

Please do not distribute without permission.

---

# 定量社会科学的因果推断

Causal Inference in Quantitative Social Sciences

---

江 艇

中国人民大学经济学院

Last updated: April 18, 2020

# Lecture 1 导论

## 2019 年《美国经济评论》的发表情况

	总篇数	部分占比	总占比
发表总文章数	127	-	-
发表长文章数	117	-	100%
非实证	37	-	32%
实证	80	-	68%
非因果推断	14	17.5%	12%
因果推断	66	82.5%	56%
实验数据	13	20%	11%
观测数据	53	80%	45%
描述性	2	4%	2%
结构式	19	36%	16%
约简式	32	60%	27%
OLS	7	22%	6%
IV	8	25%	7%
RD	4	12.5%	3%
DID	13	40.5%	11%

## 来自经济学内部的反思

- Peter Klenow:

$$\frac{\partial \text{Welfare}}{\partial \text{Research}} = \frac{\partial \text{Welfare}}{\partial \text{Knowledge}} \cdot \frac{\partial \text{Knowledge}}{\partial \text{Research}}$$

Micro	$> 0$	High	Low
Macro	$\approx 0 ?$	Low	High

- Andrew Gelman:

Randomized experiments give you accurate estimates of things you don't care about; Observational studies give you biased estimates of things that actually matter.

- James Heckman (2005):

In some quarters of our profession, the level of discussion has sunk to the level of a *New Yorker* article: coffee-table articles about “cute” topics, papers using “clever” instruments. The authors of these papers are usually unclear about the economic questions they address, the data used to support their conclusions and the econometrics used to justify their estimates. This is a sad development that I hope is a passing fad. Most of this work is without substance, but it makes a short-lived splash and it’s easy to do. Many young economists are going for the cute and the clever at the expense of working on hard and important foundational problems.

- Deaton and Cartwright (2010):

RCTs do indeed require minimal assumptions and can operate with little prior knowledge. This is an advantage when persuading distrustful audiences, but it is a disadvantage for cumulative scientific progress, where prior knowledge should be built upon, not discarded. RCTs can play a role in building scientific knowledge and useful predictions but they can only do so as part of a cumulative program, combining with other methods, including conceptual and theoretical development, to discover not “what works”, but “why things work”.

- Rust (2016):

[L]eading economics journals err by imposing an unrealistic burden of proof on empirical work: there is an obsession with establishing causal relationships that must be proven beyond the shadow of a doubt. ... [T]he overabundance of econometric theory has not paid off in terms of empirical knowledge, and may paradoxically hinder empirical work by obligating empirical researchers to employ the latest methods that are often difficult to understand and use and fail to address the problems that researchers actually confront. ... [T]here should be more room for descriptive empirical work in our journals. It should not be necessary to establish a causal mechanism or a non-parametrically identified structural model that provides an unambiguous explanation of empirical phenomena as a litmus test for publication.

## 《经济学家》杂志的访谈

- 1988.

“Economics in America’s Cambridge reached its peak of influence in the 1960s; in the 1970s it fell quiet; now it is flourishing again ... Economics has lost patience with the paradigm of perfect competition ... Out will go the presumption that markets clear smoothly ... In will stay the idea that agents are not stupid ...”

- 1998.

“Unlike the stars of the 1980s, today’s impressive young academics are using the tools of economics in fields on or beyond the traditional borders of their discipline ... [They] tend to eschew the big traditional themes of economics ... The erosion of traditional barriers within economics and the increased meshing with other disciplines will continue. A decade hence it will not only be hard to distinguish empirical economists from theorists; it may also be hard to disentangle economics from other strands of social science ...”



- 2008.

“Today’s economists show no great attachment to the rational model of behaviour that guided Mr Becker. Economic theory has become so eclectic that ingenious researchers can usually cook up a plausible model to explain whatever empirical results they find interesting. Economics is now defined neither by its subject matter nor by its method ... Economists still share a taste for the Greek alphabet: they like to provide formal, algebraic accounts of the behaviour they explain. And they pride themselves on the sophistication of their investigative methods. They are usually better at teasing confessions out of data than their rivals in other social sciences. What defines economics? Economics is what economists do — the best of them, anyway.”

- 2018.

“The critics lodged three related objections. The first was a neglect of theory: the new empiricists were not always particularly interested in testing formal models of how the world worked ... The second objection was a lack of seriousness. ‘Freakonomics’ had encouraged an emerging generation of economists to trivialize their subject ... The new empiricists were also accused of looking for keys under lampposts. Some showed more allegiance to their preferred investigative tools than to the subject or question under investigation.”

“The 2018 cohort’s combination of clever methods and dogged snuffling out of data comes along with a rejection of some of the more frolicsome manifestations of earlier new empiricists. Many of them display an admirable millennial earnestness. They are mostly tackling subjects that are both in line with long-standing economic concerns and of grave public importance ... Because they want to change the world, not just delight in its perversity, many of these economists engage closely with policy ... Intriguingly, this concern for real-world outcomes is pushing some of these young economists back towards theory.”

## 这门课的定位

- 与其说是一门方法课，不如说是一门方法论课。
- 只能告诉你什么是对的，听不听是你的事。
- 很多时候连什么是对的都没法告诉，因为没有不随研究情境而变的标准流程。
- 故事永远比技术重要。从拙劣到平庸的门槛是技术，从平庸到卓越的门槛是故事。

## 这门课的口号

入门须正  
取法须高  
立志须远

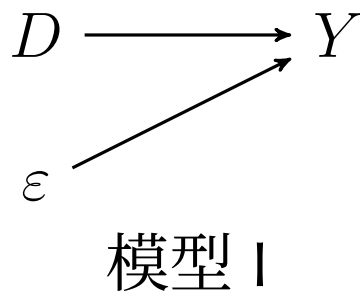
## 何为因果推断？

- 探究现象之间的因果联系。因果性，相对于描述性、相关性而言。
  - 医学的因果问题：新药疗效
  - 经济学的因果问题：扶贫项目的成效
  - 社会学的因果问题：父亲的职业如何影响子女的职业选择
  - 政治学的因果问题：识字率如何影响民主参与程度
- 用  $Y$  表示我们感兴趣的结果 (outcome)，或反应 (response)。
- 用  $D$  表示我们感兴趣的原因 (cause)，或处理 (treatment)、干预 (intervention)。

- 对原因最简单的刻画是非有即无、非是即否、非此即彼，用二元变量 (binary variable) 表示：

$$D = \begin{cases} 1 & \text{实施一项处理（进入处理组、实验组）} \\ 0 & \text{不实施一项处理（进入控制组、对照组）} \end{cases}$$

- 用  $\varepsilon$  表示影响结果的其它因素。
- 我们感兴趣的因果关系可以用如下的基本因果模型来刻画：



- 这个基本因果模型对整个定义良好的总体中的每个观测单位都成立，因此称之为总体因果关系 (population causal relationship)，研究者只观测到这个总体的某个样本。可以表示为线性模型：

