因果推断，观察性研究和 2021 年诺贝尔经济学奖

苗旺

北京大学概率统计系，统计中心

mwfy@pku.edu.cn

November 6, 2021

https://www.math.pku.edu.cn/teachers/mwfy

Outline

1. Nobel Prize in Economic Sciences 2021, 1989, and 2000
2. Causal inference in statistics
3. Recent developments in confounding adjustment
4. Other fields of causal inference

Outline

Nobel Prize in Economic Sciences 2021

诺贝尔经济学奖 2021 年授予 Card, Angrist, 和 Imbens，以表彰他们在经济学的实证研究和因果推断方法方面的贡献。

- Card: for his empirical contributions to labor economics;
- Angrist and Imbens: for their methodological contributions to the analysis of causal relationships.

Labor economics: earning, labor supply and demand, employment, etc.

Scientific background

三位经济学家获奖的科学背景是观察性数据的因果推断: Answering causal questions using observational data.

探索事物之间的因果关系和因果作用是很多科学研究的重要目的。

"Most applied science is concerned with uncovering causal relationships. "
Nobel Prize in Economic Sciences (2021)

科学例子, e.g., the impact of school closures on student learning and the spread of the COVID-19 virus?
The impact of low-skilled immigration on employment and wages?
The effect of the imposition of a minimum wage affect employment? The effect of investments in education on earnings?

Scientific background

哲学上对因果关系的思考。

- 古希腊哲学家 Democritus (约公元前 400 年) 认为: 发现一个因果关系胜过做国王.
- 亚里士多德《物理学》写道:“Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it (which is to grasp its primary cause)”,
- Hume (1711-1776) 认为人类仅仅凭经验，只能认识事物之间恒定的前后相继关系，并不能认识任何因果关系。
- 培根 (1561–1626 年): “真正的知识是根据因果关系得到的知识.”
- 《墨经》: 中国古代第一个比较完整的逻辑体系，将因果关系蕴含于逻辑关系，句子 3025 个，因果复句共 290 个。例如，“夫桀无待汤之备，故放; 纣无待武之备，故杀”(七患)。

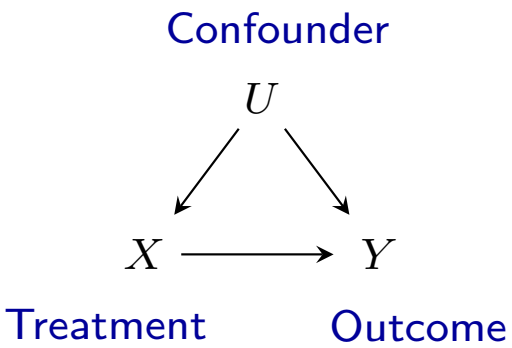
Confounding in observational studies

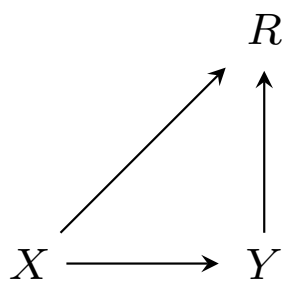
观察性研究是现代推断因果作用的主要数据来源。

Cochran & Chambers (1965): 根据经验观察推断因果作用的研究，但不能采用有控制的试验，也不能随机地分配处理。

混杂因素 (confounder)：忽略混杂因素会导致因果推断的偏差和决策的错误，甚至悖论。

选择偏差/缺失数据 (selection bias/missing data)：选择偏差导致观测数据不能代表关心的总体。





Examples: PM2.5, morbidity, other air pollutants; Genes, cancer, batch effects; 教育，收入，社会和家庭背景。

Simpson (1951) paradox

统计相关性不能代表因果关系。

一个假想的例子

	男性			女性		
	有效	无效	有效比例	有效	无效	有效比例
药	35	15	**7/10**	45	105	**3/10**
安慰剂	90	60	6/10	10	40	2/10

这种药有效吗?

	Total		
	有效	无效	有效比例
药	80	120	4/10
安慰剂	100	100	**5/10**

一种药对男人和女人都有效，但是对人类无效?

Simpson paradox, Berkson's paradox, Lord's paradox, low birth weight paradox, surrogate paradox, reverse regression paradox.

Simpson 悖论

Simpson (1951): X, Y, U 都是二值变量，Simpson 悖论是如下不等式悖论的体现.

Simpson's Reversal of Inequalities：存在一组实数，使得

$$a/b < A/B, \quad c/d < C/D,$$
$$(a+c)/(b+d) \geq (A+C)/(B+D).$$

Simpson 悖论

Simpson (1951): X, Y, U 都是二值变量，Simpson 悖论是如下不等式悖论的体现.

Simpson's Reversal of Inequalities：存在一组实数，使得

$$a/b < A/B, \quad c/d < C/D,$$
$$(a+c)/(b+d) \geq (A+C)/(B+D).$$

	$U=1$			$U=0$		
	$Y=1$	$Y=0$	比例	$Y=1$	$Y=0$	比例
$X=1$	a	b-a	a/b	c	d-c	c/d
$X=0$	A	B-A	A/B	C	D-C	C/D

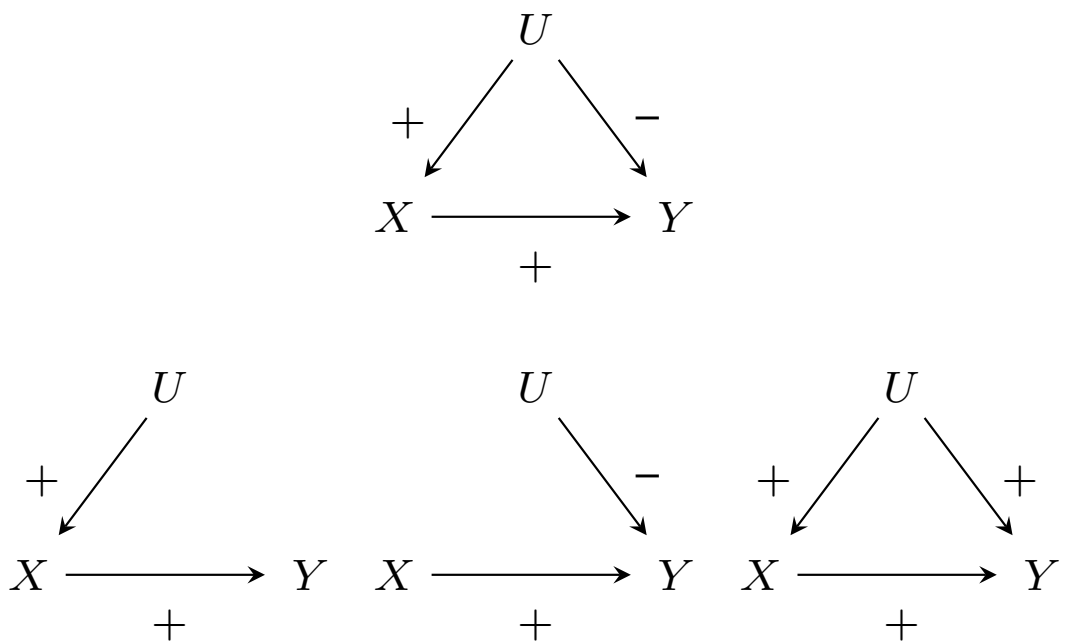
	Total		
	$Y=1$	$Y=0$	比例
$X=1$	a+c	b+d	(a+c)/(b+d)
$X=0$	A+C	B+D	(A+B)/(C+D)

