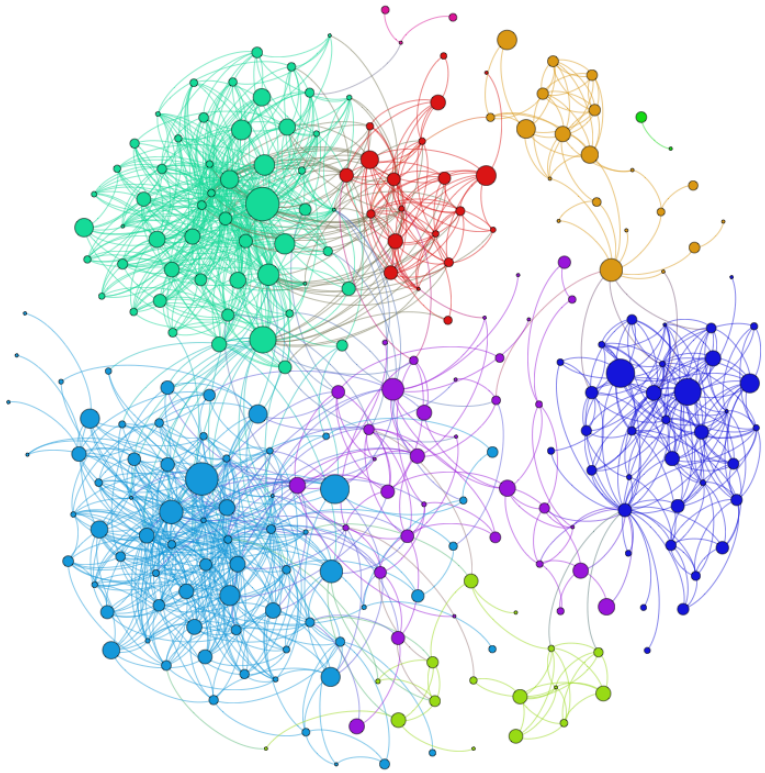
Foundations of Causal Inference

Jiaming Mao

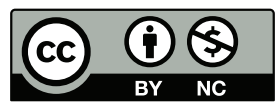
Xiamen University



This version: Spring 2019

Contact: jmao@xmu.edu.cn

Course homepage: jiamingmao.github.io/data-analysis



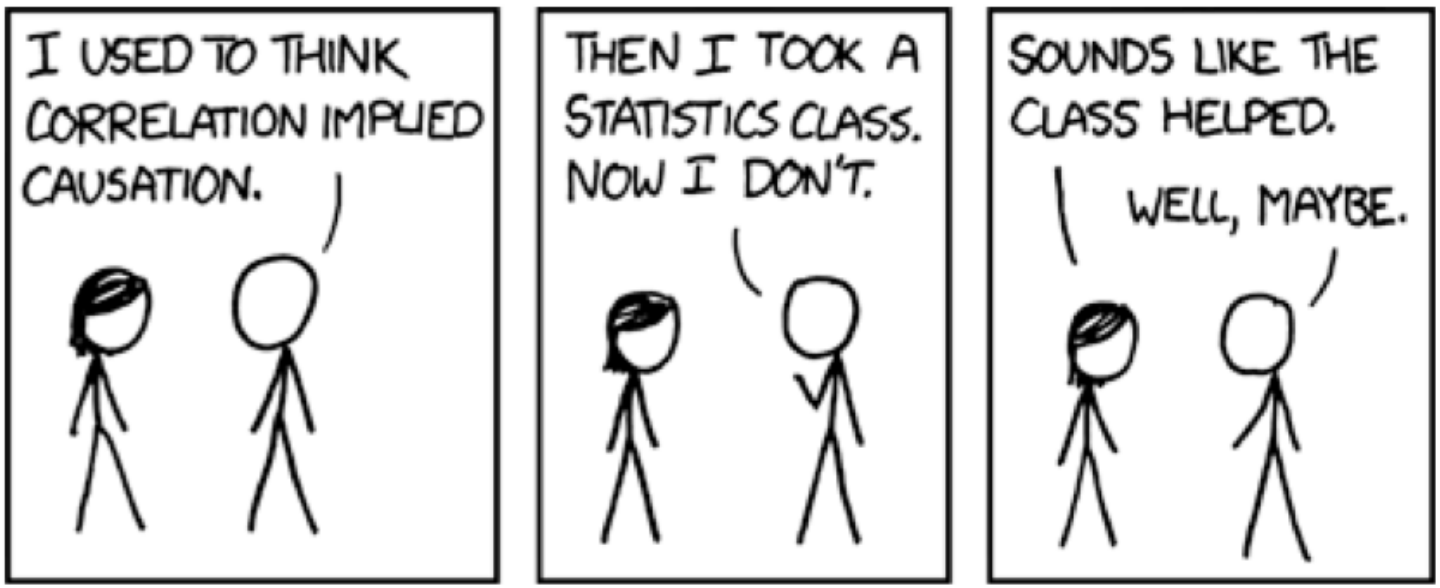
名人名言

"Causa latet: vis est notissima (The cause is hidden, but the result is known)" – Ovid, Metamorphoses, IV. 287.

"Felix qui potuit rerum cognoscere causas (happy be the man who has been able to learn the causes of things)" – Virgil, Georgics, II, 490.

"Shallow men believe in luck or in circumstance. Strong men believe in cause and effect." — Ralph Waldo Emerson

