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: March 7, 2021

课前寄语

入门须正

取法须高

立志须远


發上等願
結中等緣
享下等福
擇高處立
尋平處住
向寬處行
左宗棠題江蘇無錫梅園
歲次甲午年冬月

Prerequisite equivalency.

- 《计量经济学》(*Introduction to Econometrics*),斯托克 (James H. Stock)、沃森 (Mark W. Watson) 著。中文第三版，格致出版社，2012 年（中国人民大学出版社，2014 年）。
- 《计量经济学导论：现代观点》(*Introductory Econometrics: A Modern Approach*)，伍德里奇 (Jeffrey M. Wooldridge) 著。中文第六版，中国人民大学出版社，2018 年。

Required software. STATA®

Recommended texts.

- 《基本无害的计量经济学》(*Mostly Harmless Econometrics: An Empiricist's Companion*), 安格里斯特 (Joshua D Angrist)、皮施克 (Jorn-Steffen Pischke) 著。格致出版社，2012 年。
- 《精通计量：从原因到结果的探寻之旅》(*Mastering 'Metrics: The Path from Cause to Effect*), 安格里斯特、皮施克著。格致出版社，2019 年。
- 《用 Stata 学计量经济学》(*An Introduction to Modern Econometrics Using Stata*), 鲍姆 (Christopher F. Baum) 著。中国人民大学出版社，2012 年。
- 《用 Stata 学微观计量经济学》(*Microeconometrics Using Stata, Revised Edition*), 卡梅伦 (A. Colin Cameron)、特里维迪 (Pravin K. Trivedi) 著。重庆大学出版社，2015 年。

- 《横截面与面板数据的计量经济分析》(*Econometric Analysis of Cross Section and Panel Data*), 伍德里奇著。中文第二版，中国人民大学出版社，2016 年。
- 《计量经济分析》(*Econometric Analysis*), 格林 (William H Greene) 著。中文第六版，中国人民大学出版社，2011 年。英文第七版，中国人民大学出版社，2013 年。英文最新版，8th edition, 2017.
- *Econometrics*, Bruce E. Hansen, manuscript, 2021.

条件期望与回归

- 我们花了很多精力学习回归理论。回归就是最小二乘，是一种估计一个变量在给定其它变量下的条件期望的工具。条件期望为什么重要？

示例 1. 教育与收入 (CFPS, 2018)

- 条件期望函数 (conditional expectation function, CEF) $\mathbb{E}(y|\mathbf{x})$ 是给定 $\mathbf{x}$ 对 y 的最佳预测。

$$\mathbb{E}(y|\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}\left(y - f(\mathbf{x})\right)^2$$

- 定义期望残差 $\tilde{\varepsilon} \triangleq y - \mathbb{E}(y|\mathbf{x})$，具有如下性质：
  - $\tilde{\varepsilon}$ 均值独立于 $\mathbf{x}$，即 $\mathbb{E}(\tilde{\varepsilon}|\mathbf{x}) = 0$.
  - $\tilde{\varepsilon}$ 期望为零，即 $\mathbb{E}(\tilde{\varepsilon}) = 0$.
  - $\tilde{\varepsilon}$ 与 $\mathbf{x}$ 不相关，即 $\mathbb{E}\left(\mathbf{x}\tilde{\varepsilon}\right) = 0$.
  - $\tilde{\varepsilon}$ 均值独立于 $\mathbf{x}$ 的任意函数，即 $\mathbb{E}\left(\tilde{\varepsilon}|f(\mathbf{x})\right) = 0$.
  - $\tilde{\varepsilon}$ 与 $\mathbf{x}$ 的任意函数不相关，即 $\mathbb{E}\left(f(\mathbf{x})\tilde{\varepsilon}\right) = 0$.

- 一种重要的特殊情形是线性条件期望函数：

$$\mathbb{E}(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

- 定义总体最小二乘问题：

$$\min_{\boldsymbol{\beta}} \mathbb{E}\left(y - \mathbf{x}'\boldsymbol{\beta}\right)^2$$

- 显然，当 CEF 确为线性时，总体最小二乘问题的解即为 CEF。
- 但我们通常并不知道 CEF 的函数形式，因此线性只是对其的近似。可以证明，当 CEF 为非线性时，总体最小二乘问题的解是 CEF 的**最佳线性近似**。

$$\arg\min_{\boldsymbol{\beta}} \mathbb{E}\left(y - \mathbf{x}'\boldsymbol{\beta}\right)^2 = \arg\min_{\boldsymbol{\beta}^*} \mathbb{E}\left(\mathbb{E}(y|\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}^*\right)^2$$

- 我们把 $\mathbf{x}'\boldsymbol{\beta}$ 称作总体回归函数。定义总体回归残差 $\tilde{\tilde{\varepsilon}} \triangleq y - \mathbf{x}'\boldsymbol{\beta}$，得到如下线性回归模型：

$$y = \mathbf{x}'\boldsymbol{\beta} + \tilde{\tilde{\varepsilon}}$$

求解总体最小二乘问题，可得[1]

$$\mathbb{E}\left(\mathbf{x}(y - \mathbf{x}'\boldsymbol{\beta})\right) = \mathbb{E}\left(\mathbf{x}\tilde{\tilde{\varepsilon}}\right) = 0$$

$$\boldsymbol{\beta} = \left[\mathbb{E}(\mathbf{x}\mathbf{x}')\right]^{-1}\mathbb{E}(\mathbf{x}y)$$

[1] 注意，$\mathbb{E}\left(\tilde{\tilde{\varepsilon}}|\mathbf{x}\right) = 0$ 未必成立，只有当 CEF 确为线性时才成立。而 $\mathbb{E}\left(\tilde{\varepsilon}|\mathbf{x}\right) = 0$ 始终成立。