Simpson 悖论出现的原因

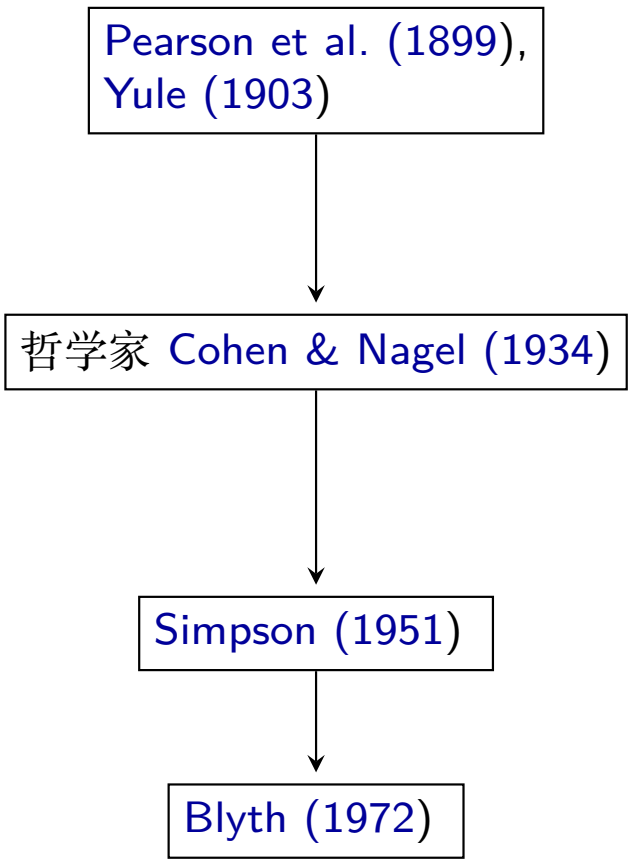
- (X, Y) 与一个共同的背景变量 U 有相关性。

  在流行病学中 U 称为混杂因素，在生物统计中称为 batch effect/unwanted variation, 在经济学中称为内生性（endogeneity）。

- (X, Y) 通过 U 产生的相关性和不通过 U 的相关性反号。



Simpson 悖论的实例



注意到了相关性由于调整第三个变量而消失的现象

注意到相关性逆转的现象
Death rates in 1910 from tuberculosis:
For African Americans, Richmond $<$ New York.
For Caucasians, Richmond $<$ New York.
For the combined population, Richmond $>$ New York

Simpson 的研究使人重视起这一现象

正式提出 Simpson 悖论这一名称

Simpson 悖论的实例

Bickel et al. (1975) 关于 1973 年伯克利研究生入学考试中是否存在性别歧视的研究. 女性的录取率低于男性的录取率，但是按照 6 个主要的专业将考生分层后，发现各个专业的女生录取率不低于男生，而且很多专业高于男生。

出现这个现象的原因是男生普遍选择了容易录取的专业，专业是一个混杂因素。

	录取	未录取	总和	录取率
男生	3738	4704	8442	**44%**
女生	1494	2827	4321	35%

专业	A	B	C	D	E	F
男生申请人数	825	560	325	417	191	373
录取率	62%	63%	37%	33%	28%	6%
女生申请人数	108	25	593	375	393	341
录取率	82%	68%	34%	35%	24%	7%
申请人数之和	933	585	918	792	584	714

Simpson 悖论的实例

Wagner (1982) 描述了美国在 1974 年和 1978 年的税收。比较 1974 年的总税率 0.141 和 1978 年的总税率 0.152，似乎税率提高了 1.1%。但是，可以看出在各个收入水平上，1978 年的税率比 1974 年的税率都低。

Year	Level	Income	Tax	Tax Rate
1974	0–5k	41,651,643	2,244,467	0.054
	5k–10k	146,400,740	13,646,348	0.093
	10k–15k	192,688,922	21,449,597	0.111
	15k–100k	470,010,790	75,038,230	0.160
	>100k	29,427,152	11,311,672	0.384
	Total	880,179,247	123,690,314	0.141
1978	0–5k	19,879,622	689,318	0.035
	5k–10k	122,853,315	8,819,461	0.072
	10k–15k	171,858,024	17,155,758	0.100
	15k–100k	865,037,814	137,860,951	0.159
	>100k	62,806,159	24,051,698	0.383
	Total	1,242,434,934	188,577,186	**0.152**

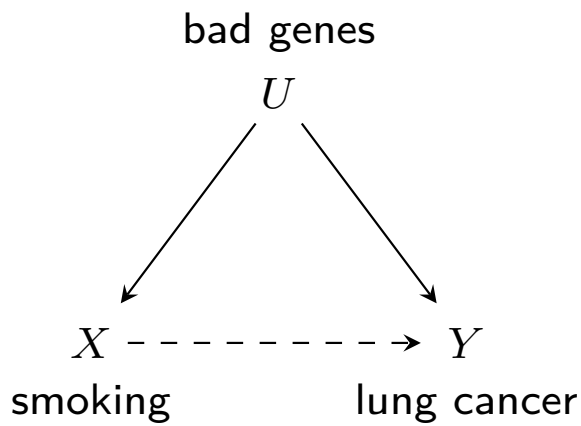
Simpson 悖论的实例

Doll 和 Hill 在 1948–1949 年对伦敦的 20 所医院进行了吸烟与肺癌的病例对照研究, 发现吸烟与肺癌有显著的关联。

Fisher (1958) 指出相关关系不能简单地解释为吸烟与肺癌之间的因果关系, 吸烟与肺癌的相关性也可以由另外两个备择理论来解释:

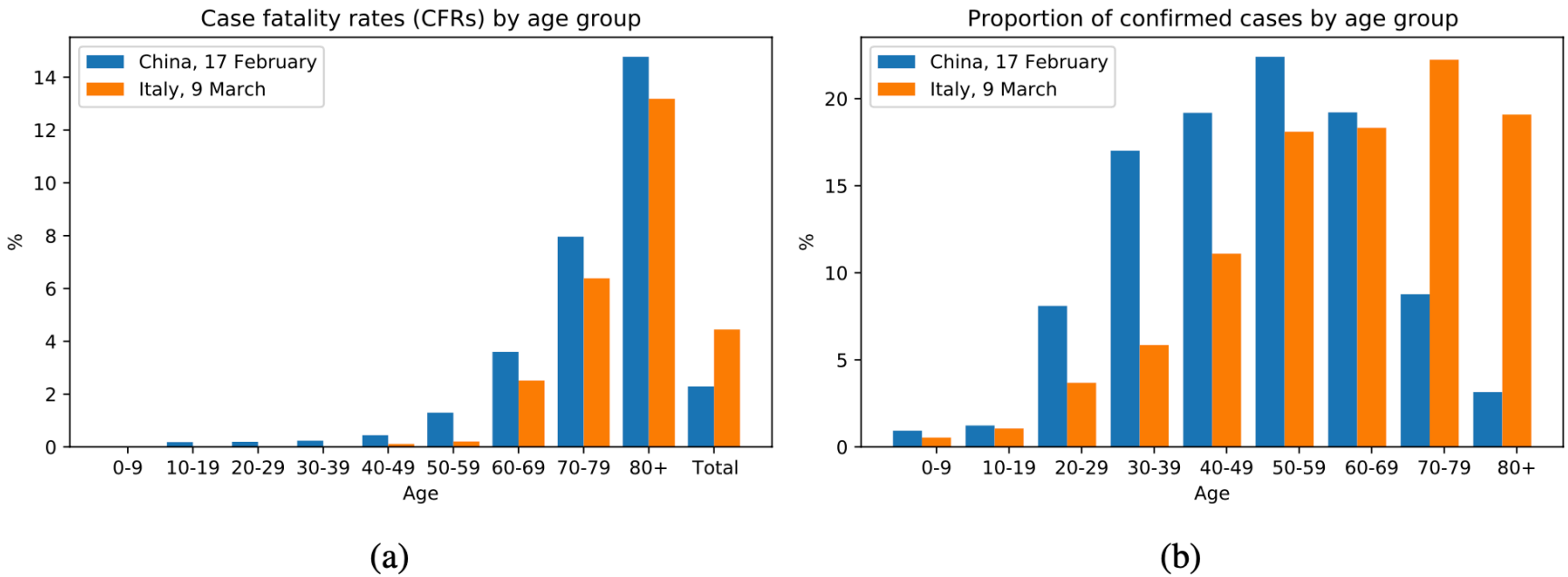
1. 癌症引起吸烟;