Causal Inference



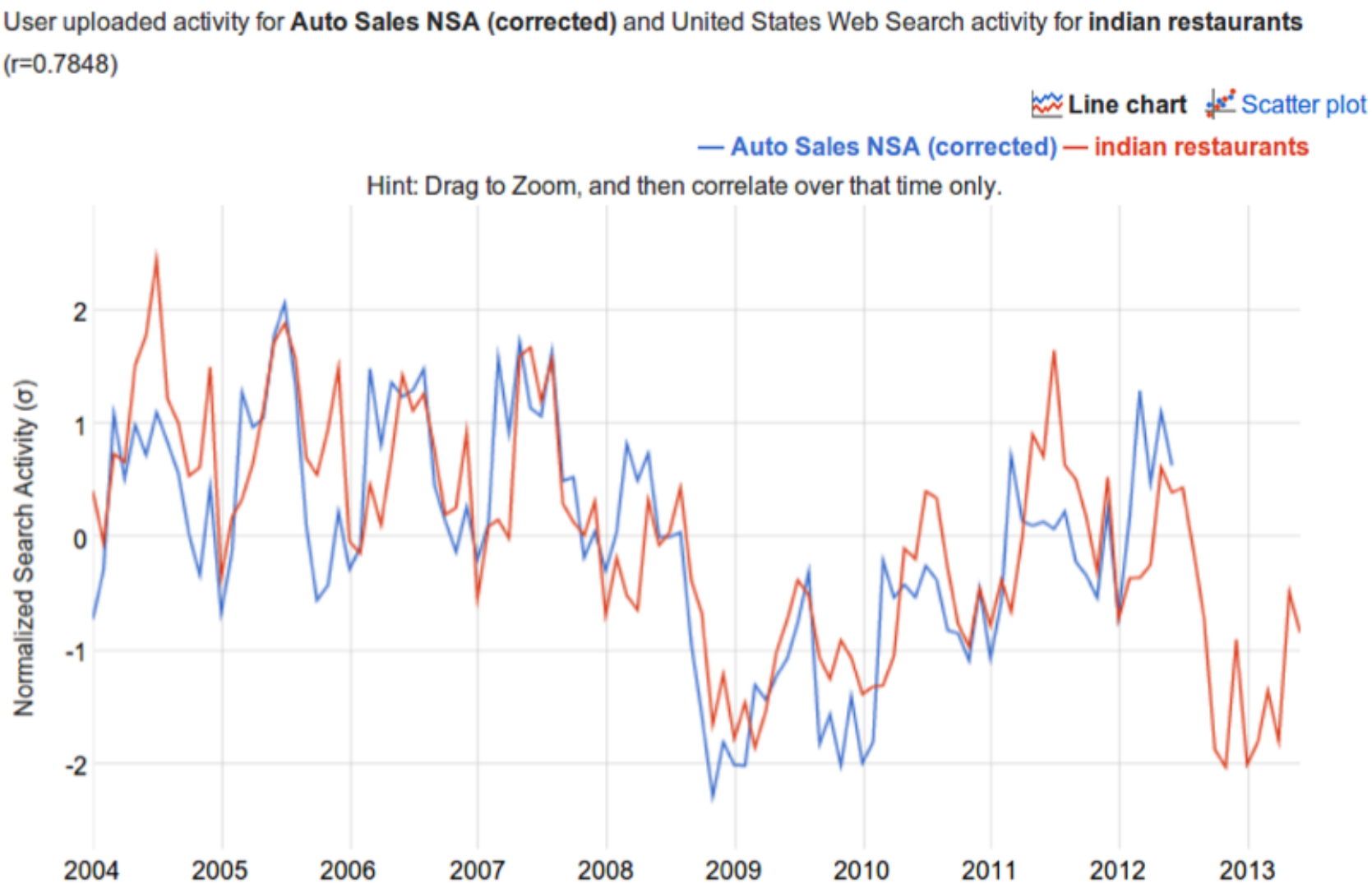
Causal Inference

- Learning $p(x, y)$ or $p(y|x)$ tells us nothing about whether there exists a causal relationship between x and y.

- **Causal inference** is concerned with the following questions:

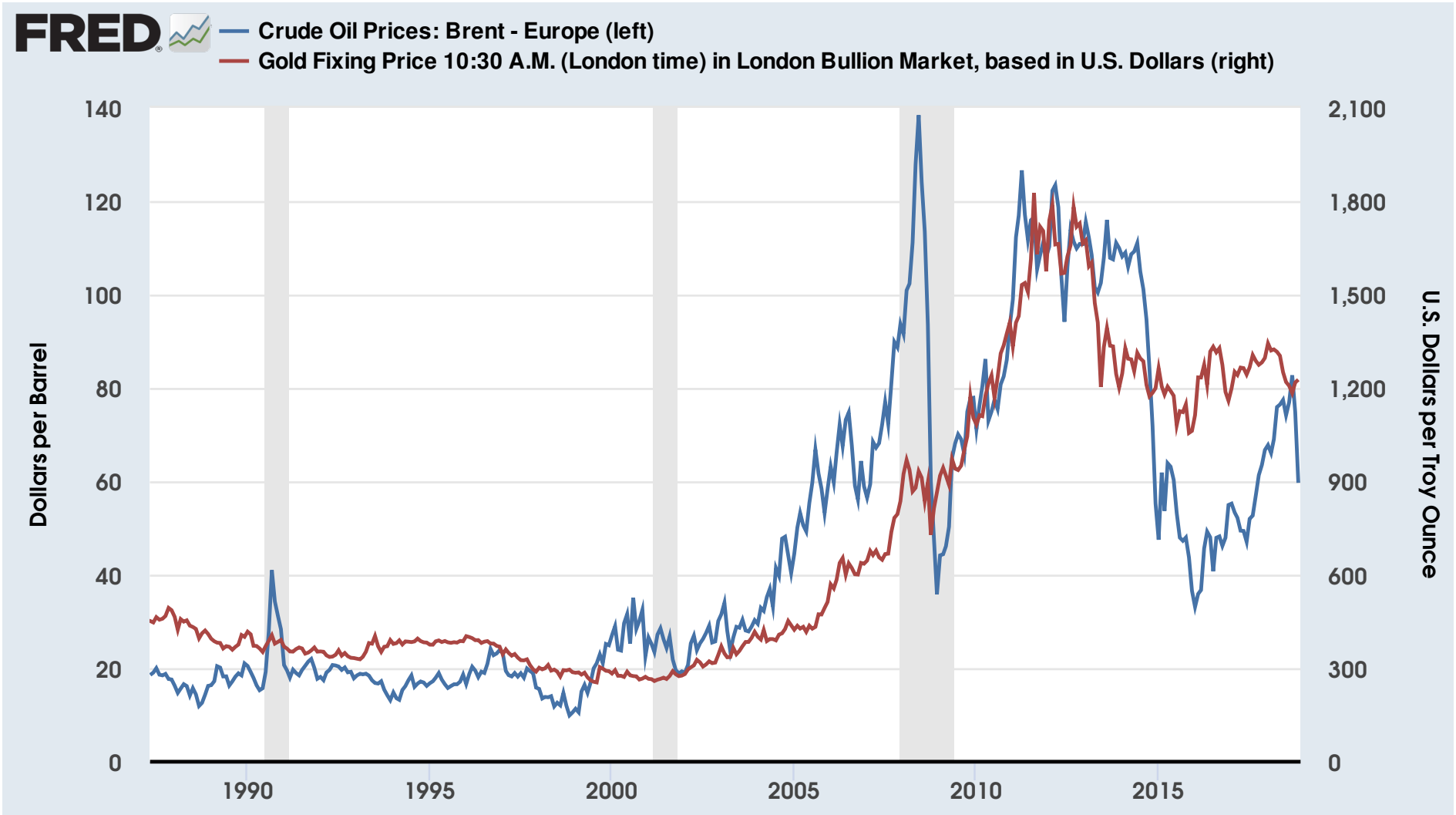
  1. Does x have a causal effect on y? If so, how large is the effect? (**causal effect learning**)
  2. If a causal effect exists, what is the mechanism by which it occurs? (**causal mechanism learning**)

Correlation does not imply Causation



Auto Sales and Search for Indian Restaurants. Source: Google Correlate

Correlation does not imply Causation



Why Causal Inference

- The urge and the capacity to find causal explanations for observed phenomena has been an essential characteristic of human beings since the very beginning of human development and is the very goal of modern science and social science.
- Why do we want to know how things work? – an obvious answer is that it makes a big difference in how we act. If the rooster's crow causes the sun to rise, we could make the night shorter by waking up our rooster earlier.
- Ultimately, every question related to the effect of actions must be decided by causal considerations. Statistical information alone is insufficient.
- True understanding enables predictions under a wide range of circumstances, including new hypothetical situations.

Russell's Chicken

Bertrand Russell[a] told the following cautionary tale of the perils of not understanding causal mechanisms:

> *A chicken infers, on repeated evidence, that when the farmer comes in the morning, he feeds her. The inference serves her well until Christmas morning, when he wrings her neck and serves her for dinner.*

[a] Russell (1912), via Deaton and Cartwright (2018).

Simpson's Paradox

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Should a doctor prescribe this drug?