何为因果推断？

- 我们关注 CEF 的目的在于，它能帮助我们理解变量之间的因果关系。我们真正关心的因果问题是：读书有没有用？一个人如果多上一年学，他的工资水平预期能增长多少？

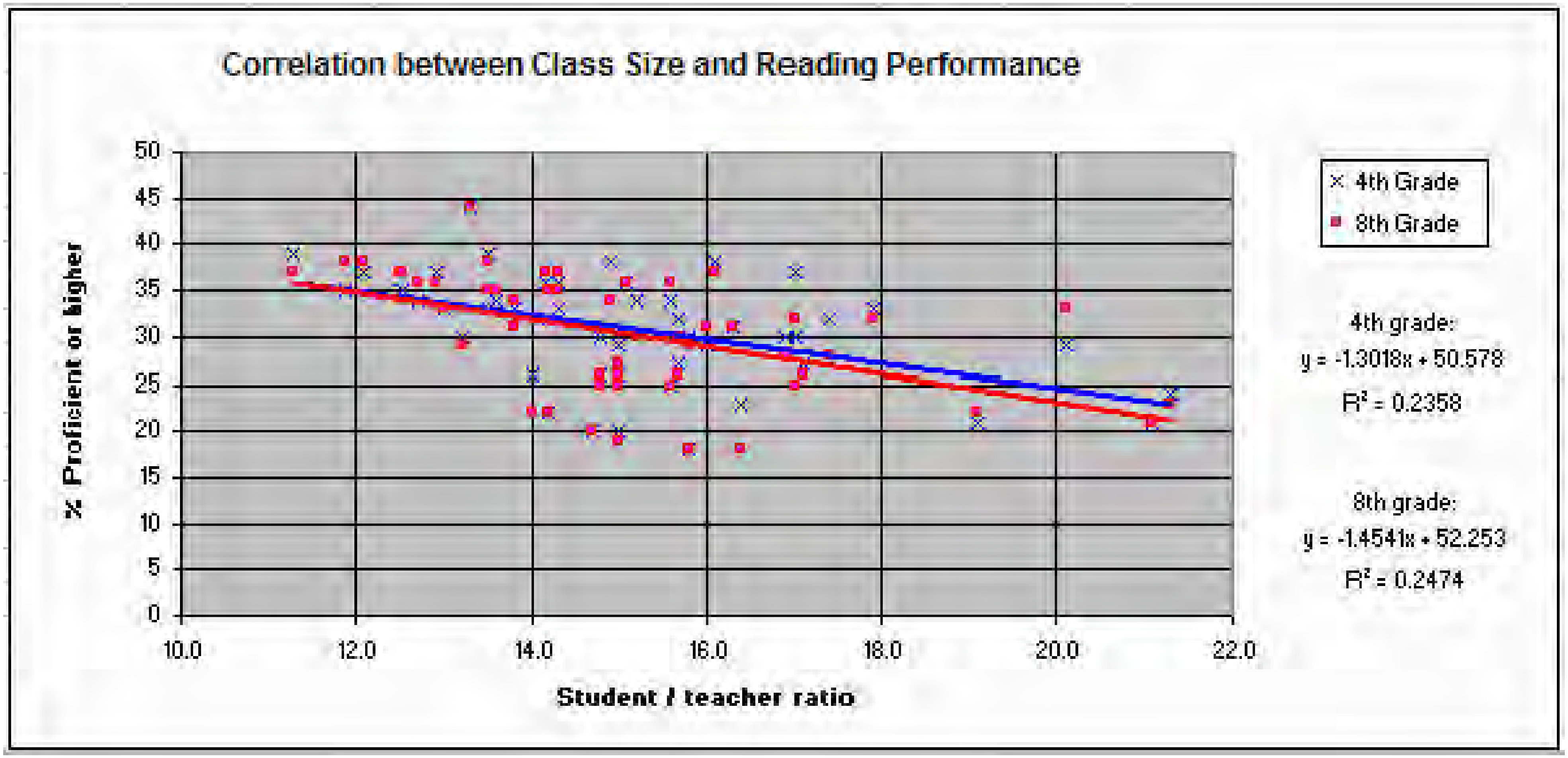
- 用 Y 表示我们感兴趣的结果 (outcome)，或反应 (response)。
- 用 D 表示我们感兴趣的原因 (cause)，或处理 (treatment)、干预 (intervention)。D 可以是离散的，也可以是连续的。
- 用 ε 表示影响结果的其它因素。
- 我们感兴趣的因果关系可以用如下的基本因果模型来刻画：