$$Y_i = \beta_1 D_i + \varepsilon'_i$$

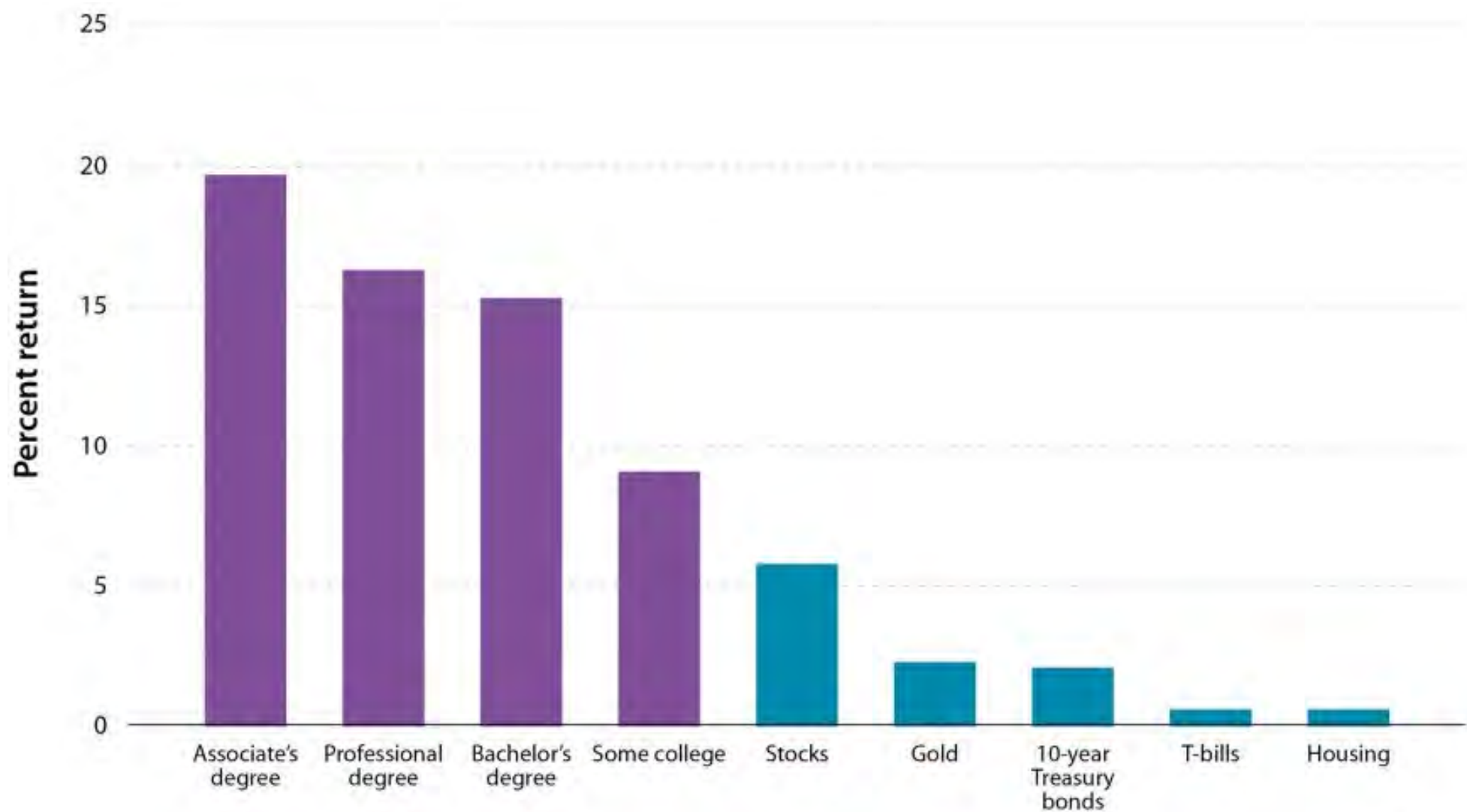
做一下技术处理：定义  $\beta_0 \equiv \mathbb{E}(\varepsilon'_i)$ ，定义  $\varepsilon_i \equiv \varepsilon'_i - \beta_0$ ，则

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \mathbb{E}(\varepsilon_i) = 0$$

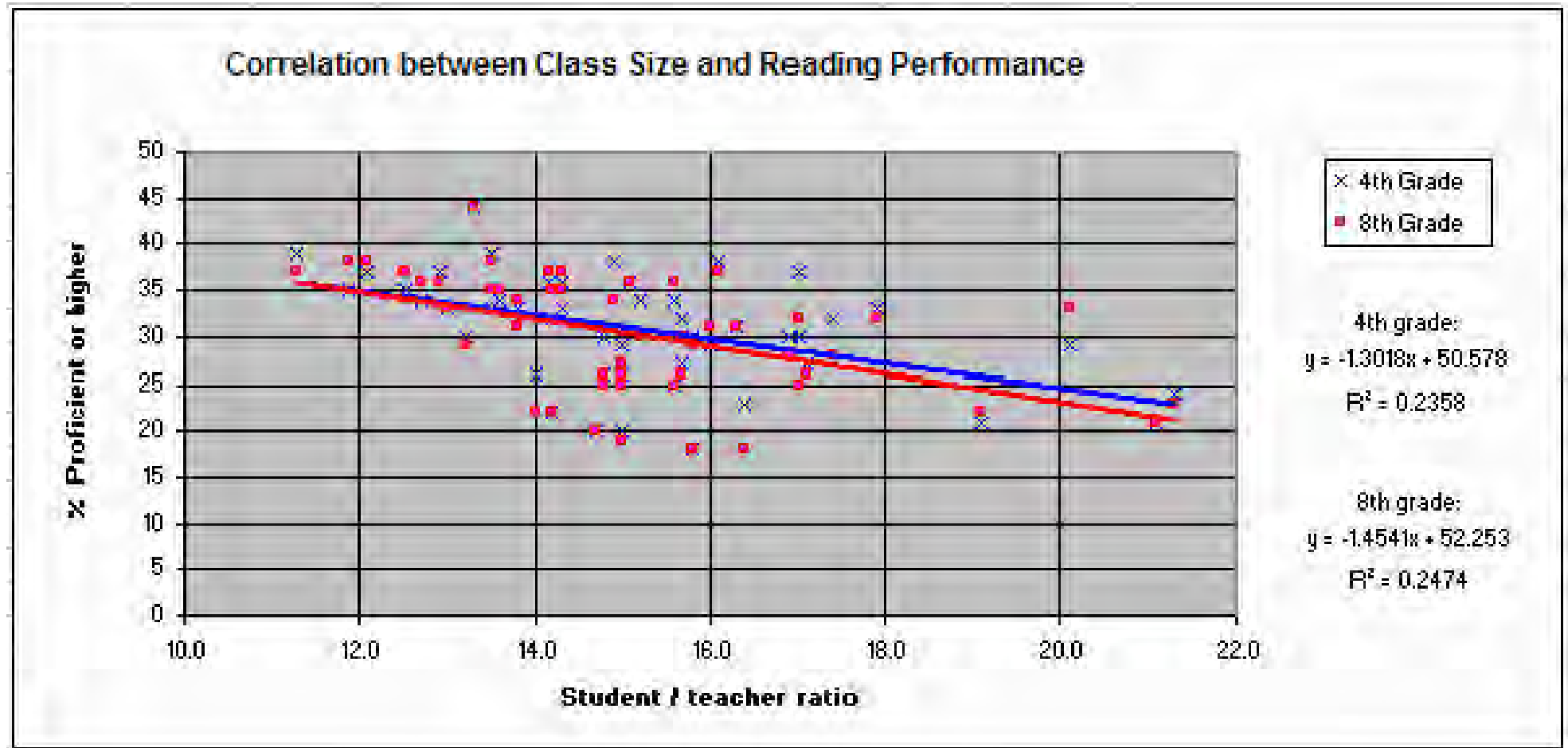
- $\beta_1$  是我们重点关注的（未知的）总体因果参数 (population causal parameter)，其含义是，保持  $\varepsilon$  不变，一项处理的实施，导致结果变化  $\beta_1$ ，称之为因果效应 (causal effect) 或处理效应 (treatment effect)。
- 当  $D$  为二元变量时，这一线性模型的限制性假设在于
  - 每个观测单位的因果效应都相同： $\beta_{1i} \equiv \beta_1$ 。
  - $D$  和  $\varepsilon$  是可分的。
- 我们来看四个例子，刻画了四组相关性事实，体会它们试图讲述什么因果故事？ $Y, D, \varepsilon$  分别是什么？