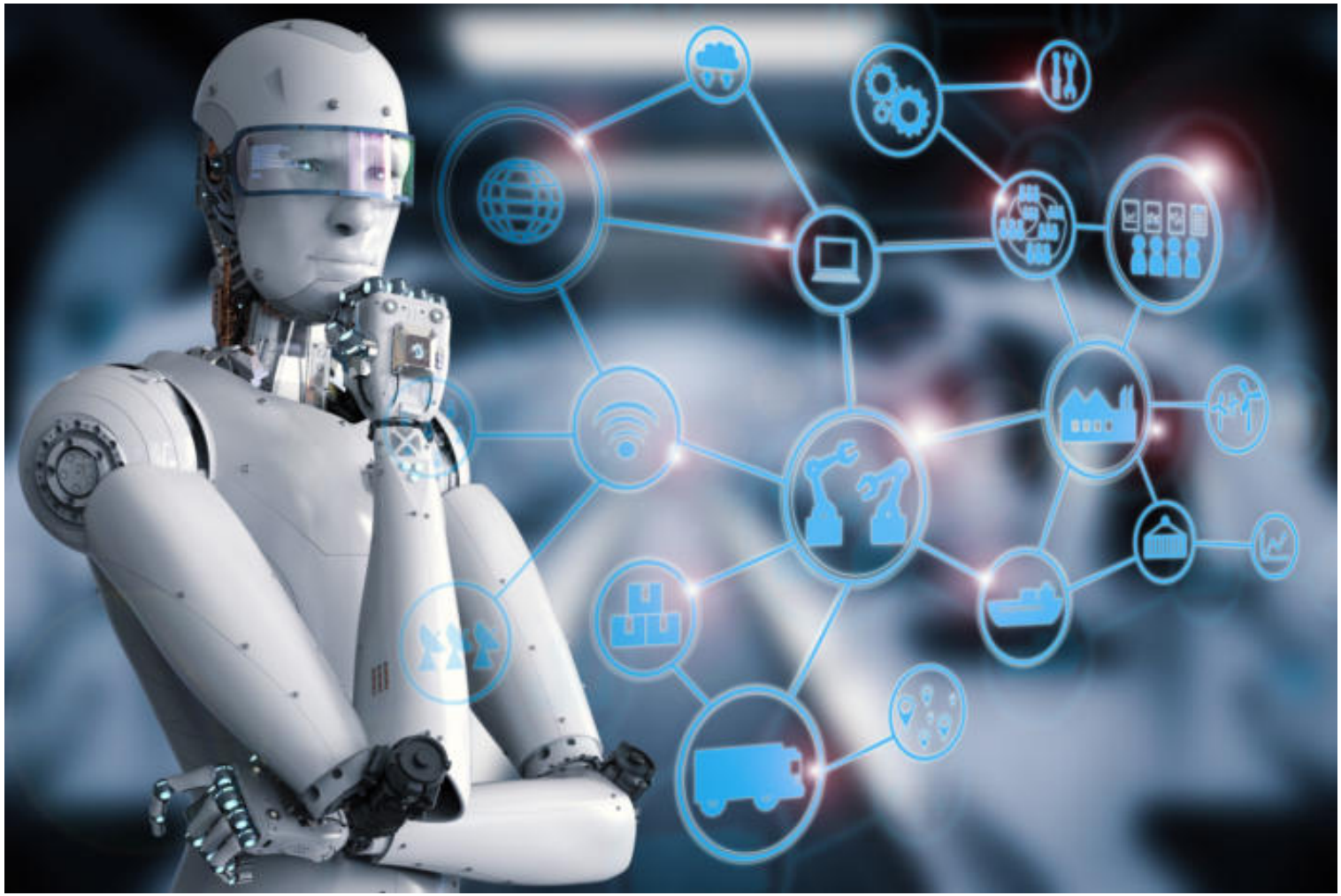
- Any statistical relationship between two variables may be reversed by including additional factors in the analysis.
- Causal relationships are more stable than statistical relationships.

Why Causal Inference

Research on causal inference methodologies has taken on new importance with the development of artificial intelligence (AI).

- How should a robot acquire causal information through interaction with its environment?
- How should a robot receive causal information from humans?

Artificial Intelligence



Seeing vs. Doing

The do operator:

$$\text{do}(x = a) : \text{set } x = a$$

- Barometer readings are useful for predicting rain:

$$\Pr(\text{rain} \mid \text{barometer} = \text{low}) > \Pr(\text{rain} \mid \text{barometer} = \text{high})$$

- But hacking a barometer won't change the probability of raining:

$$\Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{low})) = \Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{high}))$$

Seeing vs. Doing

- Doing: if x has a causal effect on y, then we can change x and expect it to cause a change in y.
- Seeing: If x is correlated[1] with y but does not have a causal effect on y, then we can only observe the correlation without the ability to change y by manipulating x.
- Holland (1986): "*No causation without manipulation*."[2]

---

[1] In this lecture, we use the term "correlation" in its broad sense to mean statistical dependence (association).

[2] i.e., the ability for a cause to be manipulated, at least in principle, is essential to the concept of causality. A thought experiment that is often used to determine whether a variable x is manipulable *in principle* is to imagine a hypothetical experiment that assigns different values to x. Angrist and Pischke (2009): "research questions that cannot be answered by *any* experiment are **FUQ'd**: Fundamentally Unidentified Questions."

Causal vs. Statistical Predictions

- **Causal prediction**: What will y be if I set $x = a$?
  - $E[y|\text{do}(x = a)]$[3]
- **Statistical prediction**: What will y be if I observe $x = a$?
  - $E[y|x = a]$

[3] Assuming we minimize the expected L2 loss in prediction.

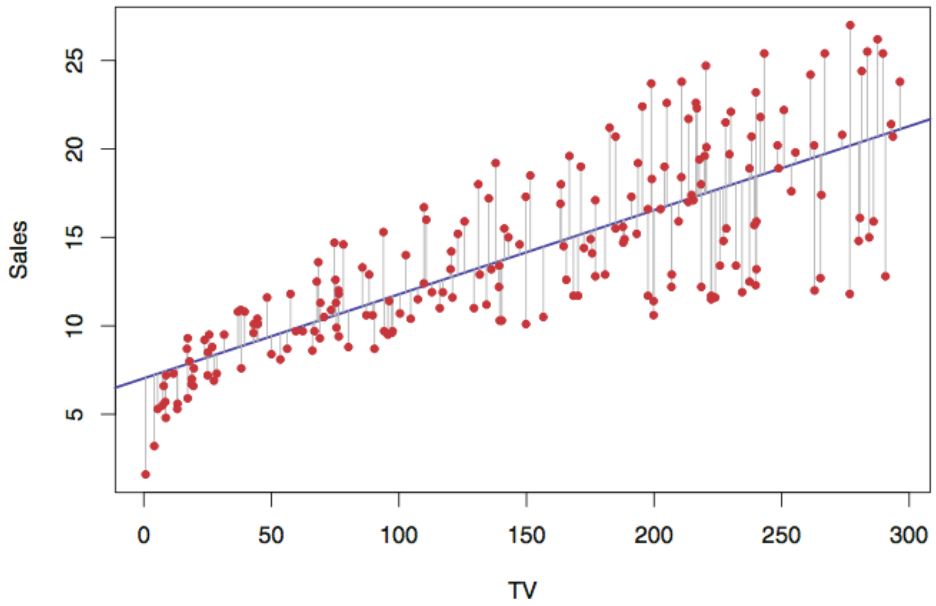
Hospitalization and Health

Average health (assigning a 1 to poor health and a 5 to excellent health) contrasting those who have been an inpatient in the past 12 months and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	3.21	0.014
No Hospital	90049	3.93	0.003

- Q1: what is the expected health status of someone who has received hospitalization? (statistical prediction)
- Q2: what will my health status be if I receive hospitalization? (causal prediction)

Advertising and Sales



- Q1: what is the expected sales of a company with a given amount of TV ad spending? (statistical prediction)
- Q2: how much will my sales increase if I increase my TV ad spending by a certain amount? (causal prediction)

The Potential Outcomes Framework

- The **potential outcomes framework**, also called the **Rubin causal model** (**RCM**)[4], is a framework for causal inference that conceptualizes observed data as if they were outcomes of experiments, conducted either by the researcher – as in actual experiments, or by the subjects of the research themselves – as in observational studies.

- Using the analogy of an experiment, when investigating the causal effect of x on y, x is referred to as **treatment** or **intervention** and y as **outcome**.

[4] due to Neyman (1923) and Rubin (1974, 1978). The RCM represents an effort to use the language of statistical analysis of experiments to model causality.

The Potential Outcomes Framework

- Suppose x takes on a set of discrete values $\{1, \ldots, A\}$. The RCM posits that $y \in \left\{\mathcal{Y}^1, \ldots, \mathcal{Y}^A\right\}$, where $\{\mathcal{Y}^a\}_{a=1}^A$ is a set of random variables, with each $\mathcal{Y}^a$ being the **potential outcome** under the treatment $x = a$:

$$\mathcal{Y}^a \equiv y | \text{do}\,(x = a)$$

- Thus under RCM, the relationship between treatment x and outcome y is described by the joint distribution $p\left(x, \mathcal{Y}^1, \ldots, \mathcal{Y}^A\right)$ and $y = \sum_{a=1}^A \mathcal{Y}^a \mathcal{I}\,(x = a)$.