$$\begin{tikzcd} D \arrow[r] & Y \\ \varepsilon \arrow[ur] & \end{tikzcd}$$

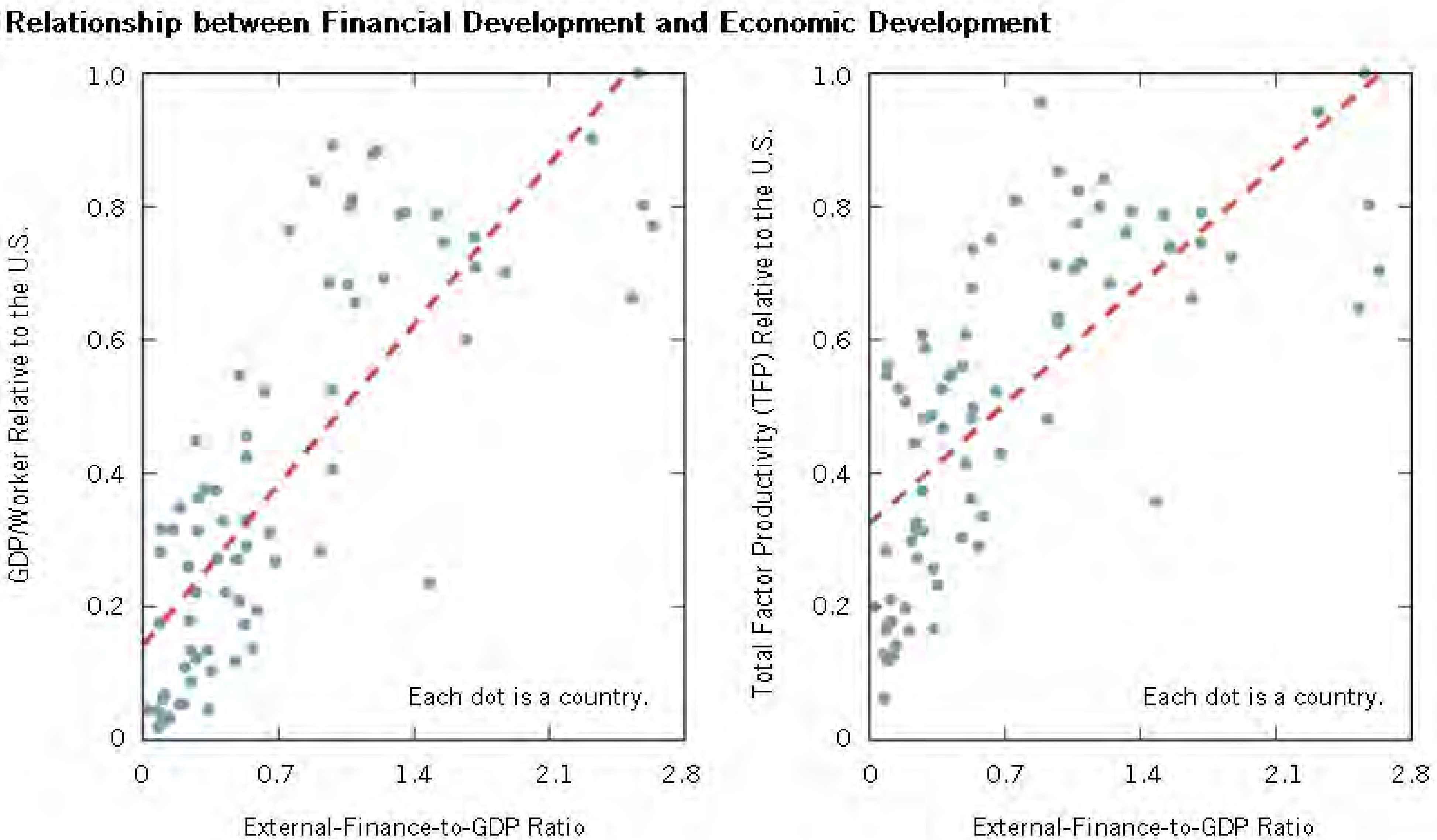
模型 I

- 再来看三个例子，刻画了三组相关性事实，体会它们试图讲述什么因果故事？

示例 2. 班级规模与教育产出

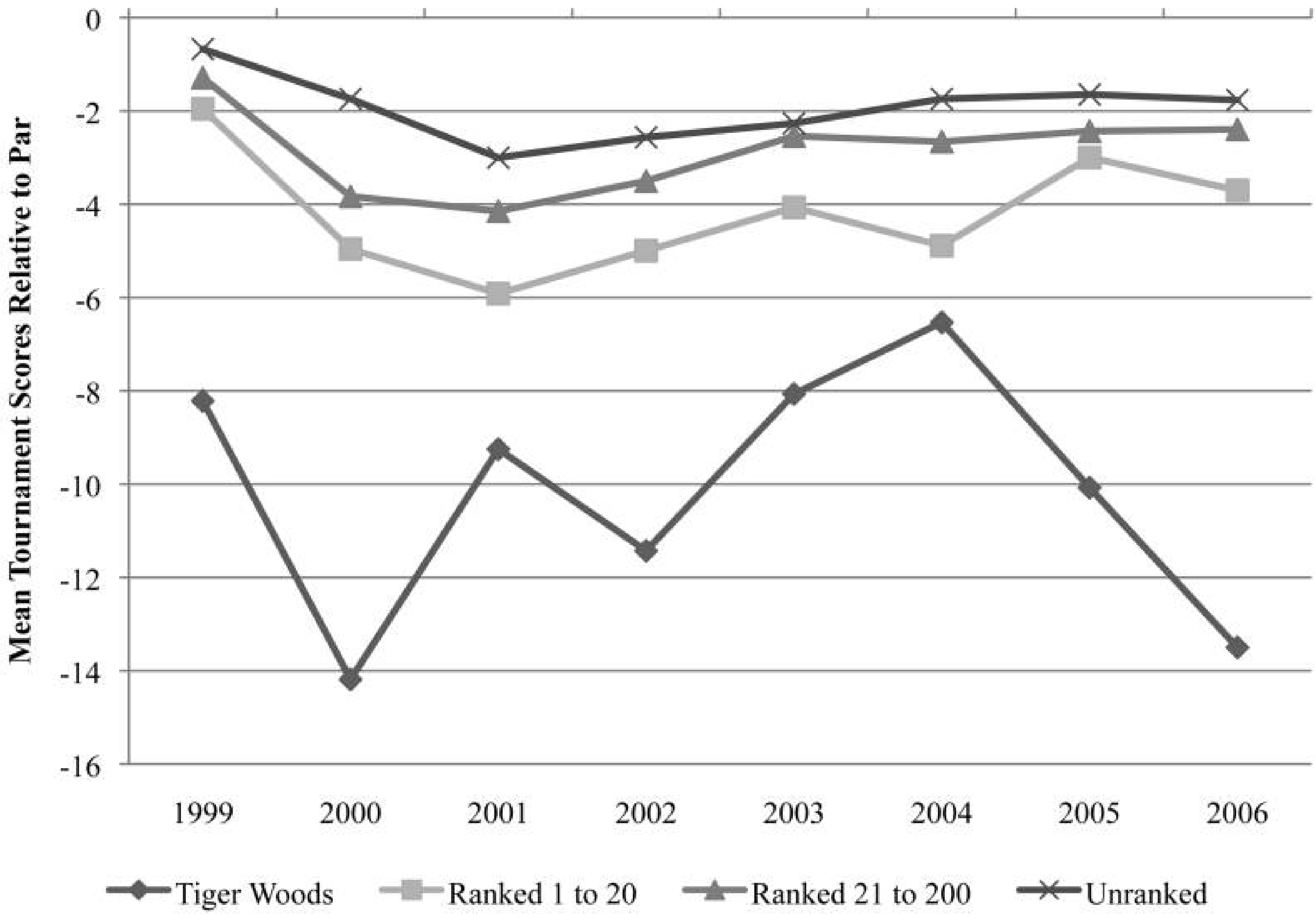


示例 3. 金融发展与经济增长

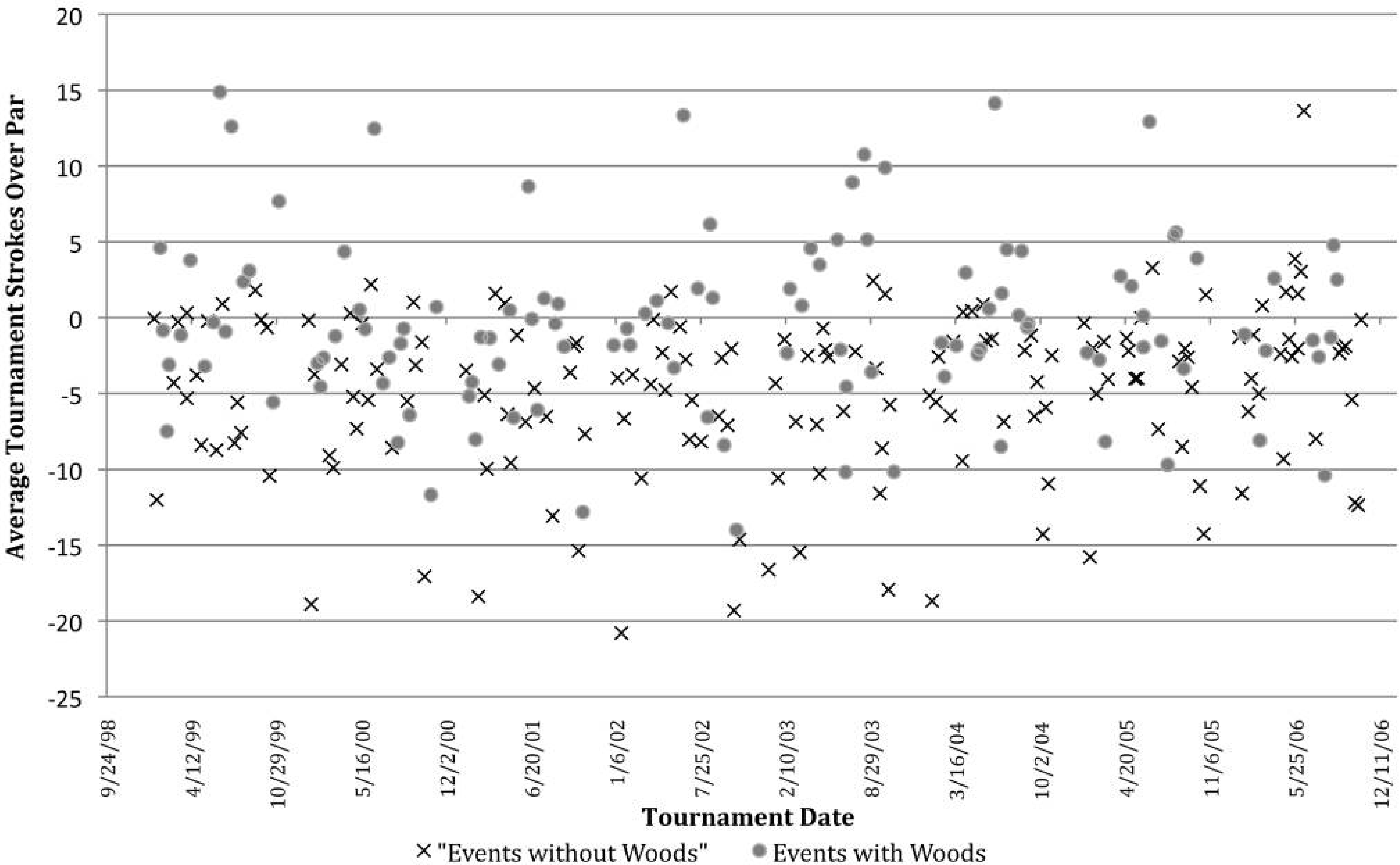


示例 4. 超级明星效应 (Brown, 2011, *JPE*).