## 示例 1. 教育回报率

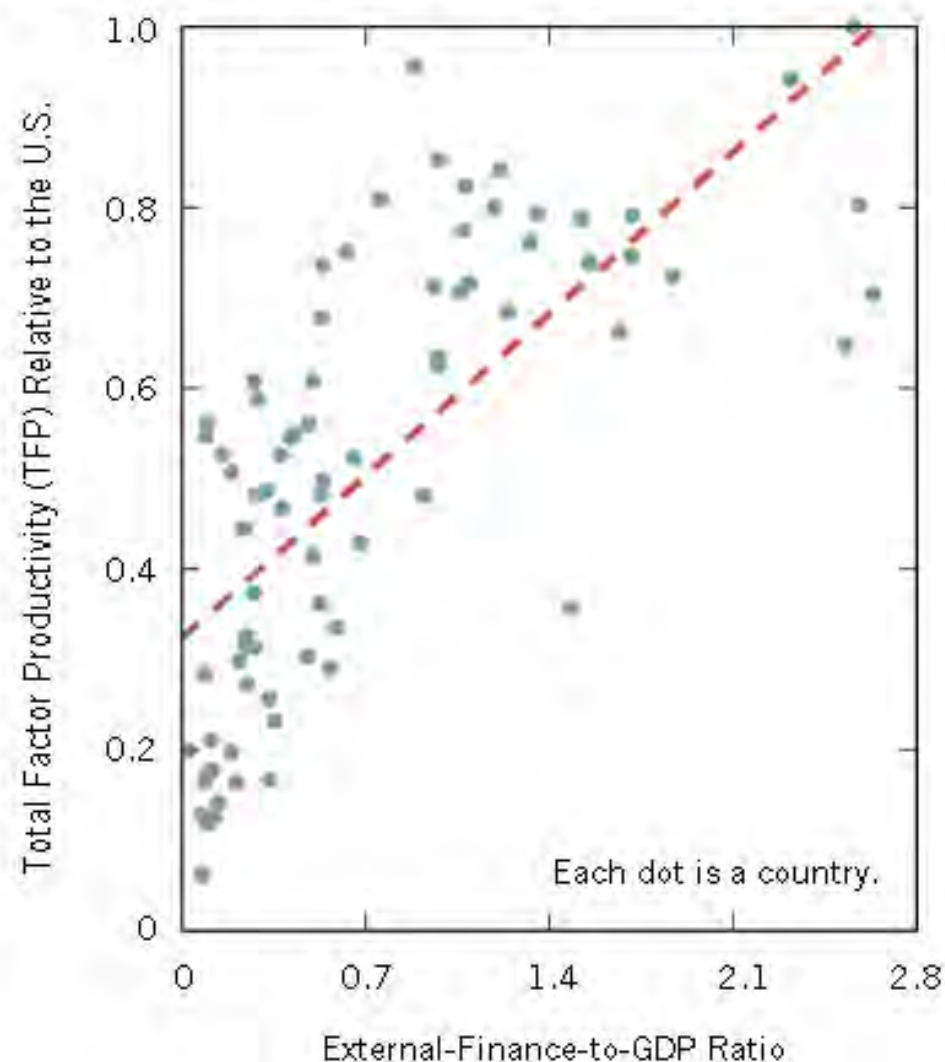
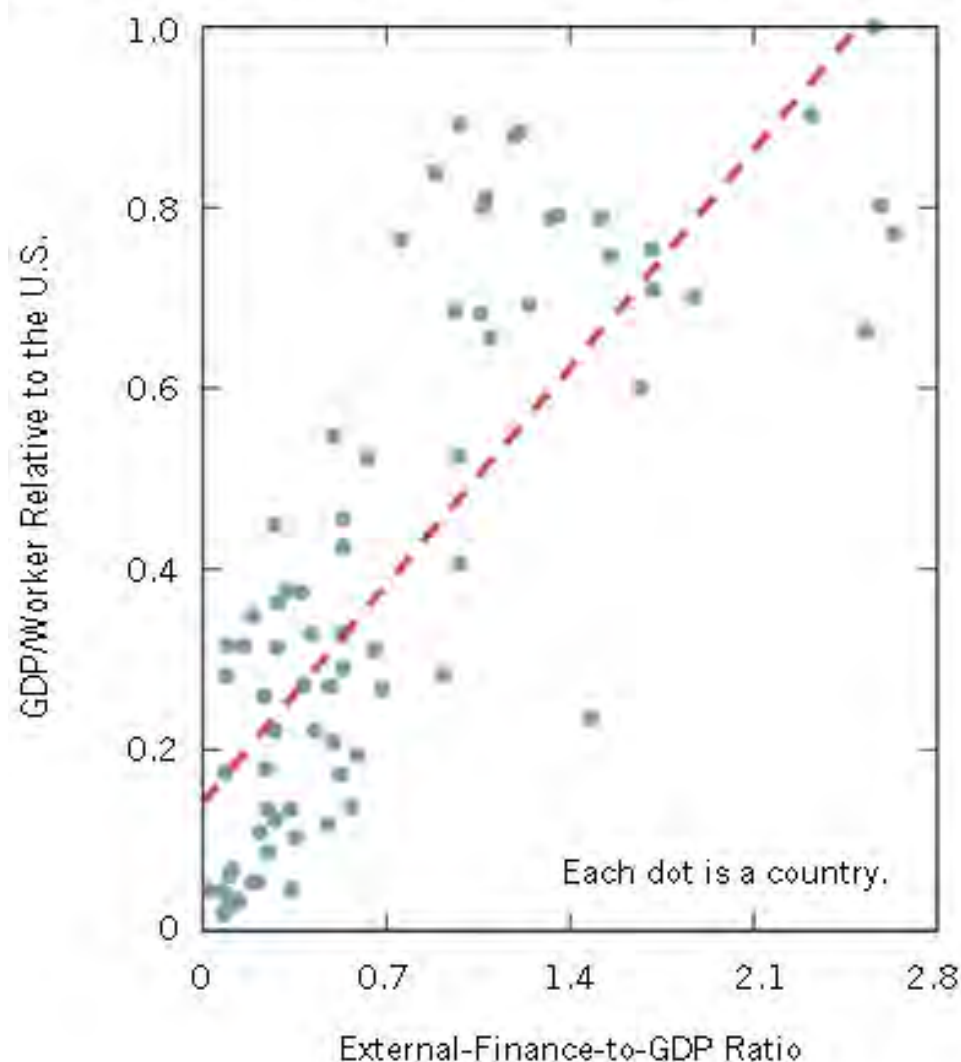