- If the goal is to predict y when we set $x = a$, then $E\,[\mathcal{Y}^a]$ is the best predictor[5].

[5] Assuming we minimize the expected L2 loss in prediction.

The Potential Outcomes Framework

- Now consider a binary treatment $x \in \{0, 1\}$. The potential outcomes are $\{\mathcal{Y}^0, \mathcal{Y}^1\}$ and the outcome y can be written as:

$$y = x\mathcal{Y}^1 + (1 - x)\,\mathcal{Y}^0$$

- x is said to have a **causal effect** on y if $p\left(\mathcal{Y}^0\right) \neq p\left(\mathcal{Y}^1\right)$[6].
  - Causal effects are defined by comparing potential outcomes.
  - How do we measure the *size* of causal effects?

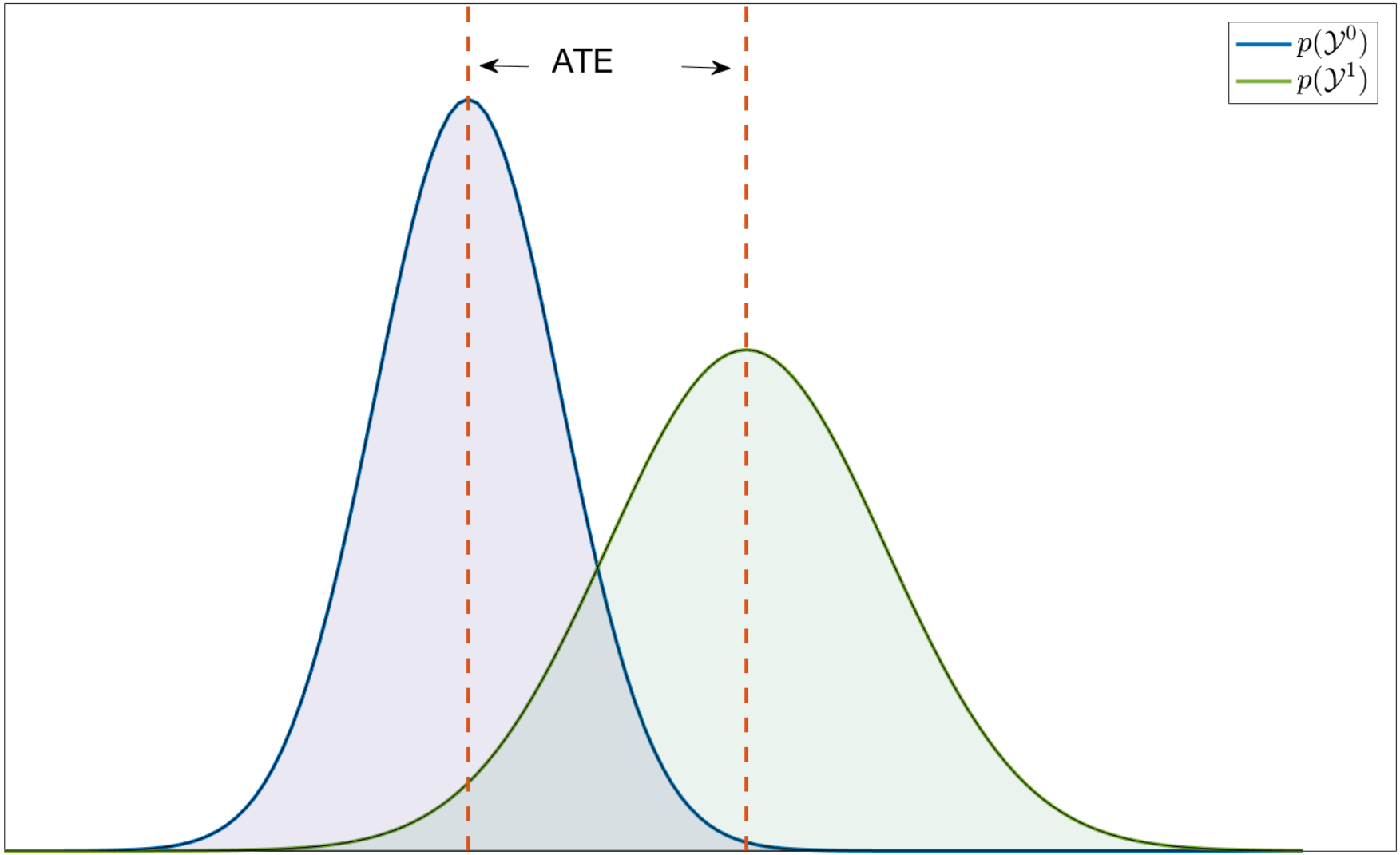
[6] Causal effects are also called **treatment effects** in the RCM. In this lecture, we use the two terms interchangeably.

The Potential Outcomes Framework

Let $\tau \equiv \mathcal{Y}^1 - \mathcal{Y}^0$.

- **Average treatment effect** (**ATE**): $E[\tau]$
- **Average treatment effect on the treated** (**ATT**): $E[\tau | x = 1]$
- **Average treatment effect on the untreated** (**ATU**): $E[\tau | x = 0]$

The Potential Outcomes Framework



The Potential Outcomes Framework

Let the observed data be $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i \in \{0, 1\}$,

$$y_i = x_i \mathcal{Y}_i^1 + (1 - x_i) \mathcal{Y}_i^0$$

, and[7]

$$\left\{\left(x_1, \mathcal{Y}_1^0, \mathcal{Y}_1^1\right), \ldots, \left(x_N, \mathcal{Y}_N^0, \mathcal{Y}_N^1\right)\right\} \overset{i.i.d.}{\sim} p\left(x, \mathcal{Y}^0, \mathcal{Y}^1\right)$$

---

[7] Equivalently, $\{(x_1, y_1), \ldots, (x_N, y_N)\} \overset{i.i.d.}{\sim} p(x, y)$, where

$$p(x, y) = x \int p\left(x, \mathcal{Y}^0, \mathcal{Y}^1\right) d\mathcal{Y}^0 + (1 - x) \int p\left(x, \mathcal{Y}^0, \mathcal{Y}^1\right) d\mathcal{Y}^1$$

The Potential Outcomes Framework

- $\tau_i \equiv \mathcal{Y}_i^1 - \mathcal{Y}_i^0$ is sometimes referred to as the **individual treatment effect**. τ_i is never observed[8]. For each individual, we only observe either $y_i = \mathcal{Y}_i^0$ or $y_i = \mathcal{Y}_i^1$[9].
- The potential outcomes that are not observed are called **counterfactual outcomes**[10].

---

[8] If we consider treatments assigned at different times to the same individual as either different treatments or as assigned to different subjects.

[9] This led Rubin (1974, 1978) to claim that causal inference is fundamentally a *missing data problem*: given any treatment assigned to an individual, the potential outcomes associated with any alternate treatments are missing. Note, however, that this is not a unique problem facing causal inference: any predictive task can be thought of as predicting what y_i will be if x_i takes on some other value.