- 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。



0
-2
-4
-6
-8
-10
-12
-14
-16
Mean Tournament Scores Relative to Par
1999
2000
2001
2002
2003
2004
2005
2006
Tiger Woods
Ranked 1 to 20
Ranked 21 to 200
Unranked



Average Tournament Strokes Over Par
20
15
10
5
0
-5
-10
-15
-20
-25
9/24/98
4/12/99
10/29/99
5/16/00
12/2/00
6/20/01
1/6/02
7/25/02
2/10/03
8/29/03
3/16/04
10/2/04
4/20/05
11/6/05
5/25/06
12/11/06
Tournament Date
"Events without Woods"
Events with Woods

- 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)：

$$\begin{tikzcd} D \arrow[r] & Y \\ \varepsilon \arrow[ur] & \end{tikzcd}$$

模型 I

$$\begin{tikzcd} D & Y \\ \varepsilon \arrow[u] \arrow[ur] & \end{tikzcd}$$

模型 II

$$\begin{tikzcd} D & Y \arrow[l] \end{tikzcd}$$

模型 III

- 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？
- 如果在特定的研究情境下，变量之间满足一定的假设条件，使得一个特定的因果模型没有与之竞争的观测上等价 (observationally equivalent) 的因果模型，我们就说这个特定的因果模型被识别 (identified)。
- 这样的假设被称作识别假设，我们马上就会看到，识别假设是永远无法严格证明的，只能根据社会科学理论加以论证。

- 因此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关系？这是社会科学理论的任务。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。
  - **统计推断 (statistical inference)**：如何从样本数据获取关于总体因果关系的信息？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。

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

$$
\begin{tikzcd}
D \arrow[rr] & & Y \\
\varepsilon \arrow[u] \arrow[urr] & &
\end{tikzcd}
$$

模型 IV

- 模型 I 和模型 IV 都可以用如下的线性模型来表示：

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

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

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

- β_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 ε 不变，一项处理的实施（D 由 0 变到 1，或 D 变化一个单位），导致结果变化 β_1，称之为因果效应 (causal effect) 或处理效应 (treatment effect)。

- 这个模型看起来和线性回归模型长得很相似，但含义截然不同。这个模型被称为结构模型，因为它包含了我们关于因果关系的先验知识：ε 的含义，模型的线性性质，以及 D 和 ε 之间的关系，这些知识将帮助我们识别 β_1。因此 β_1 被称为结构参数，ε 被称为结构误差项或结构扰动项。
- 若 ε 均值独立于 D，即 $\mathbb{E}(\varepsilon|D)=0$，则 $\mathbb{E}(Y|D)=\beta_0+\beta_1 D$，因此线性回归能够识别 β_1；反之，若 $\mathbb{E}(\varepsilon|D)=h(D)\neq 0$，

$$\mathbb{E}(Y|D)=\beta_0+\beta_1 D+h(D)$$

  线性回归仍然能够得到 $\mathbb{E}(Y|D)$ 或其最佳线性近似，但是无法识别 β_1。
- 注意，

$$Y=\beta_0+\beta_1 D+h(D)+\tilde{\varepsilon},\ \tilde{\varepsilon}=\varepsilon-h(D)$$

  $\mathbb{E}(\tilde{\varepsilon}|D)=0$ 自动成立，但 $\mathbb{E}(\varepsilon|D)=0$ 是否成立却需要借助社会科学理论加以判断，后者就是区分模型 I 和模型 IV 的识别假设，它是无法从数学上或统计上加以证明的，因为 ε 是未加观测或无法观测的。

- 在教育回报率的例子中,用 D 表示是否上过大学,Y 表示工资水平,由于 D 为二元变量,因此 $\mathbb{E}(Y|D)$ 必然可以表示为 $\mathbb{E}(Y|D)=\gamma_0+\gamma_1 D$,事实上

$$
\begin{aligned}
\gamma_0 &= \mathbb{E}(Y|D=0) \\
\gamma_1 &= \mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)
\end{aligned}
$$

由此得到线性回归模型

$$
Y=\gamma_0+\gamma_1 D+\tilde{\varepsilon}
$$

$$
\mathbb{E}(\tilde{\varepsilon}|D)=0
$$

但 γ_1 并不具有因果含义,它只表示总体中上过大学人群和没上过大学人群的平均工资差异。

线性回归模型试图回答的是如下的**预测性问题**:“**如果我们观测到** D **的取值为** D_0,我们预期 Y 的取值为何?”