2. 存在基因，它们既引起癌症, 又引起吸烟。



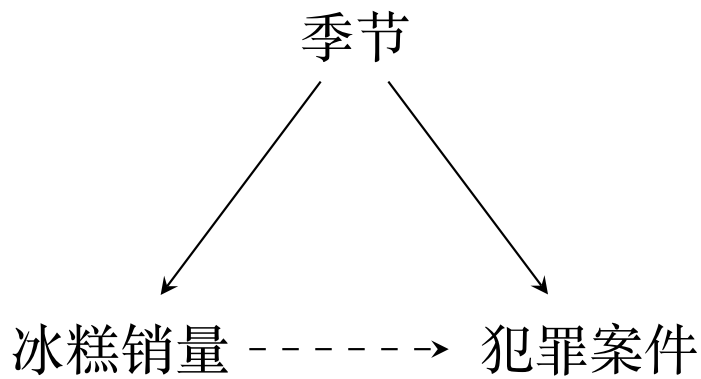
Simpson 悖论的实例

von Kügelgen et al. (2020) 新冠肺炎疫情中的 Simpson 悖论例子。



Simpson 悖论的实例

生活或研究领域中的 Simpson 悖论例子



Simpson 悖论的意义

- Simpson 悖论对使用相关性刻画因果关系提出挑战：边际相关性和分组相关性哪个代表真正的因果关系？
- 如果相信分组后的相关性代表因果关系，那么经过第四个变量分组后，相关性可能再次发生逆转。
- 人们无法知道确切的混杂因素，也不可能测量所有的混杂因素，只要有一个未被测量，就会对因果推断造成极大偏差。
- 相关性不能代表因果关系。因果关系如何定义？

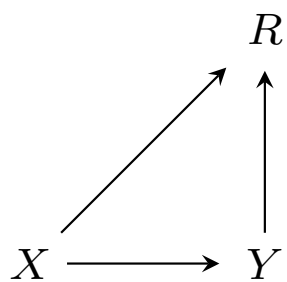
> “Problems involving causal inference have dogged the heels of statistics since its earliest days.”(Holland et al., 1986)
>
> “I am convinced that many discoveries have been delayed in our century for lack of a mathematical language that can handle causation.” (Pearl, 2009, page 427)

Berkson (1946) paradox

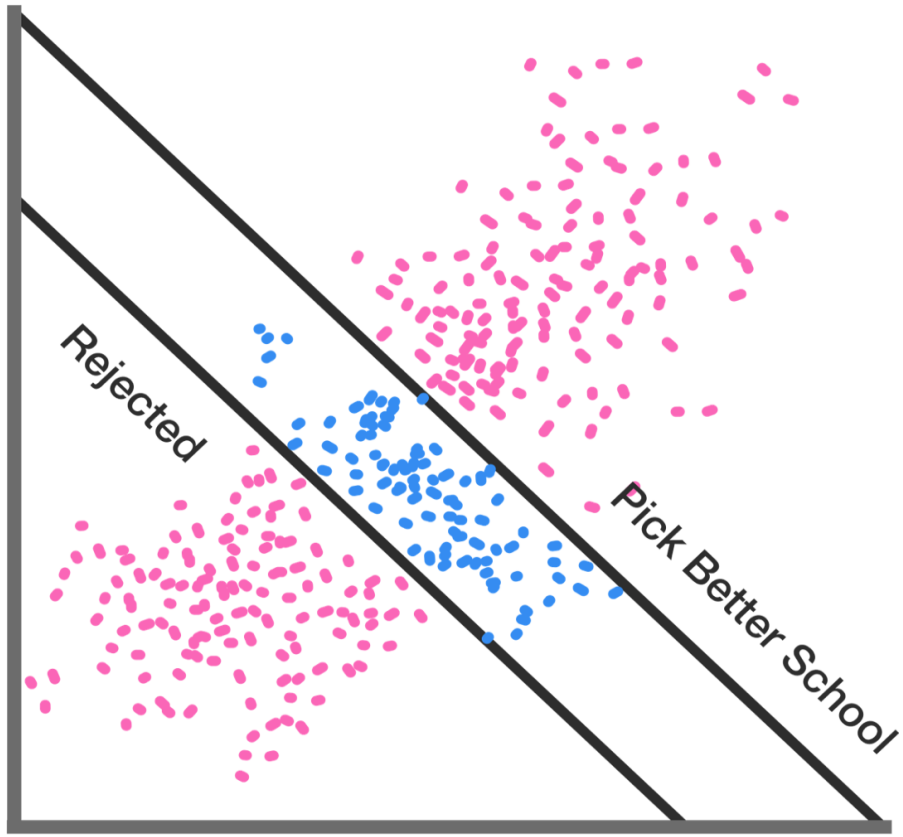
Berkson 悖论，也称为 Berkson 偏差，1946 年由医生 Berkson 提出。Berkson 从数学上解释了在医院中进行的病例–对照研究可能由于住院率导致的偏差。

Berkson 构造了一个例子，基于住院的患 diabetes 和 refractive errors 两类病人，研究 diabetes 和 cholecystitis 的关系。住院病人的数据表明 diabetes 和 cholecystitis 有正的相关性，即，cholecystitis 是 diabetes 的一个潜在危险因素，但实际上在一般人群中，这个相关性可能不存在。

Roberts et al. (1978) 关于医疗服务的调查提供了 Berkson 悖论在实际应用中的支持。



Berkson (1946) paradox



语文和数学成绩是否负相关?

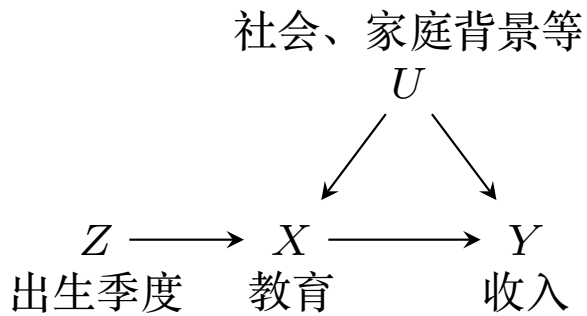
Ellenburg, Jordan. Why Are Handsome Men Such Jerks?