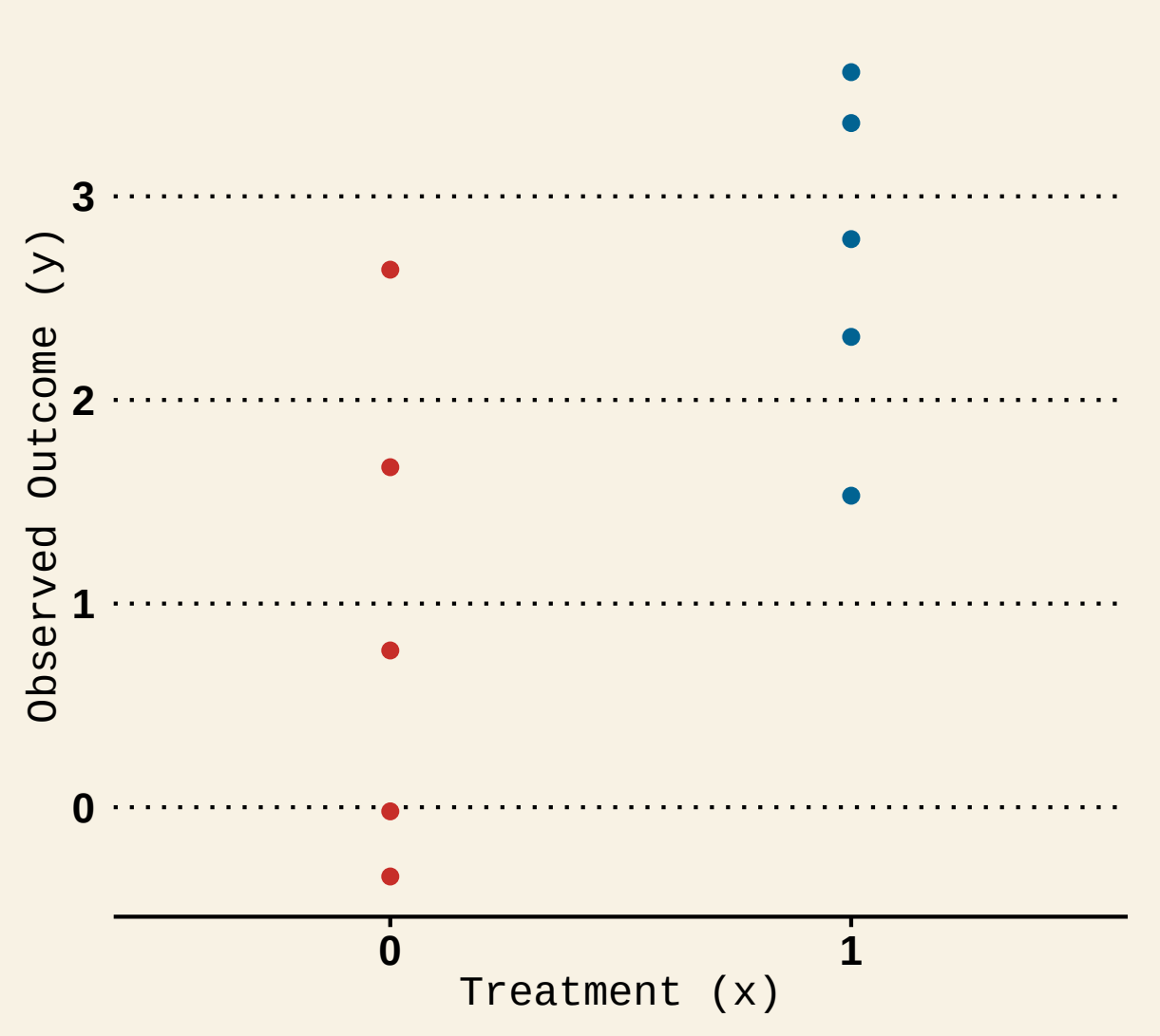
[10] Some authors define counterfactual outcome to be the same thing as potential outcome – observed or unobserved. However, I do not see the utility of having two names for the same concept.

The Potential Outcomes Framework

Treatment x	Observed Outcome y	Potential Outcomes y^0	y^1
0	-.34	-.34	3.46
0	1.67	1.67	4.03
0	-.77	-.77	3.08
0	2.64	2.64	.90
0	-.02	-.02	.96
1	2.31	-1.52	2.31
1	2.79	1.05	2.79
1	1.53	-.13	1.53
1	3.61	-1.41	3.61
1	3.36	.60	3.36

Red: observed potential outcome. Grey: unobserved potential outcome.

The Potential Outcomes Framework



Learning Causal Effects

- Given observed data $\mathcal{D}$, we can learn $p\left(\mathcal{Y}^1 \middle| x=1\right) = p\left(y \middle| x=1\right)$ and $p\left(\mathcal{Y}^0 \middle| x=0\right) = p\left(y \middle| x=0\right)$.
- To compute ATT however, we need information about $p\left(\mathcal{Y}^0 \middle| x=1\right)$:

$$\begin{aligned}\text{ATT} &= E\left[\mathcal{Y}^1 - \mathcal{Y}^0 \middle| x=1\right] \\ &= E\left[\mathcal{Y}^1 \middle| x=1\right] - E\left[\mathcal{Y}^0 \middle| x=1\right]\end{aligned}$$

- Similarly, to compute ATU, we need information about $p\left(\mathcal{Y}^1 \middle| x=0\right)$.

Learning Causal Effects

- To compute ATE, we need information about both $p\left(\mathcal{Y}^0 \middle| x=1\right)$ and $p\left(\mathcal{Y}^1 \middle| x=0\right)$:

$$\begin{aligned}
\text{ATE} &= E\left[\mathcal{Y}^1 - \mathcal{Y}^0\right] \\
&= E\left[\mathcal{Y}^1 - \mathcal{Y}^0 \middle| x=1\right] p\left(x=1\right) \\
&\quad + E\left[\mathcal{Y}^1 - \mathcal{Y}^0 \middle| x=0\right] p\left(x=0\right) \\
&= \text{ATT} \times p\left(x=1\right) + \text{ATU} \times p\left(x=0\right)
\end{aligned}$$

- We can think about causal effect learning as trying to learn these counterfactual outcome probabilities.

Random Treatment Assignment

- In **randomized controlled trials** (**RCT**), the treatment x is assigned randomly to the population.

- For a binary treatment, if x is assigned randomly, then x and $(\mathcal{Y}^0, \mathcal{Y}^1)$ are independent, denoted by $x \amalg (\mathcal{Y}^0, \mathcal{Y}^1)$. Hence

$$p\left(\mathcal{Y}^0\right) = p\left(\mathcal{Y}^0 \middle| x = 1\right) = p\left(\mathcal{Y}^0 \middle| x = 0\right) = p\left(y \middle| x = 0\right)$$
$$p\left(\mathcal{Y}^1\right) = p\left(\mathcal{Y}^1 \middle| x = 0\right) = p\left(\mathcal{Y}^1 \middle| x = 1\right) = p\left(y \middle| x = 1\right)$$

  , i.e. under random assignment, had the group that received treatment b received treatment a, the outcome probabilities would be the same as the group that actually received treatment a.

Random Treatment Assignment

As a result[11],

$$\text{ATE} = \text{ATT} = \text{ATU} = E\left[y|\, x = 1\right] - E\left[y|\, x = 0\right]$$

[11] Note: we are able to calculate

$$\begin{aligned}\text{ATE} = E\left[\mathcal{Y}^1 - \mathcal{Y}^0\right] &\overset{[1]}{=} E\left[\mathcal{Y}^1\right] - E\left[\mathcal{Y}^0\right] \\ &\overset{[2]}{=} E\left[y|\, x = 1\right] - E\left[y|\, x = 0\right]\end{aligned}$$

because of random assignment ([2]) and because the mean is a linear operator ([1]).

In general, however, *without further assumptions*, we will not be able to learn from an RCT other characteristics of the treatment effect distribution, such as its median, variance, and percentiles, as they are not linear operators. E.g.,

$$\text{Median}\left[\mathcal{Y}^1 - \mathcal{Y}^0\right] \neq \text{Median}\left[\mathcal{Y}^1\right] - \text{Median}\left[\mathcal{Y}^0\right]$$

Random Treatment Assignment

Without random assignment of x, $E[y|x=1] - E[y|x=0]$ does not give us the average treatment effect. Consider, for example, the example of hospitalization and health:

$$E[y|x=1] - E[y|x=0] = \underbrace{E[y|x=1] - E\left[y^0 \middle| x=1\right]}_{\text{ATT}} + \underbrace{E\left[y^0 \middle| x=1\right] - E[y|x=0]}_{\text{selection bias}}$$

, where y denotes health outcome and x denotes whether the individual has received hospitalization ($x=1$) or stayed home ($x=0$).