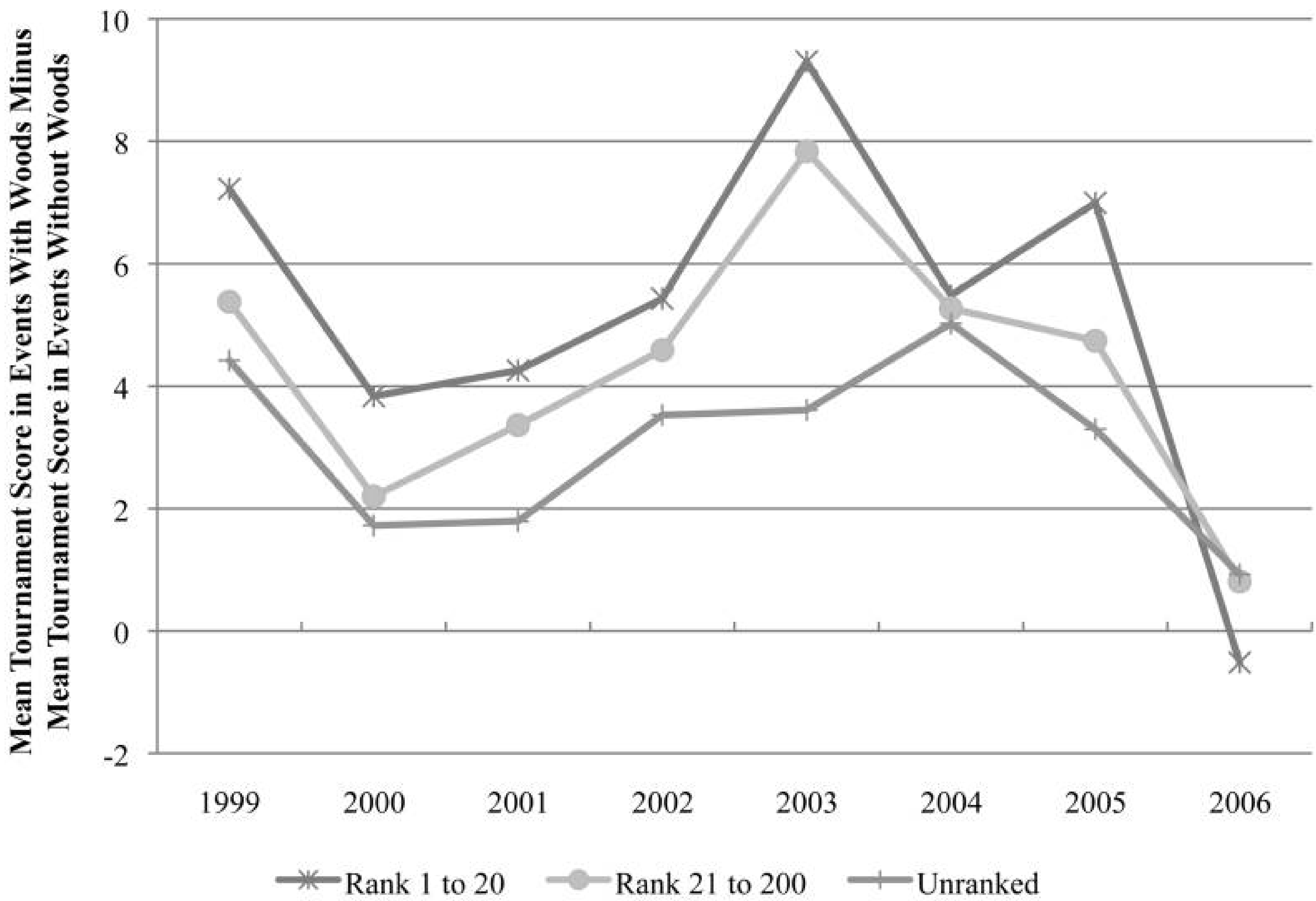
相反，当我们写下结构模型

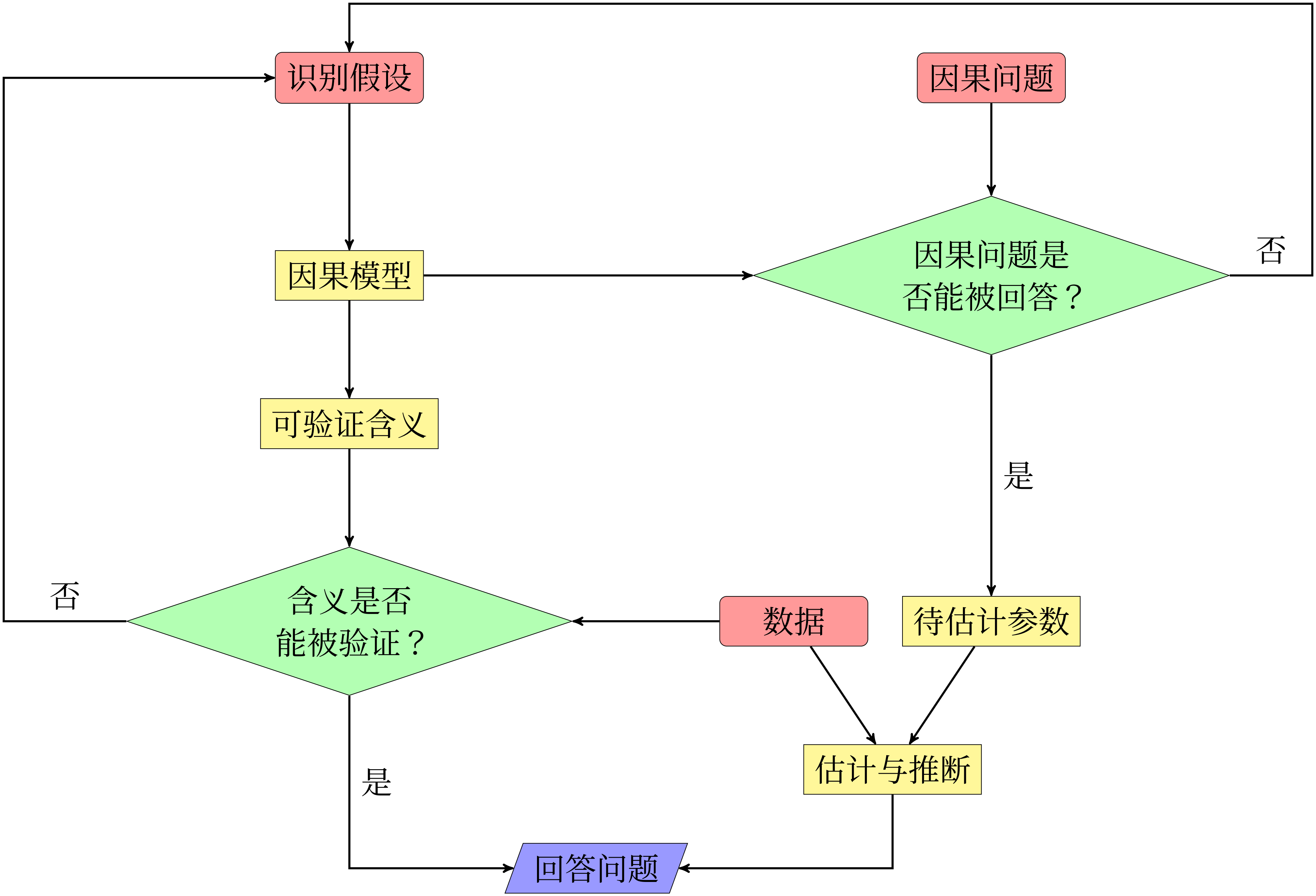
$$Y = \beta_0 + \beta_1 D + \varepsilon$$

我们是把 ε 解释为“影响工资水平的不可观测的能力或积极性”，那么 ε 很有可能和 D 相关（能力越强的人越倾向于上过大学）。此时线性回归就无法识别因果效应 β_1。

结构模型试图回答的是如下的**因果性问题**：“**如果我们干预人们的上大学行为，将 D 的取值设定为 D_0**，我们预期 Y 的取值为何？”

- 两种基本的识别策略：
  - 寻找特定的研究情境。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)。
  - 有时很难令人信服地论证识别假设的成立，此时尝试去挖掘因果模型更丰富的、可验证的相关性含义 (testable implications)，即提出这样的问题，“如果从 D 到 Y 的因果关系真的存在，那么我们还将观测到何种相关现象？”


Mean Tournament Score in Events With Woods Minus
Mean Tournament Score in Events Without Woods
10
8
6
4
2
0
-2
1999
2000
2001
2002
2003
2004
2005
2006
Rank 1 to 20
Rank 21 to 200
Unranked


识别假设
因果问题
因果模型
因果问题是
否能被回答？
否
可验证含义
是
含义是否
能被验证？
否
数据
待估计参数
估计与推断
是
回答问题

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，将其随机划分为两组，对处理组个体实施处理，对控制组个体不实施处理。

$$D_i = \begin{cases} 1 & \text{对 } i \text{ 实施处理（进入处理组、实验组）} \\ 0 & \text{不对 } i \text{ 实施处理（进入控制组、对照组）} \end{cases}$$

- 尽管每位被试的 ε_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 ε 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。

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

$$\begin{aligned}\mathbb{E}(Y_i|D_i = 1) &= \beta_0 + \beta_1 + \mathbb{E}(\varepsilon_i|D_i = 1)\\ \mathbb{E}(Y_i|D_i = 0) &= \beta_0 + \mathbb{E}(\varepsilon_i|D_i = 0)\end{aligned}$$

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) 即为随机实验的识别假设，正式表述为

$$\textbf{Assumption LS.1:} \quad \mathbb{E}(\varepsilon_i|D_i) = \mathbb{E}(\varepsilon_i)$$