## 示例 2. 班级规模与教育产出



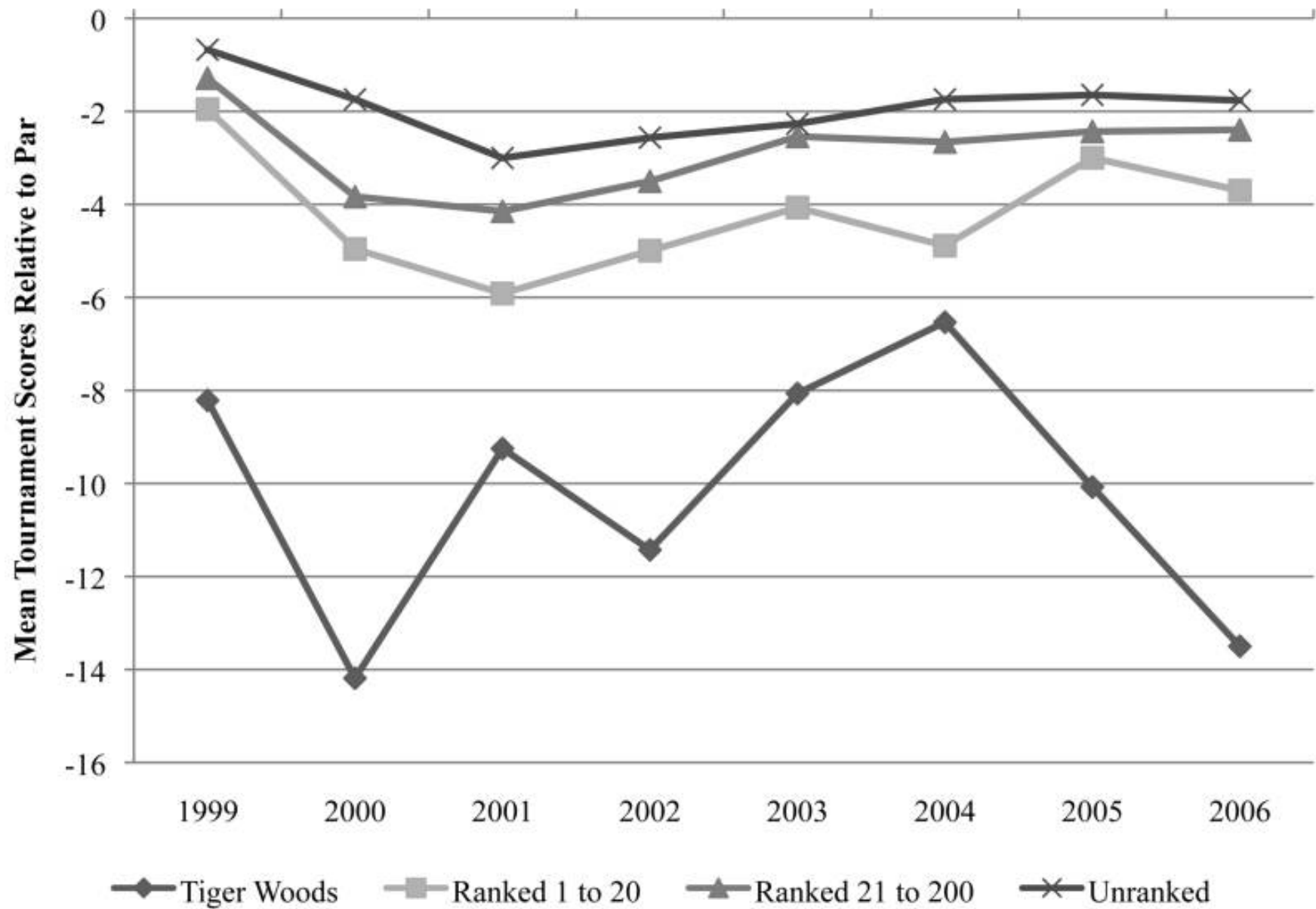
### 示例 3. 金融发展与经济增长

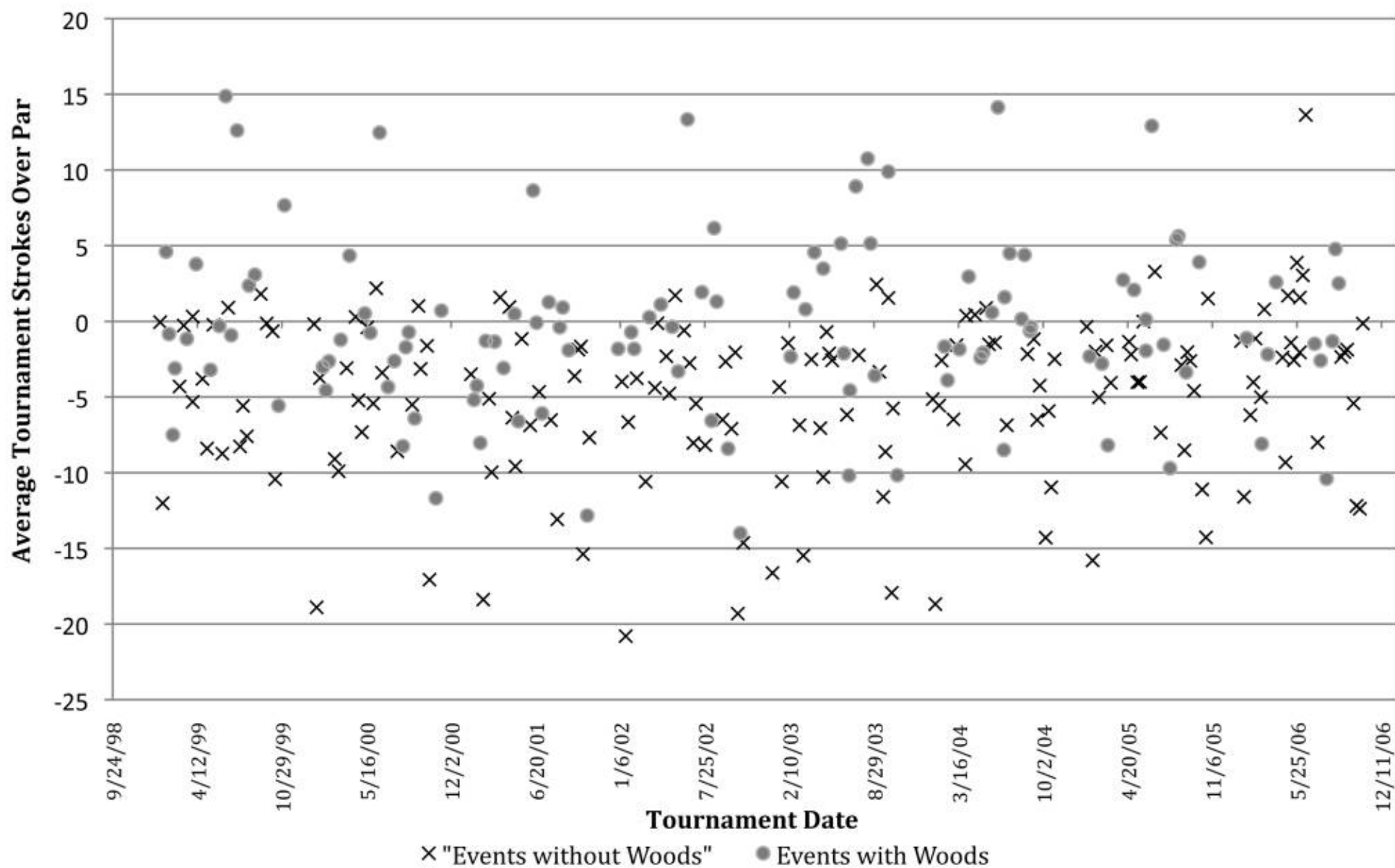
Relationship between Financial Development and Economic Development



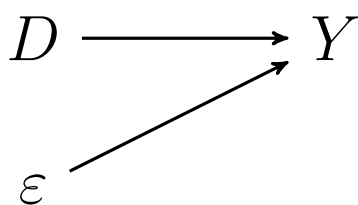
#### 示例 4. 超级明星效应 (Brown, 2011, *JPE*).

- 生活常识：竞争是一种重要的激励机制。考核相对绩效的锦标赛机制要想发挥作用，有一项重要前提——竞争者的能力必须相对均衡。存在“超级明星”时，锦标赛机制反而会产生负面效果。
- 研究情境：“老虎”伍兹，史上最伟大的高尔夫球手。1975 年出生，1996 年 20 岁时成为职业球手，职业生涯未满一年即跃居世界排名第一，在 1999 年 8 月至 2004 年 9 月以及 2005 年 6 月至 2010 年 10 月分别连续 264 周和 281 周保持世界排名第一。

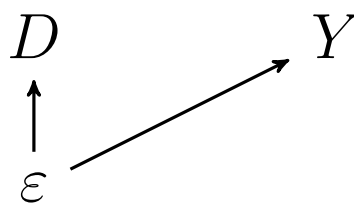




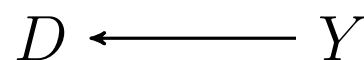
- 任何因果推断问题都包含两部分：
  - 因果识别 (causal identification)**：如果拥有整个总体，是否能够确定总体因果参数？这是社会科学理论的任务。
  - 统计推断 (statistical inference)**：如何从样本数据获取关于总体因果参数的信息？这是统计学的任务。
- 统计推断致力于发现  $Y$  和  $D$  在样本中的相关性，并由此评估其总体相关性。
- 如果我们感兴趣的（从  $D$  到  $Y$  的）因果关系真的存在，那么  $D$  和  $Y$  之间的相关性必然存在，反之则不然。 $D$  和  $Y$  相关这一事实可能被多个基本因果模型所合理化 (rationalize)：