Random Treatment Assignment

- $E[y|x=1] - E[\mathcal{Y}^0|x=1]$ = average health outcome of those who received hospitalization − their average health outcome *if* they had stayed home instead.

- $E[\mathcal{Y}^0|x=1] - E[y|x=0]$ = average health outcome of those who received hospitalization *if* they had stayed home instead − average health outcome of those who did *not* receive hospitalization.

- If individuals chose whether to receive hospitalization themselves, then it is likely that $E[\mathcal{Y}^0|x=1] - E[y|x=0] < 0$. This is called **self-selection effect** or **self-selection bias**.

Random Treatment Assignment

- Self-selection bias is a central concern to causal inference based on observed socio-economic data generated by individual choices.
- When individuals choose their own treatments (*self-selection*)[12], those who choose to receive a treatment can be *systematically* different than those who choose not to, leading to a correlation between treatment and outcome that is not due to direct causation.
- Random assignment of treatment removes such self-selection effect.

[12] Or is it really self-selection? Do we really have free will? Why did the chicken cross the road? – I *choose* not to ponder these questions here ☺

Random Treatment Assignment

More generally, for $x \in \{1, \ldots, A\}$,

- **Exchangeability**: $x \amalg \left(\mathcal{Y}^1, \ldots, \mathcal{Y}^A\right)$
- Random treatment assignment leads to exchangeability.
  - Under random assignment, groups that receive different treatment values are *ex ante* similar, or, **exchangeable**. In this case, we also say that the treatment is **exogenous**.

Random Treatment Assignment

Under random treatment assignment, correlation *implies* causation.

- Association (Correlation): $p\left(y|x=a\right) \neq p\left(y|x=b\right)$
- Causation: $p\left(\mathcal{Y}^{a}\right) \neq p\left(\mathcal{Y}^{b}\right)$
- Random assignment of $x \Rightarrow p\left(\mathcal{Y}^{a}\right) = p\left(y|x=a\right) \; \forall a$

Random Treatment Assignment

- In conditionally randomized experiments, the treatment assignment probabilities depend on the values of some variable(s) s[13].
- The treatment x is randomly assigned within each sub-population with a fixed value of s.
  - e.g., given a binary treatment $x \in \{0, 1\}$ and a population consisting of males ($s = 1$) and females ($s = 2$), a conditionally randomized experiment would assign $x = 1$ to males with probability p_1 and females with probability p_2.
- Conditional random assignment leads to **conditional exchangeability**: $x \amalg \left(\mathcal{Y}^1, \ldots, \mathcal{Y}^A\right) \Big| s$

[13] s can be multi-dimensional: $s = (s_1, \ldots, s_p)$

Random Treatment Assignment

Suppose $s \in \{1, \ldots, S\}$. Under conditional random assignment,

$$E\left[\mathcal{Y}^a\right] = \sum_{j=1}^{S} E\left[\mathcal{Y}^a | \, s = j\right] p\left(s = j\right) \tag{1}$$

$$= \sum_{j=1}^{S} E\left[y | \, x = a, s = j\right] p\left(s = j\right)$$

Random Treatment Assignment

- In the experimental design literature, s is called **nuisance factors** that the experimenter wishes to control when conducting an RCT.
- A nuisance factor is a variable that can affect y either directly or indirectly, but is not of primary interest to the experimenter. A nuisance factor is also called an **effect modifier**: consider a binary treatment x, s is an effect modifier if $p\left(\mathcal{Y}^1 - \mathcal{Y}^0\right) \neq p\left(\mathcal{Y}^1 - \mathcal{Y}^0 \middle| s\right)$.
- Various experimental designs exist to *efficiently* control for s and conduct conditionally randomized experiments.

Random Treatment Assignment

Randomized Complete Block Design (RCBD)

[a]

α_1	α_2	α_3	α_4
b	a	a	c
a	c	b	b
c	b	c	a

$$x \in \{a, b, c\},\ s \in \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$$

[a] The original use of the term **blocking** for removing sources of variation due to nuisance factors comes from agriculture, where a block is typically a set of homogeneous (contiguous) plots of land with similar fertility, moisture, and weather, which are typical nuisance factors in agricultural studies,

Random Treatment Assignment

Latin Square Design (LSD)

a, *b*

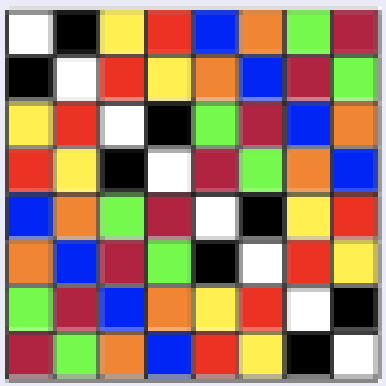
	α_1	α_2	α_3
β_1	*a*	*b*	*c*
β_2	*b*	*c*	*a*
β_3	*c*	*a*	*b*

$$x \in \{a, b, c\},\ s_1 \in \{\alpha_1, \alpha_2, \alpha_3, \},\ s_2 \in \{\beta_1, \beta_2, \beta_3\}$$

---

[a] Not to be confused with lysergic acid diethylamide.

[b] A *Latin square* of order n is an $n \times n$ array of cells in which n symbols are placed, one per cell, in such a way that each symbol occurs once in each row and once in each column.



Using RCTs for Causal Effect Learning

Because random treatment assignment results in exchangeability, RCTs can be used to produce an unbiased estimate of the ATE in the population from which the trial sample is a random sample[14].


The
Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences; and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge.

HAFNER PRESS
A Division of Macmillan Publishing Co., Inc.
New York
COLLIER MACMILLAN PUBLISHERS
London

[14] The population could be the trial sample itself.

Demand Estimation

Goal: want to know consumer demand for a product.

Using the RCM, the problem can be stated as:

- treatment: price (p)
- outcome: purchases (q)

Suppose there are two price levels: $p \in \{L, H\}$.

- Potential outcomes: $\left\{\mathcal{Q}^L, \mathcal{Q}^H\right\}$
- Desired causal effect: ATE $= E\left[\mathcal{Q}^H - \mathcal{Q}^L\right]$
- Data: $\mathcal{D} = \{(p_1, q_1), \ldots, (p_N, q_N)\}$[15]

[15] $\mathcal{D}$ can be generated by either experimental or observational studies, where

$$q_i = \mathcal{Q}_i^L \mathcal{I}(p_i = L) + \mathcal{Q}_i^H \mathcal{I}(p_i = H)$$

, and

$$\left\{\left(p_1, \mathcal{Q}_1^L, \mathcal{Q}_1^H\right), \ldots, \left(p_N, \mathcal{Q}_N^L, \mathcal{Q}_N^H\right)\right\} \overset{i.i.d.}{\sim} p\left(p, \mathcal{Q}^L, \mathcal{Q}^H\right)$$

Demand Estimation

From the data we can learn $p(q|p=a)$, $a \in \{L, H\}$[16].

Problem: without exchangeability, $p(\mathcal{Q}^a) \neq p(\mathcal{Q}^a|p=a) = p(q|p=a)$. The group that "received" the treatment $p = L$ could be systematically different than the group that "received" $p = H$[17].

- People that buy when the price is high can be richer than those who buy when the price is low.
- If we observe the person over time, then her income may be different when the price is low vs. when the price is high.

Solution: randomized experiments.

- e.g., companies could run experiments by randomly assigning prices to customers in different markets and over time.

[16] We abuse notations here by using p to denote both price and probability.

[17] Here we assume there are no other unobserved "treatments" that affect $\{\mathcal{Q}^L, \mathcal{Q}^H\}$, such as the prices of related goods.

