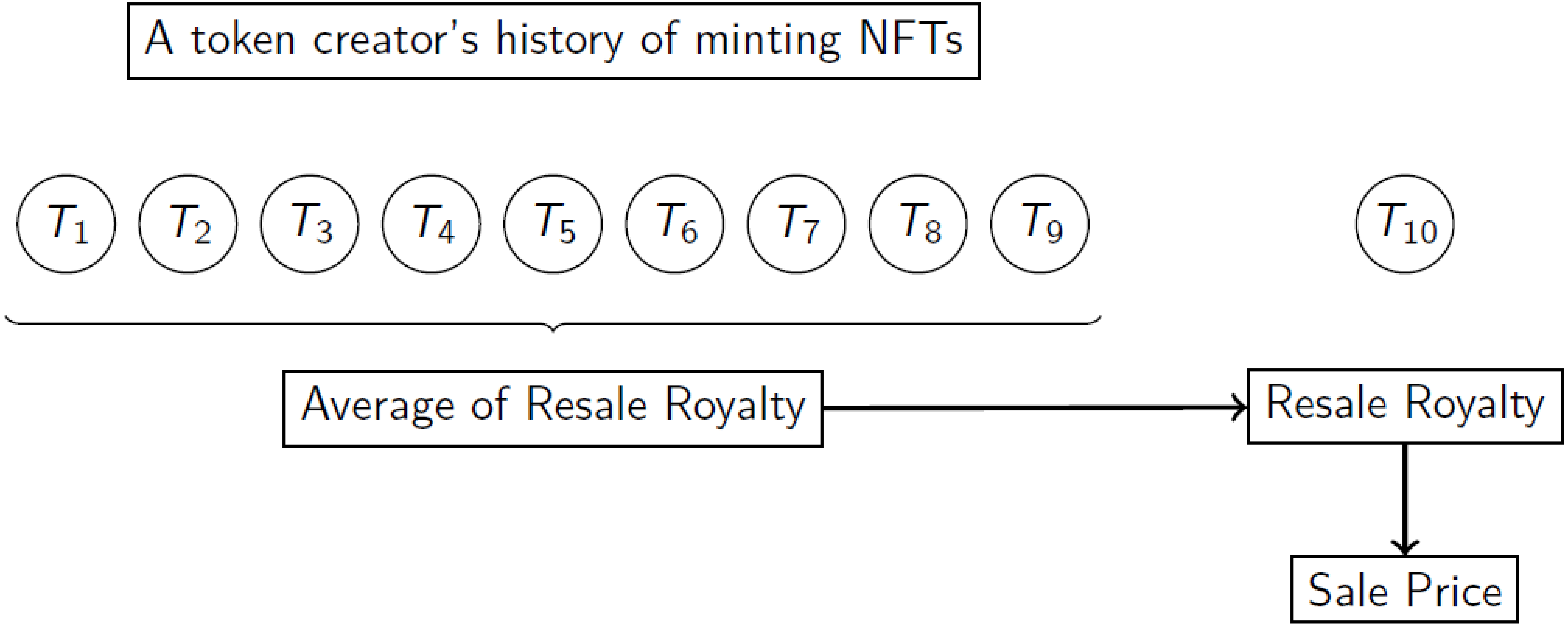
- Panel data from Rarible
  - 11 months (May 20 – April 21)
  - 16,348 token sale observations
- Impact of **resale royalty**
  - on **token sale price**
- **Accepted** at the most prestigieus IS conferences
  - **WISE 2021** (Austin, TX)
  - **CIST 2021** (Los Angeles, CA)



Tilburg University

Sample Thesis – T. v. d. Heuvel (Spring 2021)

- Instrumental variables estimation
  - **Instrument:** Historical royalty behavior of creators



Suggestions for Data-driven Thesis

- Time management
  - ~4 months
  - **Start early**
- Dataset
  - Publicly available databases, APIs
  - Ask your **"ideal"** advisor **for help**
- Math / **code is easy**
  - Design / **identification is difficult**

Thesis with me

- PhD in **Management Science (Information Systems)** - 2020
  - Jindal School of Management, The University of Texas at Dallas
- Research interests
  - **Methods: Econometrics**, Machine Learning, Game Theory
  - **Topics: FinTech**, Platform Strategy, Sharing Economy, Online Marketplaces
- If you want to write a data-driven thesis with me
  - Send me an e-mail **as early as possible**
  - **m.m.tunc@tilburguniversity.edu**

Where to find datasets?

- **Kaggle:** https://www.kaggle.com/datasets
- **Awesome Public Datasets:**
  - https://github.com/awesomedata/awesome-public-datasets
- **Google Cloud Datasets:**
  - https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset
- **EU Open Data**: https://data.europa.eu/en
- **Google Research Datasets**: https://research.google/tools/dataset/
- **Some others:**
  - https://public.opendatasoft.com/
  - https://flowingdata.com/
  - https://data.mendeley.com/
  - https://academictorrents.com/browse.php?cat=6
  - https://knoema.com/atlas/sources

Sample Theses & Data Sources on Canvas

- **Sample Thesis** by M. Abdelkaui
  - Click HERE
- **Sample Thesis** by T. v. d. Heuvel
  - Click HERE
- **Data Sources**: Economics of Digitization
  - Click HERE

Q & A

- Who has any **comments**, **inputs**, or **questions**?