模型 I

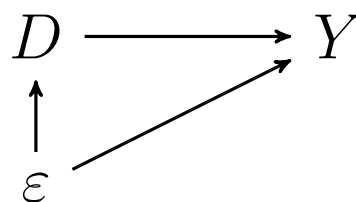


模型 II



模型 III

- 真实世界的数据生成过程 (data generating process) 甚至可能是多个基本因果模型同时作用的结果。

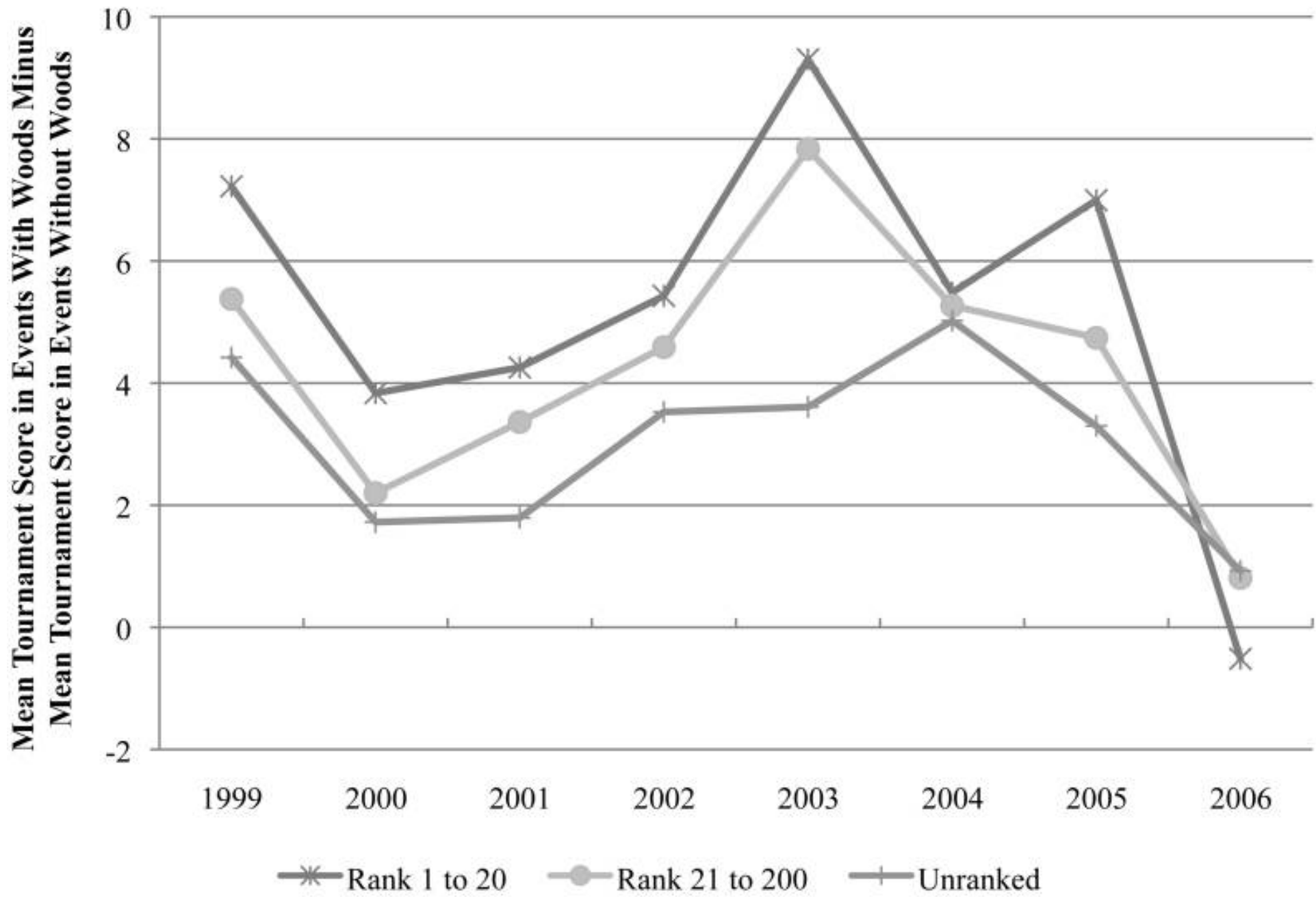


模型 IV

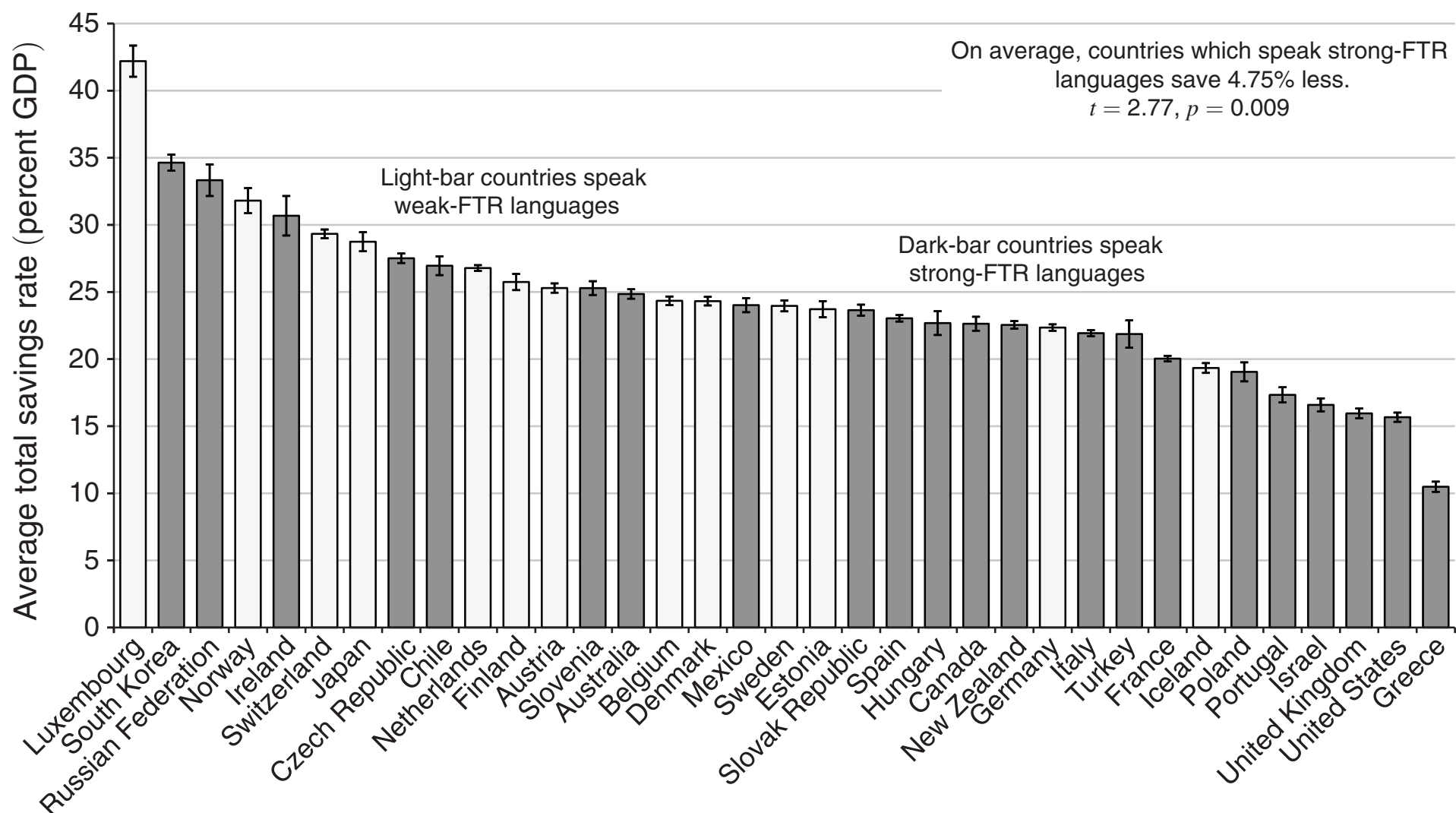
- 若一个因果模型被识别，则意味着不存在其它观测上等价 (observationally equivalent) 的因果模型，我们只能基于社会科学理论对此做出判断。重新审视前述四个例子，除了最显然的因果故事以外，还有没有其它竞争性 (alternative/competitive) 的解释？
- 因此因果识别的基本逻辑是：如果相关性不存在，则因果性不存在；如果相关性存在，且只有一种因果模型可以合理化这种相关性，则这种特定的因果性存在。

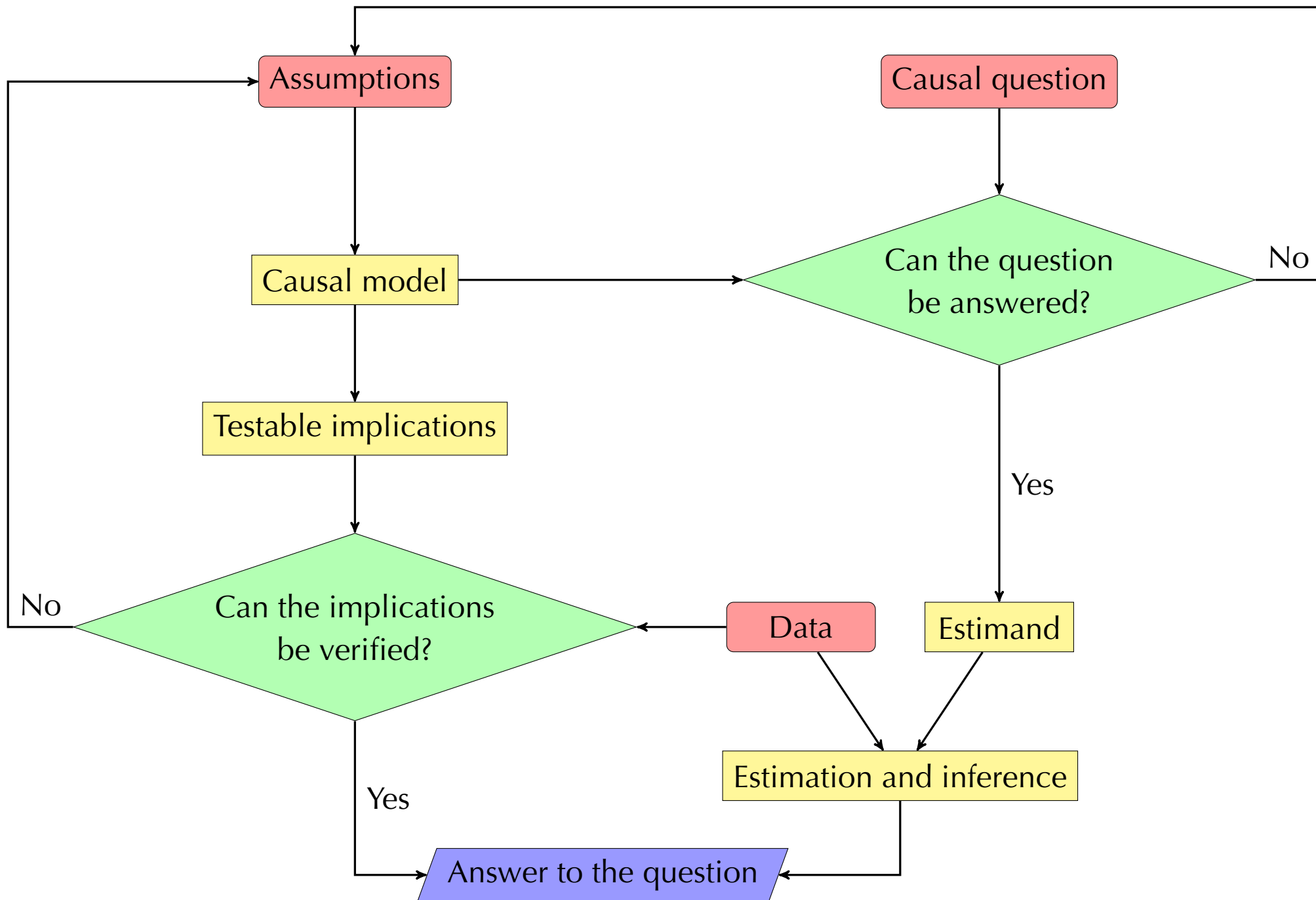


- 如何排除竞争性的因果模型？
  - 通过巧妙的研究设计，不同的方法依赖于不同的识别假设，而不同的研究情境适用不同的识别假设 (make assumptions justifiable)。
  - 挖掘因果模型更丰富的、可验证的相关性含义。论证策略可以概括为：“一根筷子易折断，十根筷子抱成团”。