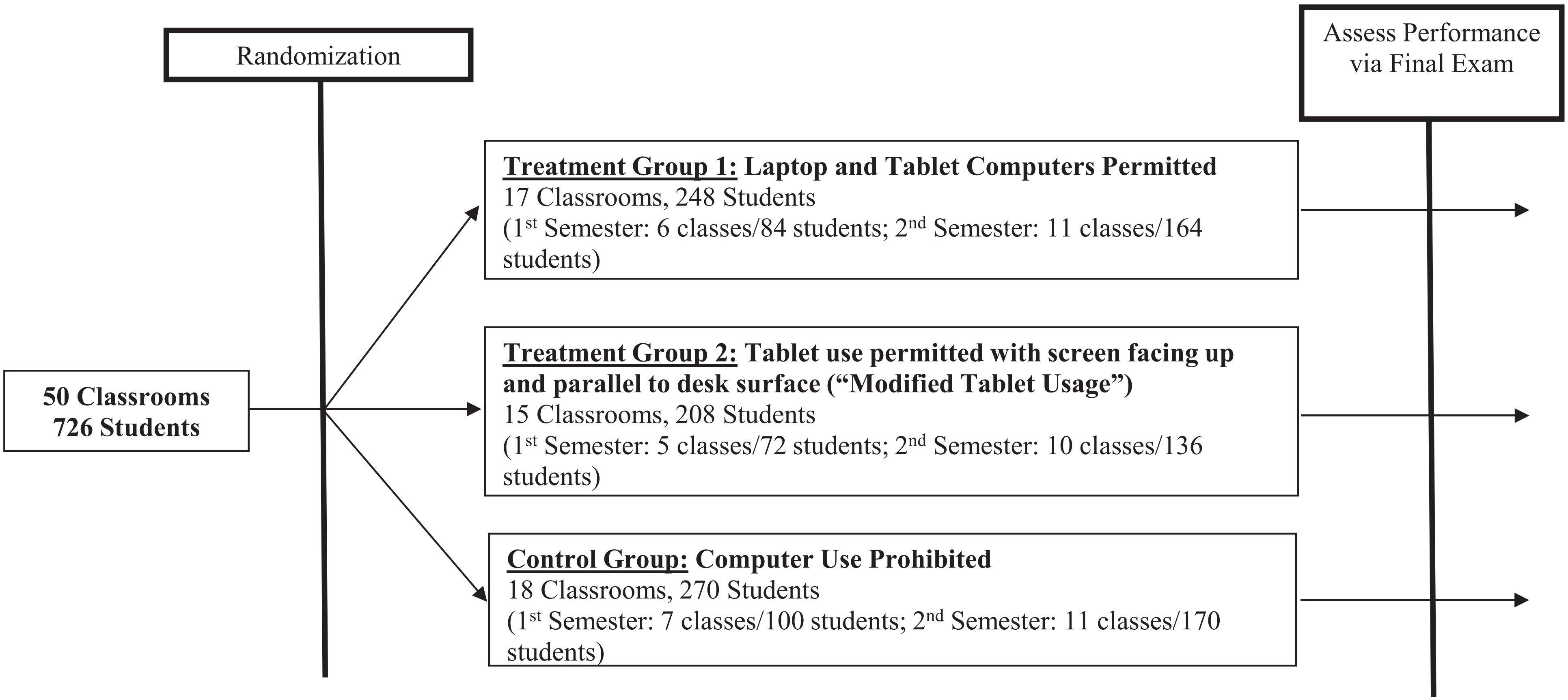
称 ε_i 与 D_i**均值独立**，或

$$\textbf{Assumption LS.1}': \quad \mathrm{Cov}(D_i, \varepsilon_i) = \mathbb{E}(D_i\varepsilon_i) = 0$$

称 ε_i 与 D_i**不相关**。

- 一般而言，假设**LS.1**比假设**LS.1′**更强，当 D_i 为二元变量时，两者等价。但这一区别通常只具有数学上的意义，不具有社会科学上的意义。
- 回忆 β_1 的含义 :"保持 ε 不变，D 由 0 变到 1 所导致的 Y 的变化"。这里存在一个因果识别的基本难题：一方面，必须在**干预性视角**下理解因果效应；但另一方面，理想的干预是不可行的。
- 假设**LS.1**使得因果识别成为可能，此时蕴含着一种视角的转换——在干预性视角和**相关性视角**之间建立联系：控制组与处理组的比较，等价于对同一组个体施加干预与否的比较。

示例 5. 使用电脑有损学习成绩 (Carter et al, 2017, *Economics of Education Review*)



	Control	Treatment 1 (laptops/tablets)	Treatment 2 (tablets, face up)	Both treatments vs. control	Treatment 1 vs. control	Treatment 2 vs. control
	(1)	(2)	(3)	(4)	(5)	(6)
A. Baseline characteristics						
Female	0.17	0.20	0.19	0.03	0.06	0.00
				(0.03)	(0.04)	(0.04)
White	0.64	0.67	0.66	0.02	0.02	0.02
				(0.04)	(0.04)	(0.05)
Black	0.11	0.10	0.11	−0.02	−0.02	−0.03
				(0.03)	(0.03)	(0.04)
Hispanic	0.13	0.13	0.09	0.00	0.02	−0.03
				(0.03)	(0.03)	(0.03)
Age	20.12	20.15	20.15	0.03	0.05	0.06
	[1.06]	[1.00]	[0.96]	(0.08)	(0.09)	(0.10)
Prior military service	0.19	0.19	0.16	−0.02	0.00	−0.01
				(0.03)	(0.04)	(0.04)
Division I athlete	0.29	0.40	0.35	0.05	0.07*	0.04
				(0.04)	(0.04)	(0.05)
GPA at baseline	2.87	2.82	2.89	−0.01	−0.05	0.03
	[0.52]	[0.54]	[0.51]	(0.04)	(0.05)	(0.05)
Composite ACT	28.78	28.30	28.30	−0.34	−0.37	−0.54
	[3.21]	[3.46]	[3.27]	(0.26)	(0.31)	(0.33)
P-Val (Joint χ^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***	0.79***	0.40***
				(0.02)	(0.03)	(0.04)
Average computer use	0.00	0.57	0.22	0.42***	0.56***	0.24***
				(0.02)	(0.02)	(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***	−0.23***	−0.19***	−0.18***
	(0.10)	(0.09)	(0.07)	(0.07)
GPA at start of course			1.09***	0.92***
			(0.07)	(0.07)
Composite ACT				0.07***
				(0.01)
Demographic controls		X	X	X
R^2	0.08	0.28	0.54	0.57
Robust SE *P*-Val	0.003	0.007	0.005	0.005
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
B. Dependent variable: Final exam multiple choice score				
Computer class	−0.25***	−0.20**	−0.16**	−0.15**
	(0.10)	(0.009)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.08	0.27	0.48	0.50
Robust SE *P*-Val	0.009	0.023	0.025	0.029
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
C. Dependent variable: Final exam short answer score				
Computer class	−0.25***	−0.21**	−0.18**	−0.17**
	(0.09)	(0.09)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.08	0.21	0.44	0.46
Robust SE *P*-Val	0.008	0.016	0.017	0.019
Wild Bootstrap *P*-Val	0.008	0.020	0.022	0.028
D. Dependent variable: Final exam essay questions score				
Computer class	−0.03	−0.01	0.02	0.02
	(0.08)	(0.08)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.32	0.37	0.50	0.51
Robust SE *P*-Val	0.705	0.912	0.801	0.755
Wild Bootstrap *P*-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*	−0.18**	−0.20***	−0.17**
	(0.10)	(0.09)	(0.07)	(0.07)
GPA at start of course			1.12***	1.01***
			(0.07)	(0.08)
Composite ACT				0.05***
				(0.01)
Demographic controls		X	X	X
R^2	0.07	0.26	0.53	0.54
Robust SE *P*-Val	0.087	0.050	0.007	0.019
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
B. Dependent variable: Final exam multiple choice score				
Computer class	−0.15	−0.15*	−0.17**	−0.14*
	(0.10)	(0.09)	(0.08)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.07	0.26	0.48	0.49
Robust SE *P*-Val	0.141	0.100	0.027	0.057
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
C. Dependent variable: Final exam short answer score				
Computer class	−0.21**	−0.22**	−0.24***	−0.21**
	(0.10)	(0.09)	(0.08)	(0.08)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.11	0.22	0.43	0.45
Robust SE *P*-Val	0.032	0.016	0.004	0.010
Wild Bootstrap *P*-Val	0.000	0.000	0.000	0.000
D. Dependent variable: Final exam essay questions score				
Computer class	−0.01	−0.01	−0.03	−0.02
	(0.08)	(0.08)	(0.07)	(0.07)
Demographic controls		X	X	X
GPA control			X	X
ACT control				X
R^2	0.37	0.41	0.54	0.54
Robust SE *P*-Val	0.882	0.853	0.682	0.742
Wild Bootstrap *P*-Val	0.687	0.727	0.318	0.426