Instrumental variable

Wright (1928); Goldberger (1972); Angrist et al. (1996); Angrist & Krueger (1991); Angrist & Evans (1998)

Instrumental variable: (i) exclusion restriction, $Z \perp\!\!\!\perp Y \mid (X, U)$; (ii) independence $Z \perp\!\!\!\perp U$; (iii) correlation $Z \not\perp\!\!\!\perp X$.

$$E(Y \mid X, U) = \gamma X + U, \quad \gamma^{\text{iv}} = \frac{\widehat{\sigma}_{zy}}{\widehat{\sigma}_{xz}}$$



Examples: fertility, parental labor supply, and gender composition of children; military service, earnings, and draft eligibility; SNPs, genes, and cancer.

在观察性研究中找到一个有效工具变量很困难，因此，人们质疑使用工具变量能否作为推断因果作用的一个普遍方式。

Card and natural experiments

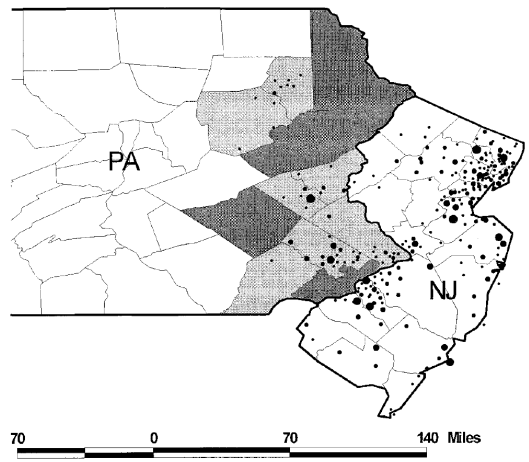
Card 与合作者使用一些自然试验分析劳动经济学中一系列重要的因果问题，重塑或加深了人们对劳动经济学中一些百年难题因果关系的认识，例如，如发现提高最低工资并不会减少就业。

并且，自然试验在劳动经济学中的成功运用也促使工具变量，重差法等成为实证研究中推断因果作用的普遍范式。

自然试验—大自然的随机化试验，是不受研究者控制的，自然发生的或政策改变等对研究的变量有类似于随机化试验影响的事件，如出生日期，基因突变，自然灾害，战争，或禁烟政策等。

The employment effects of the minimum wage

- Card & Krueger (1994) analyzed the effect of a minimum wage increase in New Jersey using the differences-in-differences method
- In February 1992, New Jersey increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.



- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase.

Difference in differences: Removing confounding with parallel trend.

$$Y_{it}(0) = \gamma_i + \lambda_t + \varepsilon_{it}$$

The DID design is usually attributed to John Snow (1855).

Structural equation model

在很长一段时间里，经济学家使用工具变量推断因果作用的主要依赖线性模型等结构方程模型 (structural equation model, Wright, 1928; Haavelmo, 1943, 1944; Goldberger, 1972)，

$$E(Y \mid X, U) = \gamma X + U,$$

结构方程模型在形式上与回归模型相似，但结构方程模型非常隐晦地包含了刻画因果关系需要的假定，以至于经常被和表示相关关系的回归模型混为一谈，而其中的因果假定难以表示和验证 (Bollen & Pearl, 2013)。

结构方程模型容易设定错误 (LaLonde, 1986)。

Potential outcome framework

统计学家提出使用潜在结果定义因果作用，潜在结果有更强的表示能力，可以更直接和清楚地定义因果作用和表述因果假定。

Potential outcome framework (Neyman, 1923; Rubin, 1974):

Y outcome, X exposure, U confounder, observed covariates are suppressed

- Potential outcome: $Y(x)$;
- $Y = Y(x)$ when $X = x$;
- Causal effect: A comparison of $Y(x)$ across different treatment levels.
  Average causal effect (ACE): $E\{Y(x) - Y(x')\}$
  Identification challenge: $Y(1)$ and $Y(0)$ can never be jointly observed.

Potential outcome framework

- 波兰华沙大学的 Neyman (1923) 在博士毕业论文"On the Application of Probability Theory to Agricultural Experiments"中提出了潜在结果 (potential outcomes) 的数学模型。

  考虑一个检测两种肥料对于产量影响的农业实验，用 n 块田做实验，$Y_i(1)$ 和 $Y_i(0)$ 表示如果第 i 块田用肥料 1 和肥料 0 分别对应的产量，那么 $Y_i(1) - Y_i(0)$ 就是肥料 1 相对于肥料 0 对第 i 块田产量的因果作用。随机地分配肥料 1 或者肥料 0 到第 i 块田，$E\{Y_i(1) - Y_i(0)\}$ 称为平均因果作用 (average causal effect)。

- 在观察性研究中，因为没有随机化，处理组和对照组不仅有可比性，Neyman 本人对观察性研究的潜在结果模型持怀疑态度，

  Rubin 认为，观察性研究也对应着一个假想的随机化实验，因此潜在结果模型可以用来定义一般的因果作用。

  Ignorability: $Y(x) \perp\!\!\!\perp X \mid U$

$$f\{Y(x) = y\} = \sum_u f(Y = y \mid u, x) f(u)$$

Angrist and Imbens and LATE

Angrist、Imbens 和合作者将工具变量与潜在结果模型结合 (Imbens & Angrist, 1994; Angrist et al., 1996)，使用潜在结果模型刻画工具变量假定和相应的统计模型，定义新的因果概念，发展新的统计推断方法。

Local average treatment effect: Individuals are affected differently by the treatment and choose whether to comply with the assignment; e.g., IV estimates of the earnings return to schooling were higher than OLS.

Assumption 1. Randomization, $Z \perp\!\!\!\perp Y(z,x)$.
Exclusion Restriction, $Y(z',x) = Y(z,x)$.
Relevance $E[X(1) - X(0)] \neq 0$.
Monotonicity, $X_i(1) \geq X_i(0)$

Complier average causal effect or local average causal effect

$$
\begin{aligned}
E[Y(1) - Y(0) \mid X(1) - X(0) = 1] &= \frac{E[Y(1, X(1)) - Y(0, X(0))]}{E[X(1) - X(0)]} \\
&= \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(X \mid Z = 1) - E(X \mid Z = 0)}.
\end{aligned}
$$

Nobel Economics Prize in 1989 and 2000

1989 年 Haavelmo 和 2000 年 Heckman 获诺贝尔奖的主要贡献都与因果研究密切相关。

- Haavelmo 将数理统计引入经济学 (Nobel Prize in Economic Sciences, 1989)，明确经济学模型如联立方程组的因果意义，为计量经济学做出奠基性的工作，被称为计量经济学之父。
- Heckman 的选择模型 (Nobel Prize in Economic Sciences, 2000) 对观察性研究处理缺失数据和选择偏差，以及因果推断消除混杂因素影响非常深远。
- 2003 年 Granger 因为在非线性时间序列方面的贡献获得诺贝尔经济学奖，但著名的 Granger causality 更接近预测问题，而被认为是过时的和不恰当的因果模型。

Connections between causal inference and missing data analysis

Table 1: Data structure of potential outcomes for a binary treatment

	Full data			Observed data			
ID	$Y_i(1)$	$Y_i(0)$	X	$Y_i(1)$	$Y_i(0)$	Y	X
1	$Y_1(1)$	$Y_1(0)$	1	$Y_1(1)$	?	$Y_1(1)$	1
2	$Y_2(1)$	$Y_2(0)$	1	$Y_2(1)$	?	$Y_2(1)$	1
3	$Y_3(1)$	$Y_3(0)$	0	?	$Y_3(0)$	$Y_3(0)$	0
4	$Y_4(1)$	$Y_4(0)$	0	?	$Y_4(0)$	$Y_4(0)$	0
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
n	$Y_n(1)$	$Y_n(0)$	1	$Y_n(1)$	?	$Y_n(1)$	1