## 示例 5. 语言与储蓄率 (Chen, 2013, *AER*).





## 随机实验：因果推断的参照系

- 在一项随机实验中，研究者主动介入了数据生成过程，以确保只有模型 I 成为可能，因此随机实验是因果推断的理想情形和参照系，**所有的研究设计都致力于使得研究情境尽量接近于随机实验。**
- 研究者招募一批被试，将其随机划分到处理组和控制组，尽管每位被试的  $\varepsilon_i$  各不相同，但随机分组保证了处理组个体和控制组个体的  $\varepsilon$  大体上保持平衡，因此两组个体结果的平均差异即反映因果效应。

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

$$\mathbb{E}(Y_i | D_i = 1) = \beta_0 + \beta_1 + \mathbb{E}(\varepsilon_i | D_i = 1)$$

$$\mathbb{E}(Y_i | D_i = 0) = \beta_0 + \mathbb{E}(\varepsilon_i | D_i = 0)$$

若

$$\mathbb{E}(\varepsilon_i | D_i = 1) = \mathbb{E}(\varepsilon_i | D_i = 0) \tag{1.1}$$

则

$$\beta_1 = \mathbb{E}(Y_i | D_i = 1) - \mathbb{E}(Y_i | D_i = 0)$$

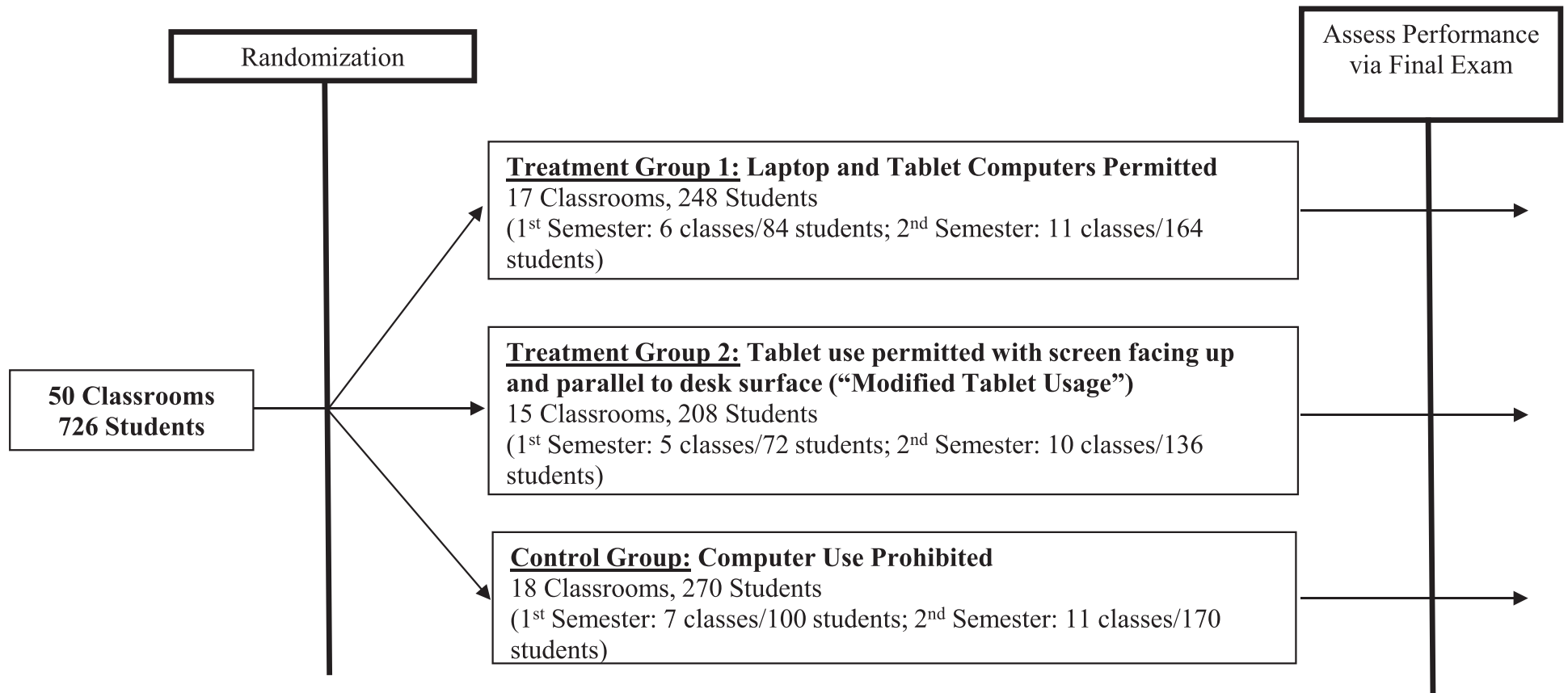
- 假设 (1.1) 还可写作

$$\mathbb{E}(\varepsilon_i | D_i) = \mathbb{E}(\varepsilon_i) = 0$$

称  $\varepsilon_i$  与  $D_i$  **均值独立**, 记作  $\varepsilon_i \perp D_i$ .

- 一般而言,  $\varepsilon_i \perp D_i \Rightarrow \text{Cov}(\varepsilon_i, D_i) = 0$ , 当  $D_i$  为二元变量时, 两者等价。不过, 这两个命题的差异在于数学上, 而不在于社会科学意义上。
- 回忆  $\beta_1$  的含义: “保持  $\varepsilon$  不变,  $D$  由 0 变为 1 所导致的  $Y$  的变化”。在随机实验中,  $\varepsilon$  和  $D$  不相关, 因此处理组与控制组的比较即可反映“保持  $\varepsilon$  不变,  $D$  由 0 变到 1”所导致的变化。

**示例 6.** 使用电脑有损学习成绩 (Carter et al, 2017, *Economics of Education Review*)



	Control (1)	Treatment 1 (laptops/tablets) (2)	Treatment 2 (tablets, face up) (3)	Both treatments vs. control (4)	Treatment 1 vs. control (5)	Treatment 2 vs. control (6)
A. Baseline characteristics						
Female	0.17	0.20	0.19	0.03 (0.03)	0.06 (0.04)	0.00 (0.04)
White	0.64	0.67	0.66	0.02 (0.04)	0.02 (0.04)	0.02 (0.05)
Black	0.11	0.10	0.11	-0.02 (0.03)	-0.02 (0.03)	-0.03 (0.04)
Hispanic	0.13	0.13	0.09	0.00 (0.03)	0.02 (0.03)	-0.03 (0.03)
Age	20.12 [1.06]	20.15 [1.00]	20.15 [0.96]	0.03 (0.08)	0.05 (0.09)	0.06 (0.10)
Prior military service	0.19	0.19	0.16	-0.02 (0.03)	0.00 (0.04)	-0.01 (0.04)
Division I athlete	0.29	0.40	0.35	0.05 (0.04)	0.07* (0.04)	0.04 (0.05)
GPA at baseline	2.87 [0.52]	2.82 [0.54]	2.89 [0.51]	-0.01 (0.04)	-0.05 (0.05)	0.03 (0.05)
Composite ACT	28.78 [3.21]	28.30 [3.46]	28.30 [3.27]	-0.34 (0.26)	-0.37 (0.31)	-0.54 (0.33)
<i>P</i> -Val (Joint $\chi^2$ Test)				0.610	0.532	0.361
B. Observed computer (laptop or tablet) use						
any computer use	0.00	0.81	0.39	0.62*** (0.02)	0.79*** (0.03)	0.40*** (0.04)
Average computer use	0.00	0.57	0.22	0.42*** (0.02)	0.56*** (0.02)	0.24*** (0.03)
Observations	270	248	208	726	518	478



**Table 4**

Unrestricted laptop/tablet classrooms vs. non-computer classrooms.