Giffen Behavior

- Economic theory has long speculated the existence of *Giffen behavior*: when poor consumers face price increase of inferior goods that are essential to them but have no close substitutes, they may end up buying *more* of these goods, not less.

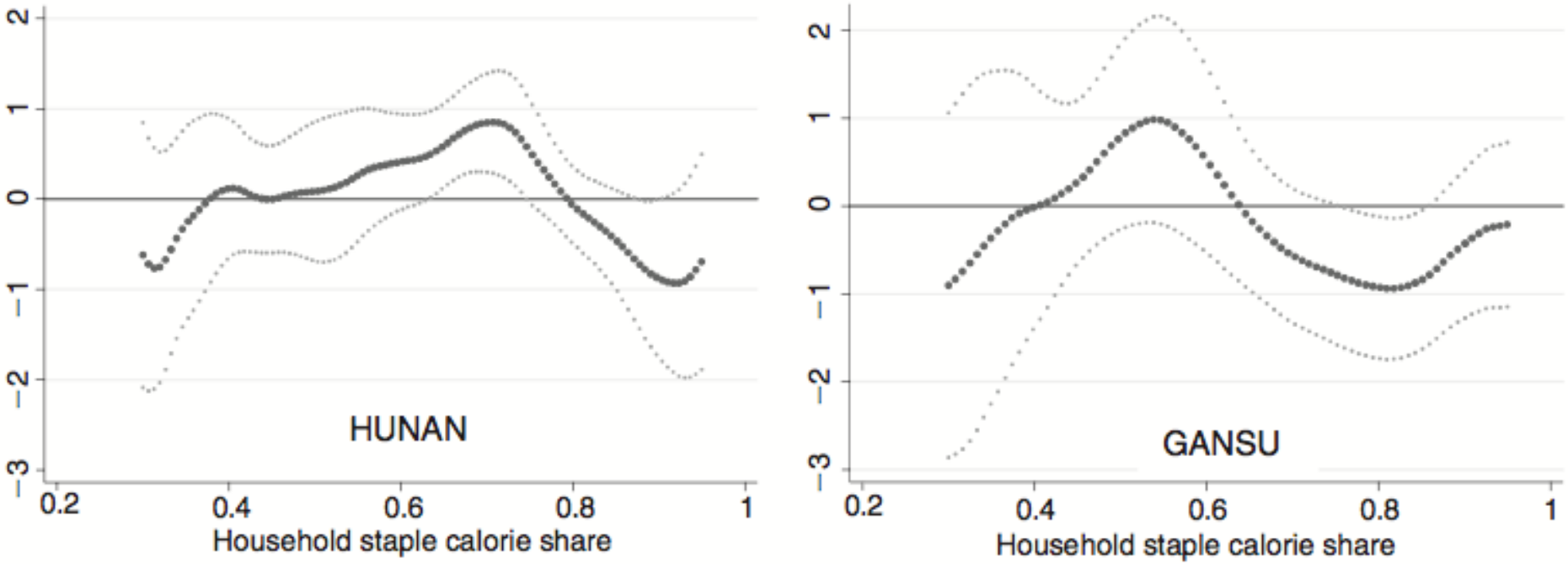
- In reality, we often observe higher prices associated with more purchases, but are they cases of
  - higher demand causing both higher prices and more purchases, or
  - higher prices causing people to buy more (Giffen behavior)?

Giffen Behavior

Jensen and Miller (2008) provides the first empirical evidence of the existence of Giffen behavior using a randomized field experiment.

- randomly selected 1,300 households (3,661 individuals) from two Chinese provinces (Hunan, Gansu) who live under the poverty line.
- Households were randomly given subsidies of .10, .20 or .30 yuan per jin (500g) for staple food consumption (Hunan: rice; Gansu: wheat).
- Theory predicts that households with high but not too high *staple calorie shares* should exhibit Giffen behavior – they would buy *less* not more staple food if the price becomes cheaper.

Giffen Behavior



Percentage decrease in consumption as a result of a one percent subsidy-induced decrease in price, plotted against household staple calorie share. Positive value indicates a decrease in consumption (Jensen and Miller, 2008).

Program Evaluation

Evaluating policy is a central concern to governments. The **program evaluation** literature in applied economic research is concerned with evaluating and predicting the effects of various government programs and economic policies:

- Effect of worker training programs on employment
- Effect of early childhood interventions on adult outcomes
- Effect of negative income taxes on labor supply
- Effect of environmental regulations on pollution emission
- ...

Classroom Size and Student Learning

- Governments operating public schools want to know whether the expense of smaller classes has a payoff in terms of higher student achievement.
- Many studies of education production using non-experimental data suggest there is little or no link between class size and student learning.
- Problem: weaker students often deliberately grouped into smaller classes.

Classroom Size and Student Learning

- The 2002 U.S. Education Sciences Reform Act mandates the use of rigorous experimental or quasi-experimental research designs for all federally-funded education studies.

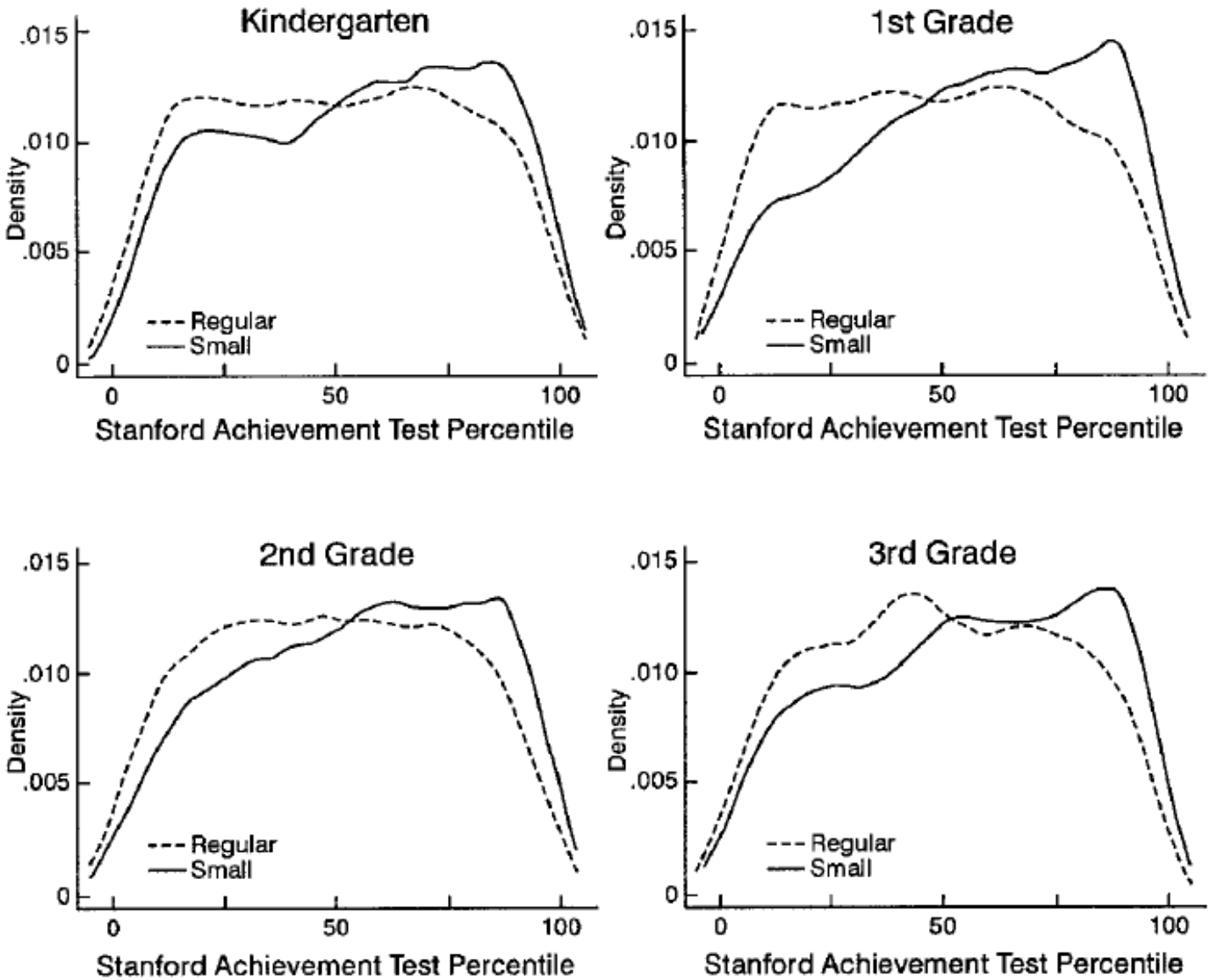
- The Tennessee STAR project: randomly assigned a cohort of kindergarten students to one of three groups: small classes (13–17 students per teacher), regular-size classes (22–25 students), and regular/aide classes (22–25 students) which also included a full-time teacher's aide. The experiment ran for 4 years and a total of 11,600 students from 80 schools were involved (random assignment took place *within* schools).