控制变量的作用

- 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 ε 和 D 相关。此时若要研究“保持 ε 不变，D 由 0 变到 1”，有两种思路：
  - 将 ε 中与 D 相关的因素剥离出来，使得剩余的 ε 和 D 不相关。
  - 考察非 ε 所带来的 D 的变化。

  这里我们先讨论前一种思路。
- 设想一个非实验情境：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。仍将使用电脑的学生归作处理组，不使用电脑的学生归作控制组。
- 在这一研究情境中，研究者无法假设 ε 和 D 不相关。例如，ε 中可能包含一个因素叫“学习习惯”。一方面，学习习惯不佳的学生考试成绩较差；另一方面，学习习惯不佳的学生更倾向于在课堂上使用电脑。因此处理组和控制组考试成绩的差异既有可能反映使用电脑的效应，也有可能反映学习习惯的效应。

- 考虑到是否按时出勤一定程度上能够反映学习习惯，因此采用出勤率 (X) 作为学习习惯的代理变量 (proxy variable)，则线性模型可以改写作

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

$$\mathbb{E}(Y_i|D_i = 1, X_i = x) = \beta_0 + \beta_1 + \beta_2 x + \mathbb{E}(\varepsilon_i''|D_i = 1, X_i = x)$$
$$\mathbb{E}(Y_i|D_i = 0, X_i = x) = \beta_0 + \beta_2 x + \mathbb{E}(\varepsilon_i''|D_i = 0, X_i = x)$$

若

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

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，原 ε 和 D 的相关性可以由它和 X 的相关性完全捕捉，在 X 相同的子总体内，新 ε'' 和 D 不再相关，新 ε'' 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 相同的子总体内以考察因果效应。[2]

[2] 在不引起混淆的前提下，此后 ε'' 仍写作 ε.

- “把比较局限在 X 相同的子总体内”这个想法，我们经常简略地说成“给定 X”或“保持 X 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"。称 X 为控制变量 (control variables) 或协变量 (covariates)。
- 假设 (1.2) 可以正式表述为

$$\textbf{Assumption LS.2:} \quad \mathbb{E}(\varepsilon_i|D_i, X_i) = \mathbb{E}(\varepsilon_i|X_i)$$

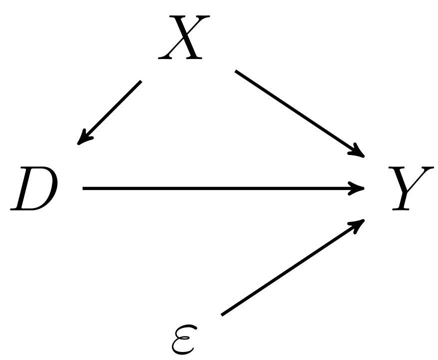
称 ε_i 与 D_i**条件均值独立**，或

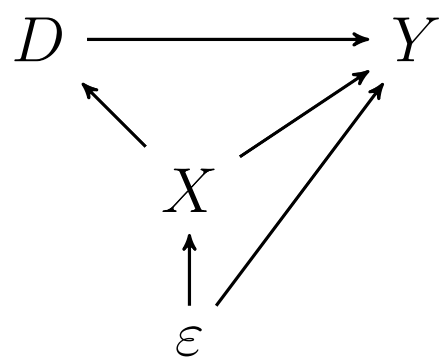
$$\textbf{Assumption LS.2}':\quad \mathrm{Cov}(D_i, \varepsilon_i|X_i) = 0$$

称 ε_i 与 D_i**条件不相关**。

- 一般而言，假设**LS.2**比假设**LS.2′**更强，当 D_i 为二元变量时，两者等价。

- 注意，假设**LS.2**并不要求 $\mathbb{E}(\varepsilon_i|X_i) = \mathbb{E}(\varepsilon_i)$ 或 $\mathrm{Cov}(X_i, \varepsilon_i) = 0$，其区别见以下两图。





- 在左图中，$\mathrm{Cov}(D_i, \varepsilon_i) = \mathrm{Cov}(X_i, \varepsilon_i) = 0$，这是传统计量教科书中的假设，但却是比假设**LS.2**更强的假设，此时 D_i 和 X_i 都是外生的，$\hat{\beta}_1^{OLS}$ 和 $\hat{\beta}_2^{OLS}$ 能够分别反映使用电脑和出勤对考试成绩的因果效应，即 $\hat{\beta}_1^{OLS} \to_p \beta_1$ 且 $\hat{\beta}_2^{OLS} \to_p \beta_2$。但考察出勤变量的外生性给研究增加了额外的困难。因此，要么实证研究者在普遍地掩耳盗铃，要么这并不是实证研究者实际采用的假设。

- 在右图中，$\text{Cov}(D_i, \varepsilon_i) \neq 0$，但 $\text{Cov}(D_i, \varepsilon_i|X_i) = 0$，则 $\hat{\beta}_1^{OLS} \to \beta_1$ 成立，而 $\hat{\beta}_2^{OLS} \to_p \beta_2$ 不一定成立。

$$\varepsilon_i = \delta_0 + \delta_1 D_i + \delta_2 X_i + v_i$$

$$\text{Cov}(D_i, v_i) = \text{Cov}(X_i, v_i) = 0$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i \\ &= (\beta_0 + \delta_0) + (\beta_1 + \delta_1) D_i + (\beta_2 + \delta_2) X_i + v_i \end{aligned}$$

$$\hat{\beta}_1^{OLS} \to_p \beta_1 \Leftrightarrow \delta_1 = 0 \Leftrightarrow \text{Cov}(D_i, \varepsilon_i|X) = 0$$