Outline

Three aspects of causal inference

- 什么是因果作用; structural equation model; causal diagram model; potential outcome framework;
- 如何推断因果作用; RCT; ignorability, IPW, REG, DR estimation; sensitivity analysis; instrumental variable; difference in difference; regression discontinuity; synthetic control; proximal inference;
- 因果作用有什么用; policy; prediction; interference; optimal treatment regime; transfer learning;

Mathematical languages for causal inference

"I am convinced that many discoveries have been delayed in our century for lack of a mathematical language that can handle causation." (Pearl, 2009, page 427)

- Structural equation model (Wright, 1928; Haavelmo, 1943, 1944)
- causal diagram model (Pearl, 1995)
- potential outcome framework (Neyman, 1923; Rubin, 1974)

These three models are mathematically equivalent (Galles & Pearl, 1998).

Inference on causal effects

- Fisher and randomized experiment (Fisher, 1937).
  随机化试验是推断因果作用的金标准: $X \perp\!\!\!\perp Y(x)$,

$$E\{Y(x)\} = E(Y \mid X = x),$$

$$ACE = E(Y \mid X = 1) - E(Y \mid X = 0).$$

- Rubin and Rosenbaum and ignorability (Rosenbaum & Rubin, 1983b).
- Definition of confounders (Greenland et al., 1999; Geng et al., 2002; VanderWeele & Shpitser, 2013)
- Pearl and causal structure learning (Pearl, 2009; He & Geng, 2008; Xie et al., 2006; Xie & Geng, 2008).
- Robins and time-varying confounding and dynamic treatment regime; A-learning. (Robins, 1986, 2004)

Instrumental variable approach in statistics

- Bounds (Robins, 1989; Balke & Pearl, 1997);
- LATE (Angrist et al., 1996);
- Instrumental inequality (Balke & Pearl, 1997);
- Surrogate/instrumental variable paradox (Chen et al., 2007; Ju & Geng, 2010; VanderWeele, 2013).

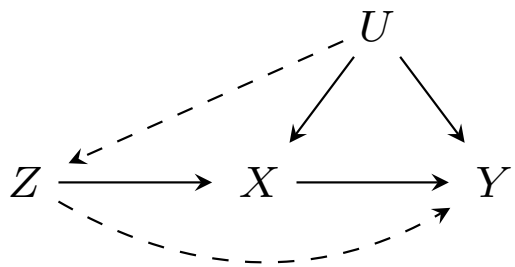
Outline

Confounding in causal revolution

"...the complete solution of the confounding problem one of the main highlights of the Causal Revolution because it ended an era of confusion that has probably resulted in many wrong decisions in the past." (Pearl & Mackenzie, 2018).

Nonparallel trends and invalid controls in DID

Weak IV, invalid IV



New methods for confounding adjustment

- New IV methods: Quantile and non-separable models (Chernozhukov & Hansen, 2005; Horowitz, 2009)
- Sensitivity analysis (Cornfield et al., 1959; Rosenbaum & Rubin, 1983a)
- Synthetic control (Abadie et al., 2010)
- Regression discontinuity (Lee & Lemieux, 2010; Thistlethwaite & Campbell, 1960)
- Proximal inference/ negative controls (Miao et al., 2018; Shi et al., 2020; Tchetgen Tchetgen et al., 2020)
- Multi-treatment confounding (Wang & Blei, 2019; Miao et al., 2020)

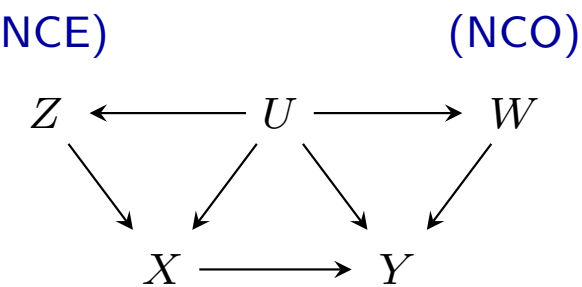
Proximal inference/ negative controls

Miao & Tchetgen Tchetgen (2017); Miao et al. (2018); Miao & Tchetgen Tchetgen (2018); Shi et al. (2020); Tchetgen Tchetgen et al. (2020)

Acknowledging covariate measurements as imperfect proxies of confounding mechanisms, covariates are decomposed to three parts: treatment associated Z, outcome associated W, and both associated C.

Exclusion restrictions:

- $W \perp\!\!\!\perp X \mid U$, $W \not\perp\!\!\!\perp U$
- $Z \perp\!\!\!\perp Y \mid (X, U)$, $Z \not\perp\!\!\!\perp U \mid X$



Negative control examples

NCO: $W \perp\!\!\!\perp X \mid U$, $W \not\perp\!\!\!\perp U$

- Trichopoulos et al. (1983) Acute stress X, deaths from heart disease Y, deaths from cancer W

  Khush et al. (2013) E. coli in water X, diarrhea Y, respiratory symptoms W

  Gagnon-Bartsch & Speed (2012); Eisenberg & Levanon (2003) Housekeeping genes

NCE: $Z \perp\!\!\!\perp Y \mid (X, U)$, $Z \not\perp\!\!\!\perp U \mid X$

- Davey Smith (2008, 2012) paternal smoking

  Flanders et al. (2017); Miao & Tchetgen Tchetgen (2017) future air pollution levels

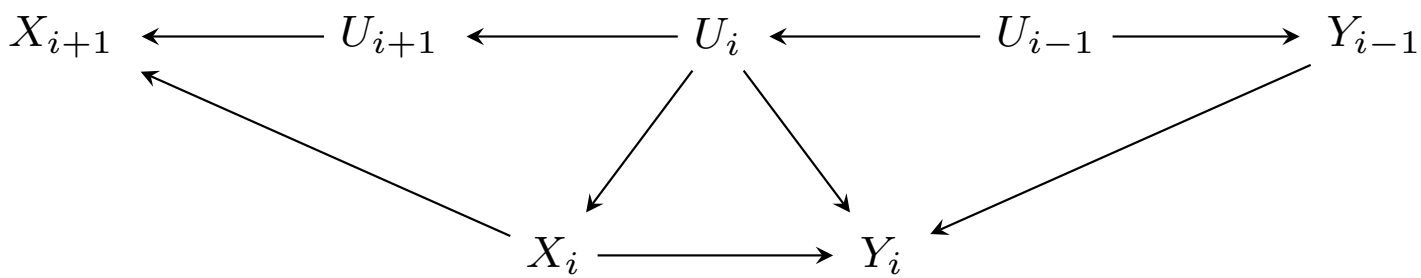
- Instrumental variable

Timeseries setting

Consider a time-series model, e.g., an air pollution study,

$$U_i = \xi U_{i-1} + (1-\xi^2)^{1/2}\varepsilon_{1i}, \quad V_i = 0.6U_i + \varepsilon_{2i},$$
$$X_i = 0.4 + 1.5V_i + \eta U_i + \varepsilon_{3i}, \quad Y_i = 0.5 + 0.7X_i + 1.5V_i + 0.9U_i + \varepsilon_{4i},$$