	(1)	(2)	(3)	(4)
A. Dependent variable: Final exam multiple choice and short answer score				
Computer class	−0.28*** (0.10)	−0.23*** (0.09)	−0.19*** (0.07)	−0.18*** (0.07)
GPA at start of course			1.09*** (0.07)	0.92*** (0.07)
Composite ACT				0.07*** (0.01)
Demographic controls		X	X	X
R <sup>2</sup>	0.08	0.28	0.54	0.57
Robust SE <i>P</i> -Val	0.003	0.007	0.005	0.005
Wild Bootstrap <i>P</i> -Val	0.000	0.000	0.000	0.000
B. Dependent variable: Final exam multiple choice score				
Computer class	−0.25*** (0.10)	−0.20** (0.009)	−0.16** (0.07)	−0.15** (0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R <sup>2</sup>	0.08	0.27	0.48	0.50
Robust SE <i>P</i> -Val	0.009	0.023	0.025	0.029
Wild Bootstrap <i>P</i> -Val	0.000	0.000	0.000	0.000
C. Dependent variable: Final exam short answer score				
Computer class	−0.25*** (0.09)	−0.21** (0.09)	−0.18** (0.07)	−0.17** (0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R <sup>2</sup>	0.08	0.21	0.44	0.46
Robust SE <i>P</i> -Val	0.008	0.016	0.017	0.019
Wild Bootstrap <i>P</i> -Val	0.008	0.020	0.022	0.028
D. Dependent variable: Final exam essay questions score				
Computer class	−0.03 (0.08)	−0.01 (0.08)	0.02 (0.07)	0.02 (0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R <sup>2</sup>	0.32	0.37	0.50	0.51
Robust SE <i>P</i> -Val	0.705	0.912	0.801	0.755
Wild Bootstrap <i>P</i> -Val	0.549	0.811	0.721	0.641

**Table 5**

Modified-tablet classrooms vs. non-computer classrooms.

	(1)	(2)	(3)	(4)
A. Dependent variable: Final exam multiple choice and short answer score				
Computer class	−0.17* (0.10)	−0.18** (0.09)	−0.20*** (0.07)	−0.17** (0.07)
GPA at start of course			1.12*** (0.07)	1.01*** (0.08)
Composite ACT				0.05*** (0.01)
Demographic controls		X	X	X
R <sup>2</sup>	0.07	0.26	0.53	0.54
Robust SE <i>P</i> -Val	0.087	0.050	0.007	0.019
Wild Bootstrap <i>P</i> -Val	0.000	0.000	0.000	0.000
B. Dependent variable: Final exam multiple choice score				
Computer class	−0.15 (0.10)	−0.15* (0.09)	−0.17** (0.08)	−0.14* (0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R <sup>2</sup>	0.07	0.26	0.48	0.49
Robust SE <i>P</i> -Val	0.141	0.100	0.027	0.057
Wild Bootstrap <i>P</i> -Val	0.000	0.000	0.000	0.000
C. Dependent variable: Final exam short answer score				
Computer class	−0.21** (0.10)	−0.22** (0.09)	−0.24*** (0.08)	−0.21** (0.08)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R <sup>2</sup>	0.11	0.22	0.43	0.45
Robust SE <i>P</i> -Val	0.032	0.016	0.004	0.010
Wild Bootstrap <i>P</i> -Val	0.000	0.000	0.000	0.000
D. Dependent variable: Final exam essay questions score				
Computer class	−0.01 (0.08)	−0.01 (0.08)	−0.03 (0.07)	−0.02 (0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R <sup>2</sup>	0.37	0.41	0.54	0.54
Robust SE <i>P</i> -Val	0.882	0.853	0.682	0.742
Wild Bootstrap <i>P</i> -Val	0.687	0.727	0.318	0.426

## 观测性研究 (observational studies) 的挑战：选择性

- 在非实验研究中，研究者是数据生成过程的被动观测者，因此影响  $Y$  的因素很可能同时影响  $D$ ，意味着  $\varepsilon$  和  $D$  相关。此时若要研究“保持  $\varepsilon$  不变， $D$  由 0 变到 1”，有两种思路：
  - 将  $\varepsilon$  中与  $D$  相关的因素剥离出来，使得剩余的  $\varepsilon$  和  $D$  不相关。
  - 考察非  $\varepsilon$  所带来的  $D$  的变化。

这里我们先讨论前一种思路。

- 设想某学校允许学生在课堂上自由使用电脑，研究者记录下学生实际是否使用电脑及其考试成绩。影响考试成绩的其中一个因素是高中学习成绩，而高中学习成绩较差的学生更倾向于在大学课堂上使用电脑，因此处理组（使用电脑组）个体的高中学习成绩系统性地低于控制组（不使用电脑组）个体，两组个体考试成绩的差异既有可能反映使用电脑的效应，也有可能反映高中学习成绩的效应。

- 令  $\varepsilon$  中与  $D$  相关的因素为  $X$ （高中学习成绩），则线性模型可以改写作

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i'', \quad \mathbb{E}(\varepsilon_i'') = 0$$

$$\mathbb{E}(Y_i | D_i = 1, X_i = x) = \beta_0 + \beta_1 + \beta_2 x + \mathbb{E}(\varepsilon_i'' | D_i = 1, X_i = x)$$

$$\mathbb{E}(Y_i | D_i = 0, X_i = x) = \beta_0 + \beta_2 x + \mathbb{E}(\varepsilon_i'' | D_i = 0, X_i = x)$$

若

$$\mathbb{E}(\varepsilon_i'' | D_i = 1, X_i = x) = \mathbb{E}(\varepsilon_i'' | D_i = 0, X_i = x) \quad (1.2)$$

则

$$\beta_1(x) = \mathbb{E}(Y_i | D_i = 1, X_i = x) - \mathbb{E}(Y_i | D_i = 0, X_i = x)$$

- 假设 (1.2) 的直观含义是，在  $X$  相同的子样本内， $\varepsilon$  在处理组和控制组之间再次达到平衡——近似随机分组，因此可以将组间比较局限在  $X$  相同的子样本内以考察因果效应。<sup>[1]</sup>

---

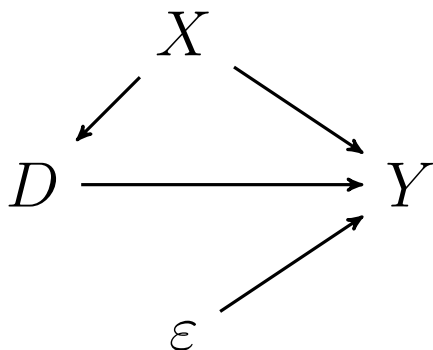
[1] 在不引起混淆的前提下，此后  $\varepsilon''$  仍写作  $\varepsilon$ .

- 假设 (1.2) 还可写作

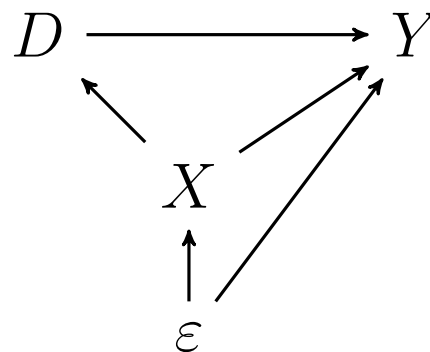
$$\mathbb{E}(\varepsilon_i | D_i, X_i) = \mathbb{E}(\varepsilon_i | X_i)$$

称  $\varepsilon_i$  与  $D_i$  **条件均值独立**, 记作  $\varepsilon_i \perp D_i | X_i$ .

- 一般而言,  $\varepsilon_i \perp D_i | X_i \Rightarrow \text{Cov}(\varepsilon_i, D_i | X_i) = 0$ , 当  $D_i$  为二元变量时, 两者等价。
- 称  $X$  为控制变量 (control variables) 或协变量 (covariates), 或者出于理解的需要, 也称之为分块变量 (blocking variable)。
- 请注意, 条件均值独立并不要求  $\mathbb{E}(\varepsilon_i | X_i) = \mathbb{E}(\varepsilon_i) = 0$ .



$$\mathbb{E}(\varepsilon_i | X_i) = \mathbb{E}(\varepsilon_i)$$



$$\mathbb{E}(\varepsilon_i | X_i) \neq \mathbb{E}(\varepsilon_i)$$

- 总结：实验数据 vs. 非实验数据（观测数据）
  - 实验数据： $D_i$  可以被看作是确定性的。研究者设定  $D_i$  并观测到  $Y_i$ 。一组新的样本即意味着一组新的  $\varepsilon_i$ （或曰一组新的  $Y_i$ ）。
  - 观测数据： $D_i$  和  $\varepsilon_i$  都是随机的。一组新的样本同时意味着一组新的  $D_i$  和  $\varepsilon_i$ （或曰一组新的  $D_i$  和  $Y_i$ ）。

- 绝大多数社会科学数据都是观测数据，各随机变量可能是联合决定的，特别而言， $\varepsilon_i$  的分布可能取决于  $D_i$ ，这是因为  $D_i$  部分地由  $\varepsilon_i$  决定，其直观含义是，绝大多数人类行动都是选择的结果而不是分配的结果，即**人们自选择 (self-select) 接受某项处理**。选择性分为两种：
  - **基于可观测变量的选择性** (selection on observables)。这意味着给定可观测变量，选择性将不复存在，这等价于是说， $\varepsilon$  与  $D$  条件均值独立。
  - **基于不可观测变量的选择性** (selection on unobservables)。但由于至少部分选择性变量是不可观测的，因此无法“给定”这些变量，即无法将组间比较局限在这些变量相同的子样本内以考察因果效应，这是观测性研究最大的挑战。
- 选择性也被称作内生性 (endogeneity) 或内生处理 (endogenous treatment)，“ $D$  具有选择性”和“ $D$  是内生的”是一回事。忽视选择性而造成对因果效应估计的偏误就称作选择性偏误（内生性偏误）。