而 δ_2 通常不为零，$\hat{\beta}_2^{OLS} \to_p \beta_2 + \delta_2 \neq \beta_2$。

- 控制变量的双重作用：首先是控制 X 的直接效应 (β_2)，使得对回归系数的估计更准确（回归模型整体拟合程度更高，从而系数估计的标准误更小）；但更重要的是切断影响 Y 的其它因素（隐藏在扰动项中）与 D 的相关性，研究者希望，这个因素与 X 相关 (δ_2)，并且一旦控制 X 以后，这个因素不再与 D 相关。
- 因此 $\delta_2 \neq 0$ 是控制变量发挥作用的题中应有之义，其伴随的结果是无法准确识别控制变量的因果效应（还好我们并不关心）。
- 正确理解控制变量，要记住三句话：
  1. **一项因果推断研究待探究的原因往往只有一个，因此只能也只需处理某个特定 D 的内生性问题。**“XXX 的影响因素研究”不会是一项好的因果推断研究。
  2. 大部分影响 Y 的因素都被打包在 ε 中，**关键的控制变量一定是既影响 D 又影响 Y 的因素，只影响 Y 而不影响 D 的因素对于探究 D 对 Y 的因果关系往往并不重要。**
  3. **在研究中不要过度解读控制变量的系数估计结果。**

- 在**示例 2**中，以学区为观测单位，Y 是学生的平均成绩，D 是班级的平均规模，X 是享受午餐补助的学生比例。ε 中包含学生的经济状况，它既影响 Y（反映课外学习机会的多寡），也影响 D（反映地区的财政实力）。如果控制 X 能够控制住经济状况与 D 的相关性，则可以一致地估计班级规模的因果效应。而 X 的系数很可能是负的，但这并不意味着取消午餐补助能够提高学生平均成绩，而是因为正的 β_2 被负的 δ_2 所抵消。

观测性研究的挑战：选择性

- 绝大多数社会科学数据都是观测数据，D 部分地由 ε 决定，因此 ε 的分布因 D 而异，原因在于，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**。D 的选择性或内生性，是观测性研究的根本挑战。选择性分为两种：
  - **基于可观测变量的选择性** (selection on observables)。这是指，个体是否接受某项处理，只受到可观测变量的影响，给定这些可观测变量，接受处理与否可视作近似随机。这等价于是说，给定这些可观测变量，假设**LS.2**成立。此时因果推断的关键，就是寻找这些造成选择性的可观测变量。
  - **基于不可观测变量的选择性** (selection on unobservables)。这是指，至少有一部分造成选择性的变量是不可观测的，因此无法“给定”这些变量，即假设**LS.2**不成立，无法将组间比较局限在这些变量相同的子总体内以考察因果效应。这相当于我们常说的遗漏变量问题。

- 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 $\pi \triangleq \mathrm{Pr}(D=1|X,\varepsilon)$ 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。

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

  - 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。

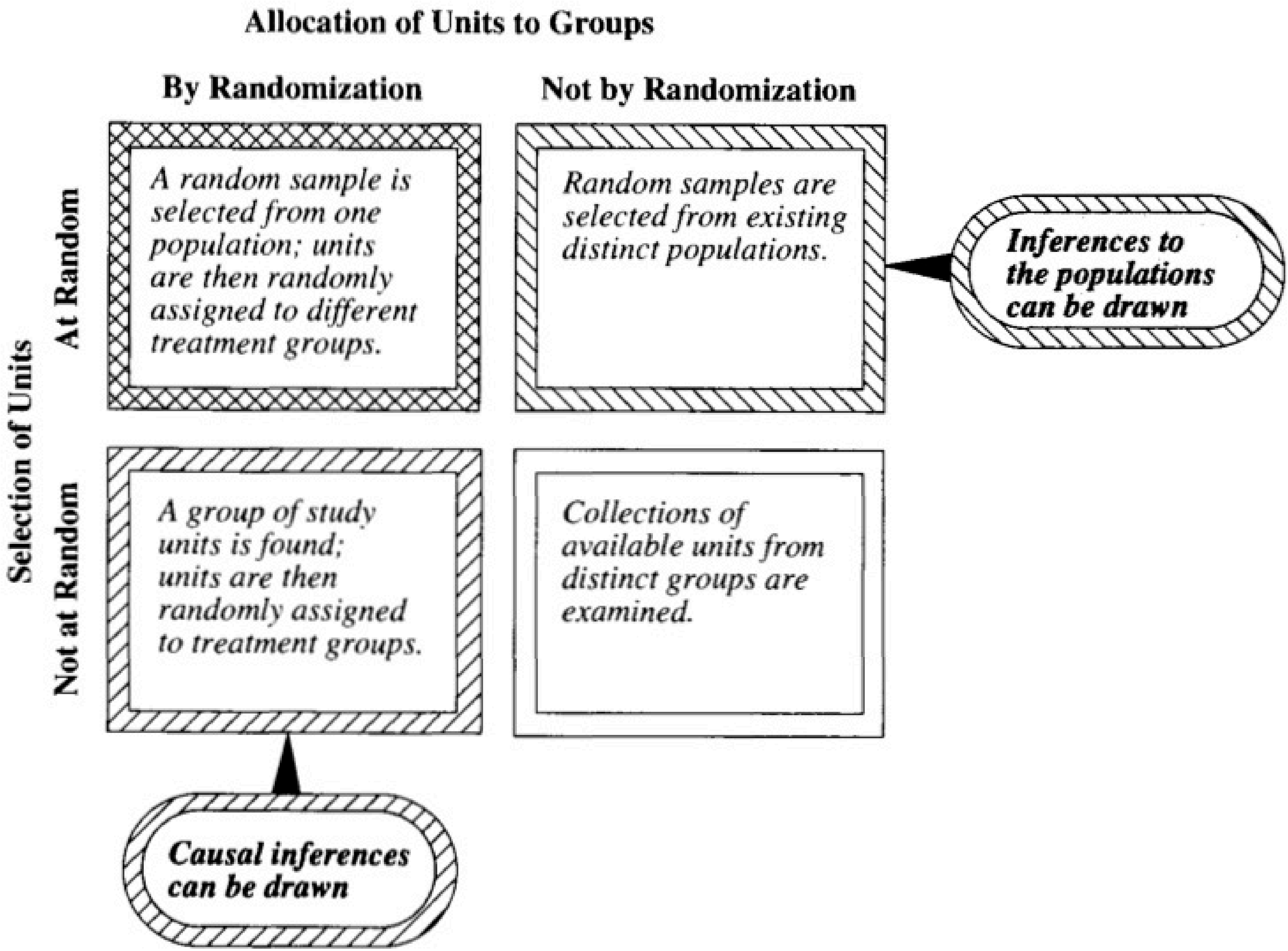
$$\mathrm{Pr}(D=1|X,\varepsilon)=\pi(X,\varepsilon)$$

  - 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)。

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

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
By Randomization
Not by Randomization
Selection of Units
At Random
Not at Random
A random sample is selected from one population; units are then randomly assigned to different treatment groups.
Random samples are selected from existing distinct populations.
Inferences to the populations can be drawn
A group of study units is found; units are then randomly assigned to treatment groups.
Collections of available units from distinct groups are examined.
Causal inferences can be drawn

D 为连续变量的一般情形

- 此时 D 被称作连续处理 (continuous treatment)。而 β_1 的含义是，保持 ε 不变，当处理强度 (treatment intensity) 变化一个单位时，结果会变化 β_1 个单位。

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

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

若

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

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 β_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，β_1 的含义是，保持 X 和 ε 不变，当处理强度变化一个单位时，结果会变化 β_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

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

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。
- 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。