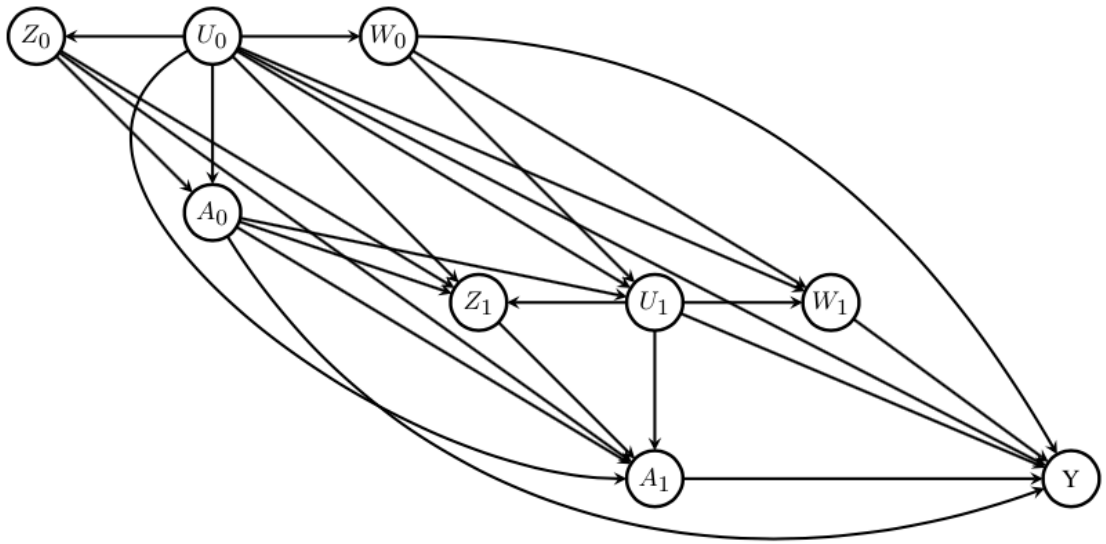
- $Z = X_{i+1}$, $W = Y_{i-1}$



- Note: Estimate the causal effect with only (X_i, Y_i)!

Extensions

Ying, Miao, Shi, and Tchetgen Tchetgen. (2021). Proximal Causal Inference for Complex Longitudinal Studies.



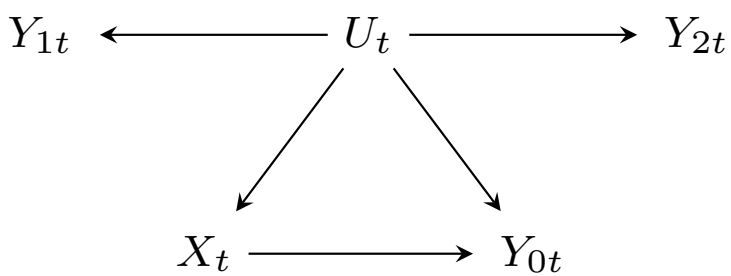
Zhengling Qi, Rui Miao, Xiaoke Zhang. (2021). Proximal Learning for Individualized Treatment Regimes Under Unmeasured Confounding.

Proximal inference framework for synthetic control

Difference in differences: $Y_{it}(0) = \gamma_i + \lambda_t + \varepsilon_{it}$; unit $i = 0$ is treated, using $i = 1$ as control for $i = 0$;

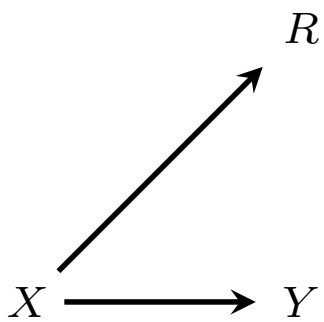
Synthetic control: $Y_{it}(0) = \gamma_i + \lambda_t + \theta_i U_t + \varepsilon_{it}$; unit $i = 0$ is treated, weighting $i \geq 1$ to create a synthetic control $\sum \alpha_{i \geq 1} Y_{it}$; estimating α by OLS on pretreatment data, which is biased.

Shi, Miao, Hu, Tchetgen Tchetgen. (2021). Theory for identification and Inference with Synthetic Controls: A Proximal Causal Inference Framework

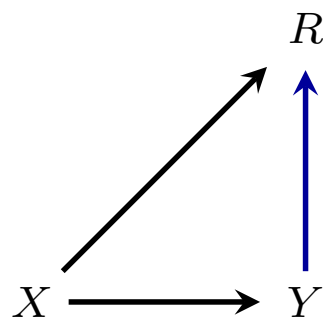


Missing data

- Y outcome of interest with missing values; X a vector of covariates; $R = 1$ for observed, 0 otherwise;
- The interest is some functional of $f(Y, X)$, e.g., $E(Y)$
- Missingness mechanisms (Rubin, 1976; Little & Rubin, 2002)
  - Missing at random (MAR) $R \perp\!\!\!\perp Y \mid X$;
  - Missing not at random (MNAR) $R \not\perp\!\!\!\perp Y \mid X$;