- 选择性就是分配机制 (assignment mechanism)：每个个体如何接受处理，可以用倾向得分  $\pi \triangleq \Pr(D = 1|X, \varepsilon)$  来表示。
  - 基于可观测变量的选择性：倾向得分是可观测变量的未知函数。

$$\Pr(D = 1|X, \varepsilon) = \pi(X)$$

- 基于不可观测变量的选择性：倾向得分是不可观测变量的未知函数。

$$\Pr(D = 1|X, \varepsilon) = \pi(X, \varepsilon)$$

- 若不存在选择性, 则意味着倾向得分是常数, 也即随机分配 (random assignment)。

$$\Pr(D = 1|X, \varepsilon) = \pi(\text{const.})$$

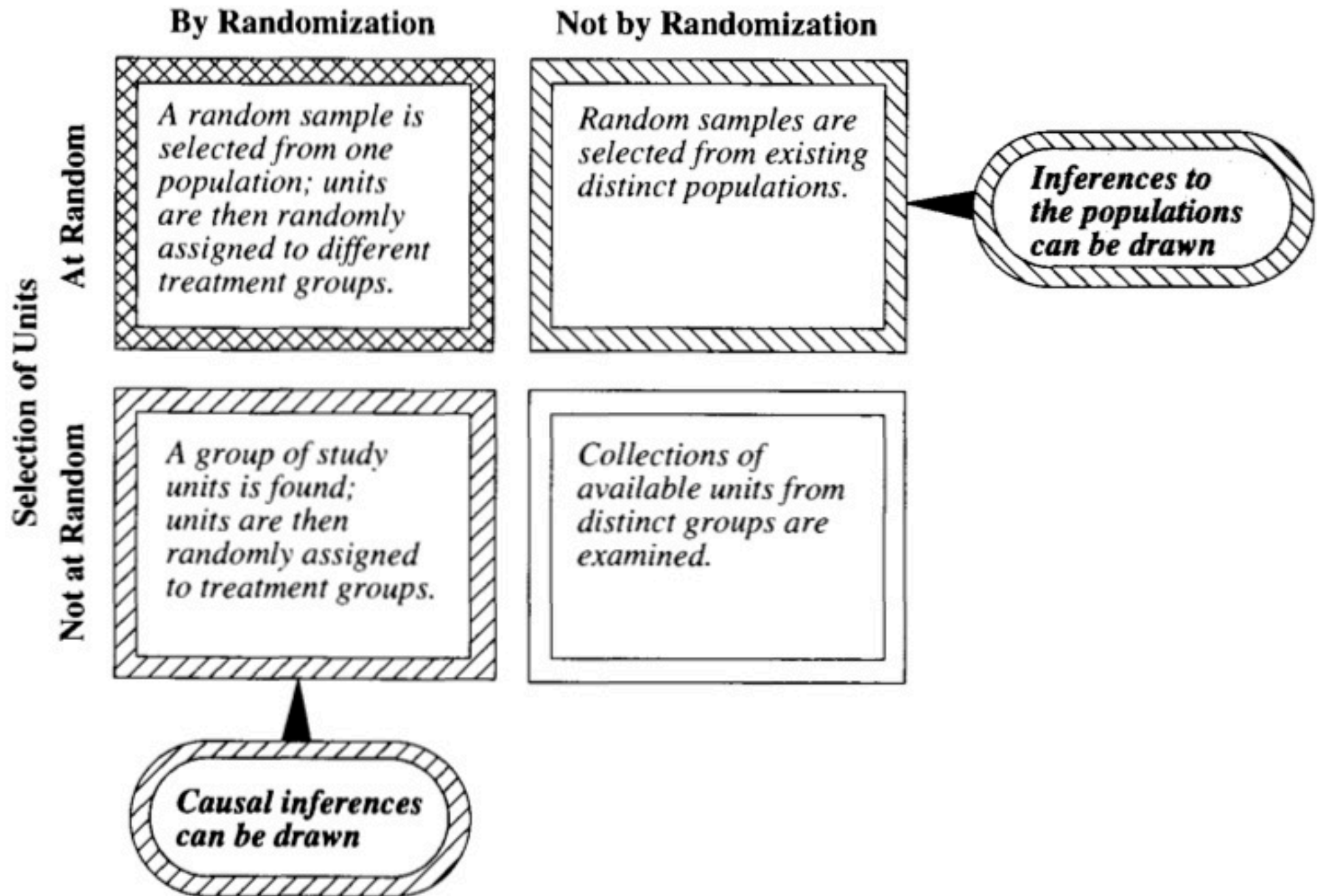
- 绝大多数因果推断实证方法都只能消除基于可观测变量的选择性，而优秀的研究设计能够通过消除基于可观测变量的选择性，达到消除基于不可观测变量的选择性的目的。
- 消除基于可观测变量的选择性，关键在于分块。
- 对于研究设计的另一项启示是：没有一项研究可以穷尽  $Y$  的所有影响因素，或者说，一项因果研究的待探究原因往往只有一个，因此只能也只需处理某个特定  $D$  的选择性或内生性问题。 $Y$  的影响因素中，绝大部分都被打包在  $\varepsilon$  中。关键的控制变量一定是造成潜在选择性的变量，而不会造成选择性的影响因素——即只影响  $Y$  而不影响  $D$  的因素——对于探究  $D$  对  $Y$  的因果效应往往并不重要。



## 自选择 vs. 样本选择

- 我们通常假定所采用的样本来自随机抽样 (random sampling), 即总体中的每个个体都以相同概率进入样本, 且抽取一个观测值不影响抽取其它观测值的概率, 此时称样本中的每个观测值满足独立同分布 (independently and identically distributed)。
- 随机抽样意味着不存在样本选择 (sample selection), 例如收入调查中富人的应答率较低, 或项目评估中处理组个体的非随机流失 (attrition)。
- 有时样本选择是由自选择引起的。例如在估计工资方程时, 尽管我们所感兴趣的总体是所有工作年龄的劳动力, 但只有实际参加工作的劳动力其工资才能被观测到, 因此存在样本选择, 其产生的原因正是劳动力自选择决定是否参加工作。这个问题应该被称作样本选择问题还是自选择问题? 这并不重要。重点在于, 我们在估计工资方程时必须正式处理样本的非随机特性。

## Allocation of Units to Groups



## **$D$ 为连续变量的一般情形**

- 此时  $D$  被称作连续处理 (continuous treatment)。而  $\beta_1$  的含义是，保持  $\varepsilon$  不变，当处理强度 (treatment intensity) 变化一个单位时，结果会变化  $\beta_1$  个单位，这样的  $\beta_1$  也被称作边际效应 (marginal effect)。

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

$$\mathbb{E}(Y_i|D_i) = \beta_0 + \beta_1 D_i + \mathbb{E}(\varepsilon_i|D_i)$$

若

$$\mathbb{E}(\varepsilon_i|D_i) = \mathbb{E}(\varepsilon_i) \equiv 0 \quad (1.3)$$

则

$$\beta_1 = \frac{d\mathbb{E}(Y_i|D_i)}{dD_i}$$

- 此时我们说，因果效应可以用“ $D_i$  变化一个单位， $Y_i$  平均变化多少个单位”来衡量。请注意，这是一种衡量手段，而不是  $\beta_1$  的定义，因为这种衡量手段本质上依赖的是相关性，而当假设 (1.3) 成立时，相关性可以揭示因果性。

- 此时的线性模型新增了一个限制性假设：边际效应不随着  $D$  的水平而变化，称这一假设为函数形式 (functional form) 假设或模型设定 (model specification) 假设。
- 对于含控制变量情形， $\beta_1$  的含义是，保持  $X$  和  $\varepsilon$  不变，当处理强度变化一个单位时，结果会变化  $\beta_1$  个单位。

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i$$

$$\mathbb{E}(Y_i | D_i, X_i) = \beta_0 + \beta_1 D_i + \beta_2 X_i + \mathbb{E}(\varepsilon_i | D_i, X_i)$$

若

$$\mathbb{E}(\varepsilon_i | D_i, X_i) = \mathbb{E}(\varepsilon_i | X_i) = f(X_i) \quad (1.4)$$

其中  $f(X_i)$  是（仅）关于  $X_i$  的未知函数。则

$$\beta_1 = \frac{\partial \mathbb{E}(Y_i | D_i, X_i)}{\partial D_i}$$

- 此时我们说，因果效应可以用“保持  $X_i$  不变， $D_i$  变化一个单位， $Y_i$  平均变化多少个单位”来衡量。“保持  $X_i$  不变”，英文的说法是“holding everything constant” 或“other things being equal”，拉丁文的说法是“ceteris paribus”。
- 此时的线性模型施加了更强的函数形式假设：边际效应不随着  $D$  或  $X$  的水平而变化。