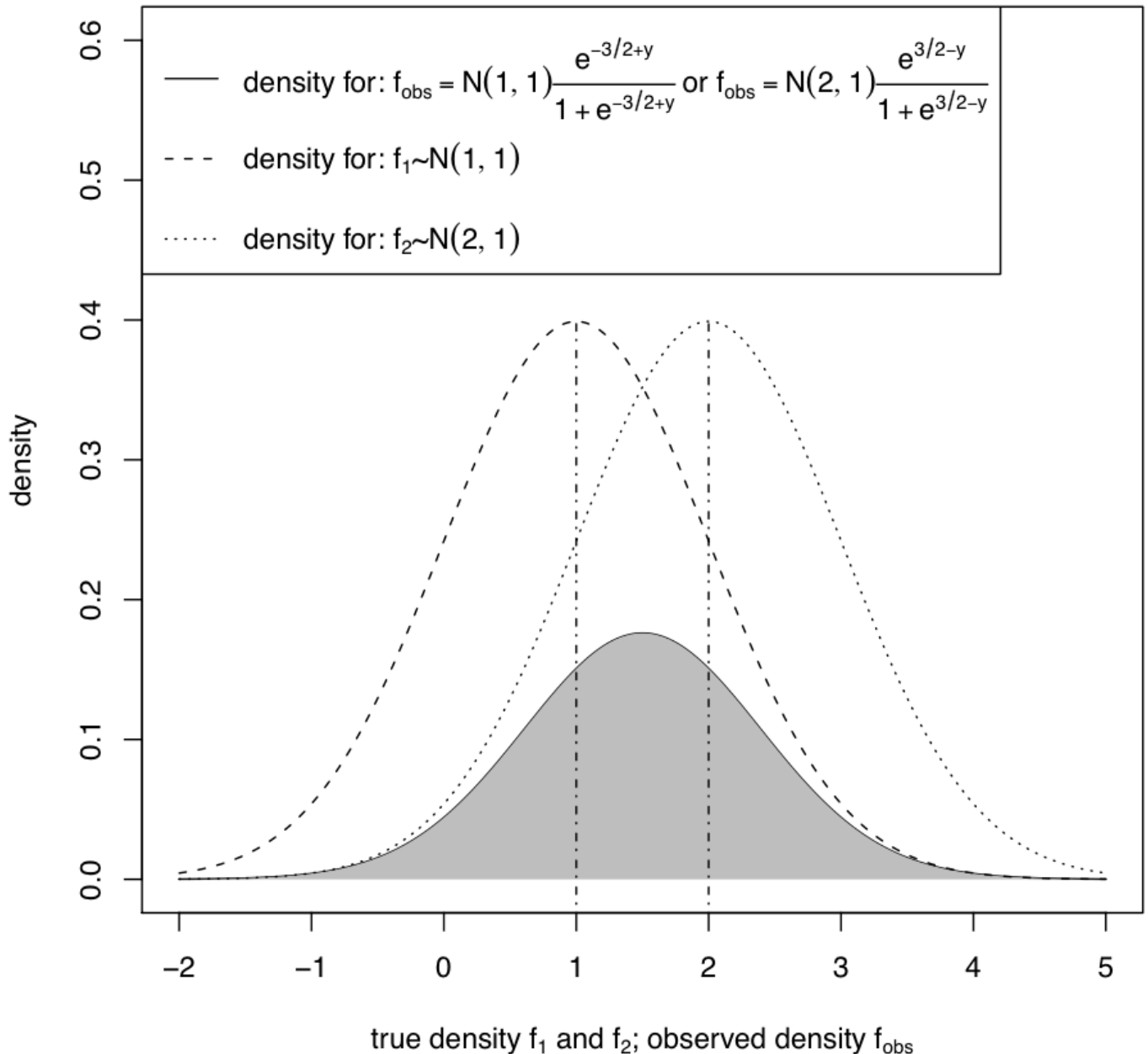
MAR



MNAR

The identification Difficulty

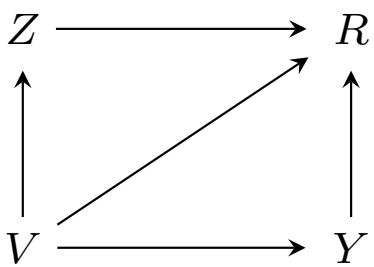
$Y \sim N(\mu, \sigma^2)$, logit $P(R = 1|Y) = \alpha + \beta Y$.

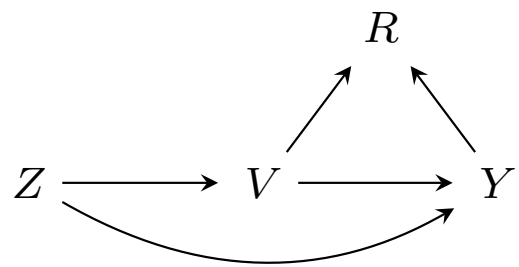


Nonignorable missing data analysis

I will not review the huge literature of missing data analysis.

- Identification of normal and normal mixture models with missing data (Miao et al., 2016);
- Instrumental variable approach (Liu et al., 2020; Sun et al., 2018);
- Shadow variable approach (Miao & Tchetgen Tchetgen, 2016; Miao et al., 2019);





- Callback design for nonresponse adjustment (Miao et al., 2021)

Callback design for nonresponse adjustment

Callback data: in many surveys interviewers continue to contact nonrespondents and the contact attempts are recorded.

Table 2: Data structure of a sampling survey with callbacks

Frame/Questionnaire data				Contact attempts			
ID	X	Y	R	R_1	R_2	$\dots$	R_K
1	x_1	y_1	1	1	1	$\cdots$	1
2	x_2	y_2	1	0	1	$\cdots$	1
:	:	NA	0	0	0	$\cdots$	0
:	:	NA	0	0	0	$\cdots$	0
n	x_n	y_n	1	0	0	$\cdots$	1

- We propose an identifying assumption that allows for nonparametric and nonlinear propensity score models,
- establish the semiparametric theory
- and propose a suite of semiparametric estimators including doubly robust ones.

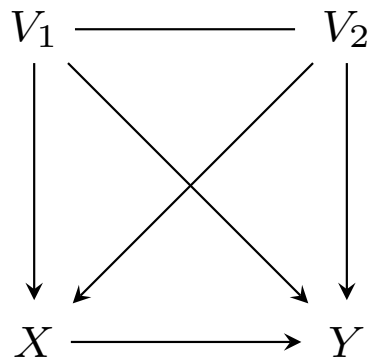
Outline

Other fields of causal inference

- Interference (Hudgens & Halloran, 2008; Liu & Hudgens, 2014; Tchetgen Tchetgen & VanderWeele, 2010);
- Mediation analysis and causal mechanism (Robins & Greenland, 1992; Pearl, 2001; VanderWeele et al., 2014)
- Individualized treatment regime (Qian & Murphy, 2011; Zhao et al., 2012)
- Data fusion (Chatterjee et al., 2016; Li et al., 2020, 2021; Sun & Miao, 2021)

Data fusion

Li et al. (2020): V_1, V_2 are observed in different datasets



Li et al. (2021): Randomized trials conducted in $H = 0$, historical controls available from $H \geq 1$.

Causal inference and AI research

2011 年图灵奖获得者 Pearl: “To Build Truly Intelligent Machines, Teach Them Cause and Effect.”

2018 年图灵奖获得者 Bengio 和 LeCun 认为“It’s a big thing to integrate causality into AI.”

Connections

- Data fusion, missing data and transfer learning, domain adaptation, semisupervised learning.
- Dynamic treatment regime and reinforcement learning.
- Individualized treatment regime and classification.
- Semiparametrics and double-debiased machine learning

Causal inference and AI research

Gaps

Keywords for ICLR 2022 submissions.

Summary

- Haavelmo, Heckman, Card, Angrist 和 Imbens 推动了经济学中的因果推断研究。
- 在工具变量研究和因果推断整个领域，统计学家做出了全方位的和首屈一指的贡献。McFadden 和 Heckman 因为在 discrete choice models and selection bias 方面的贡献获 2000 年诺贝尔经济学奖，而统计学家在流行病学和生物医学中关于缺失数据, case control 的工作是平行的，但这些贡献被轻视 (Breslow, 2003)。
- 混杂因素和缺失数据仍然是因果推断和观察性研究的重点问题。
- 因果推断和机器学习，人工智能研究有一些可结合的方向，但仍然有大的鸿沟。
- 在理论创新的同时，积极运用因果推断解决现实世界的问题，为科学和社会进步做出更大的贡献。

谢谢!

References

Abadie, A., Diamond, A. & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association* **105**, 493–505.

Angrist, J. D. & Evans, W. N. (1998). Children and their parents' labor supply: evidence from exogenous variation in family size. *The American Economic Review* **88**, 450–477.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association* **91**, 444–455.

Angrist, J. D. & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* **106**, 979–1014.

Balke, A. & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* **92**, 1171–1176.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2**, 47–53.

References

Bickel, P. J., Hammel, E. A. & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science* **187**, 398–404.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* **67**, 364–366.

Bollen, K. A. & Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research*. Springer, pp. 301–328.

Breslow, N. E. (2003). Are statistical contributions to medicine undervalued? *Biometrics* **59**, 1–8.

Card, D. & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review* **84**, 772–793.

Chatterjee, N., Chen, Y.-H., Maas, P. & Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111**, 107–117.

References

Chen, H., Geng, Z. & Jia, J. (2007). Criteria for surrogate end points. *Journal of the Royal Statistical Society: Series B* **69**, 919–932.

Chernozhukov, V. & Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica* **73**, 245–261.

Cochran, W. G. & Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A* **128**, 234–266.

Cohen, M. F. & Nagel, E. (1934). *An Introduction to Logic and Scientific Method*. New York: Harcourt, Brace and Co.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22**, 173–203.

Davey Smith, G. (2008). Assessing intrauterine influences on offspring health outcomes: can epidemiological studies yield robust findings? *Basic & Clinical Pharmacology & Toxicology* **102**, 245–256.

Davey Smith, G. (2012). Negative control exposures in epidemiologic studies. *Epidemiology* **23**, 350–351.

References

Eisenberg, E. & Levanon, E. Y. (2003). Human housekeeping genes are compact. *TRENDS in Genetics* **19**, 362–365.

Fisher, R. A. (1937). *The design of experiments*. London: Oliver And Boyd.

Fisher, R. A. (1958). Cancer and smoking. *Nature* **182**, 596–596.

Flanders, W. D., Strickland, M. J. & Klein, M. (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *American Journal of Epidemiology* **185**, 941–949.

Gagnon-Bartsch, J. A. & Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539–552.

Galles, D. & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science* **3**, 151–182.

Geng, Z., Guo, J. H. & Fung, W. (2002). Criteria for confounders in epidemiological studies. *J. R. Stat. Soc. B* **64**, 3–15.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica* **40**, 979–1001.

References

Greenland, S., Robins, J. M. & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica* , iii–115.

He, Y.-B. & Geng, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.* **9**, 2523–2547.

Holland, P. W., Glymour, C. & Granger, C. (1986). Statistic and causal inference. *Journal of the American Statistical Association* **81**, 945–960.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*, vol. 12. Springer.

Hudgens, M. G. & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association* **103**, 832–842.

Imbens, G. W. & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467–475.

References

Ju, C. & Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *J. R. STAT. SOC. B* **72**, 129–142.

Khush, R. S., Arnold, B. F., Srikanth, P., Sudharsanam, S., Ramaswamy, P., Durairaj, N., London, A. G., Ramaprabha, P., Rajkumar, P., Balakrishnan, K. et al. (2013). H2S as an indicator of water supply vulnerability and health risk in low-resource settings: a prospective cohort study. *The American Journal of Tropical Medicine and Hygiene* **89**, 251–259.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* **76**, 604–620.

Lee, D. S. & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* **48**, 281–355.

Li, H., Miao, W., Cai, Z., Liu, X., Zhang, T., Xue, F. & Geng, Z. (2020). Causal data fusion methods using summary-level statistics for a continuous outcome. *Statistics in Medicine* **39**, 1054–1067.

Li, X., Miao, W., Lu, F. & Zhou, X.-H. (2021). Improving efficiency of inference in clinical trials with external control data. *Biometrics* , in press.

References

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley: New York.

Liu, L. & Hudgens, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association* **109**, 288–301.

Liu, L., Miao, W., Sun, B., Robins, J. & Tchetgen, E. T. (2020). Identification and inference for marginal average treatment effect on the treated with an instrumental variable. *Statistica Sinica* **30**, 1517–1541.

Miao, W., Ding, P. & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Miao, W., Geng, Z. & Tchetgen Tchetgen, E. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105**, 987–993.

Miao, W., Hu, W., Ogburn, E. L. & Zhou, X. (2020). Identifying effects of multiple treatments in the presence of unmeasured confounding .

Miao, W., Li, X. & Sun, B. (2021). The role of callback in survey data for nonresponse adjustment .

References

Miao, W., Liu, L., Tchetgen, E. T. & Geng, Z. (2019). Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556* .

Miao, W. & Tchetgen Tchetgen, E. (2017). Invited commentary: Bias attenuation and identification of causal effects with multiple negative controls. *American Journal of Epidemiology* **185**, 950–953.

Miao, W. & Tchetgen Tchetgen, E. (2018). A confounding bridge approach for double negative control inference on causal effects .

Miao, W. & Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **103**, 475–482.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. *Translated in Statistical Science* **5**, 465–480 (1990).

Nobel Prize in Economic Sciences (1989). The sveriges riksbank prize in economic sciences in memory of alfred nobel 1989. *https://www.nobelprize.org/prizes/economic-sciences/1989/haavelmo/facts/* .

References

Nobel Prize in Economic Sciences (2000). The sveriges riksbank prize in economic sciences in memory of alfred nobel 2000. *https://www.nobelprize.org/prizes/economic-sciences/2000/heckman/facts* .

Nobel Prize in Economic Sciences (2021). Scientific background on the sveriges riksbank prize in economic sciences in memory of alfred nobel 2021: Answering causal questions using observational data. *https://www.nobelprize.org/uploads/2021/10/advanced-economicsciencesprize2021.pdf* .

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82**, 669–688.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th conference on Uncertainty in artificial intelligence*. San Francisco, CA.: Morgan Kaufmann, pp. 411–420.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press, 2nd ed.

Pearl, J. & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic books.

References

Pearson, K., Lee, A. & Bramley-Moore, L. (1899). Genetic (reproductive) selection: Inheritance of fertility in man, and of fecundity in thoroughbred racehorses. *Philosophical Transactions of the Royal Society of London, Series A* **192**, 257–330.

Qian, M. & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics* **39**, 1180–1210.

Roberts, R. S., Spitzer, W. O., Delmore, T. & Sackett, D. L. (1978). An empirical demonstration of berkson's bias. *Journal of chronic diseases* **31**, 119–128.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: A Focus on AIDS* , 113–159.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. Springer.

Robins, J. M. & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3**, 143–155.

References

Rosenbaum, P. R. & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B* **45**, 212–218.

Rosenbaum, P. R. & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika* **63**, 581–592.

Shi, X., Miao, W., Nelson, J. C. & Tchetgen Tchetgen, E. (2020). Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B* **82**, 521–540.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B* **13**, 238–241.

References

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. & Tchetgen, E. J. T. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica* **28**, 1965.

Sun, B. & Miao, W. (2021). On semiparametric instrumental variable estimation of average treatment effects through data fusion. *Statistica Sinica* , in press.

Tchetgen Tchetgen, E. & VanderWeele, T. (2010). On causal inference in the presence of interference. *Statistical Methods in Medical Research* **21**, 55–75.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X. & Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982* .

Thistlethwaite, D. L. & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology* **51**, 309.

Trichopoulos, D., Zavitsanos, X., Katsouyanni, K., Tzonou, A. & Dalla-Vorgia, P. (1983). Psychological stress and fatal heart attack: The athens (1981) earthquake natural experiment. *The Lancet* **321**, 441–444.

VanderWeele, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics* **69**, 561–565.

References

VanderWeele, T. J. & Shpitser, I. (2013). On the definition of a confounder. *Annals of statistics* **41**, 196–220.

VanderWeele, T. J., Vansteelandt, S. & Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* **25**, 300–306.

von Kügelgen, J., Gresele, L. & Schölkopf, B. (2020). Simpson's paradox in covid-19 case fatality rates: a mediation analysis of age-related causal effects. *arXiv preprint arXiv:2005.07180* .

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician* **36**, 46–48.

Wang, Y. & Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association* **114**, 1574–1596.

Wright, P. G. (1928). *Tariff on Animal and Vegetable Oils*. New York: Macmillan.

Xie, X. & Geng, Z. (2008). A recursive method for structural learning of directed acyclic graphs. *J. Mach. Learn. Res.* **9**, 459–483.

Xie, X., Geng, Z. & Zhao, Q. (2006). Decomposition of structural learning about directed acyclic graphs. *Artif. Intell.* **170**, 422–439.

References

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika* **2**, 121–134.

Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.