Classroom Size and Student Learning

		Students who entered STAR in kindergarten			
	Variable	Small	Regular	Regular/Aide	Joint P-value
1.	Free lunch	.47	.48	.50	.09
2.	White/Asian	.68	.67	.66	.26
3.	Age in 1985	5.44	5.43	5.42	.32
4.	Attrition rate	.49	.52	.53	.02
5.	Class size in kindergarten	15.10	22.40	22.80	.00
6.	Percentile score in kindergarten	54.70	48.90	50.00	.00

Krueger (1999)

Classroom Size and Student Learning



Krueger (1999)

But wait! What have we learned?

- Have we learned from the STAR project that small classes are better for student learning everywhere, regardless of culture, student composition, teacher quality, etc.? Probably not. So what is the meaning of this causal effect estimate?
  - In particular, does the effect apply to inner-city students in Chicago or Harry Porter and his friends at Hogwarts? Does it apply if the subject of study is economics, or if the teachers come from China?
  - The answer depends on our understanding of the underlying causal mechanism: if we believe small class affects learning by allowing students to see the blackboard more clearly and hence does not depend on student composition, then the effect probably *should* apply to inner-city students in Chicago. But since professors at Hogwarts do not use a blackboard, the effect probably does not apply there.
    Importantly, if we believe the observed effect was due to the initial discomfort felt by Kindergarten students enrolled in large classes, then the effect may not even apply to the same cohort of Tennessee students if we were to conduct the experiment on them again.

Meaningful Causal Effects

- Causal effects do not exist in a vacuum. In social scientific research, the causal effects of interest are often the results of complex social and economic processes.
- Causal effects are population-specific: when we talk about "the causal effect of x on y", it is always with respect to a specific population with a specific social, cultural, and economic environment.
- The set of populations in which a causal effect applies is called its **scope**. A causal effect is only meaningful if we can define its scope.

Meaningful Causal Effects

Consider a binary treatment x. In a specific population with a specific social, cultural, and economic environment,

$$E\left[\mathcal{Y}^1 - \mathcal{Y}^0\right] = \int E\left[\mathcal{Y}^1 - \mathcal{Y}^0 \middle| s\right] p(s)\, ds \tag{2}$$

, where $s = s(\mathcal{M})$ is the set of effect modifiers according to the causal mechanism $\mathcal{M}$ by which x affects y in that population.

The causal effect of x on y can differ in two populations because:

1. The causal mechanism is different.
   - The effect of decreasing oil supply on gas price will be different in countries where gas prices are determined by market and countries where gas prices are controlled by governments.
2. Or, the mechanism is the same, but $p(s)$ is different.
   - The impact of raising the retirement age on the economy will be different for two countries with different population age structures.

Meaningful Causal Effects

- An experiment or observational study is conducted on a specific sample, drawn likewise from a specific population with a specific social, cultural, and economic environment – call it the **study population**.
- Often this study population is not we are interested in. Instead, we are interested in learning the causal effect in a **target population**.
- The result of an experimental or observational study is said to have **internal validity** if it is valid for the study population. It is said to have **external validity** if it can be generalized or extrapolated to other populations. The ability to be extrapolated from one population to another is also referred to as the **transportability** of a result.
- A causal effect that we have learned is transportable from the study population to the target population if both are within its scope.

Meaningful Causal Effects

- To understand the meaning and scope of a causal effect, we need an understanding of the underlying causal mechanism[18], which should be based on prior information and analyses, i.e., our **prior knowledge**.

- Understanding the underlying causal mechanism is not only necessary for understanding a causal effect – what it means, where it applies, but also for determining whether a causal effect estimated from a study population applies to the target population – whether the same mechanism holds in both populations and whether the relevant effect modifiers have the same distribution[19].

---

[18] Heckman and Vytlacil (2007) thus criticize the RCM: "Rooted in biostatistics, they are motivated by the experiment as an ideal. They do not clearly specify the mechanisms determining how hypothetical counterfactuals are realized or how hypothetical interventions are implemented except to compare "randomized" with "nonrandomized" interventions."

[19] This is in fact true for *all* statistical inferences, causal or non-causal.

Meaningful Causal Effects

- Understanding the causal mechanism also helps us to raise more interesting questions and design studies to learn more useful effects.
  - Have we learned from the STAR project why small classes were better? There are many ways class size could affect student learning. Teachers could employ different teaching methods in small classes. Students may interact more with each other, therefore learning better in small classes. Or it could be simply the fact that sitting closer to the blackboard allows students to see and therefore learn better. The overall effect of small class on learning masks the effect of each of these possible channels.
  - An understanding of the existence of these different channels allows us to design our studies specifically to learn their respective effects, which would be more interesting than the overall effect.

Meaningful Causal Effects

Fumigation and Yield

Fumigation is the use of fumigants to control eelworms which affects crop yield. Suppose we conduct an RCT to study the effect of fumigation on barley yield by randomly selecting N barley fields in Llanfairpwllgwyngyllgogerychwyrndrobwllllantysiliogogogoch (short: Llanfairpwll)[a] and randomly applying fumigation to M of them. The result shows that fumigation increases barley yield by 20%.

What does this result mean? Where does it apply? Is the result valid for barley fields in China? Does it apply to Okay, Oklahoma[b]? Or can we even say that it applies to the same fields in Llanfairpwll, in the sense that we will observe the same effect if we are to conduct the same experiment again next year?

---

[a] A village in Wales. See here.

[b] Another actual town.

Meaningful Causal Effects

Fumigation and Yield (cont.)

The understanding of the result depends on our understanding of the causal mechanism and its implied effect modifiers.

Let $\widetilde{\tau}$ denote the possible effects in the study population and let τ denote the possible effects for any barley field in the world. Suppose we believe the effect of fumigation depends on the season in which a field is fumigated, and suppose the experiment is conducted in the summer, then the result we get from the experiment is really $E[\widetilde{\tau}] = E[\tau|\text{summer}]$.

If, in addition, we believe the effect also depends on what crops were grown last year and in Llanfairpwll, 50% of the barley fields grew barley last year and the other 50% grew wheat, then the result we get is $E[\widetilde{\tau}] = 0.5 \times E[\tau|\text{summer}, \text{barley}] + 0.5 \times E[\tau|\text{summer}, \text{wheat}]$, where we condition on season and last year's crop.

External Validity

"Psychology is the study of psychology students." – Anonymous

Vol 466|1 July 2010 nature

OPINION

Most people are not WEIRD

To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners, argue **Joseph Henrich**, **Steven J. Heine** and **Ara Norenzayan**.

A 2008 survey of the top psychology journals found that 96% of subjects were from Western, educated, industrialized, rich and democratic (WEIRD) societies – particularly American undergraduates.

The Experimental Ideal and Its Limitations

- RCTs are often regarded as the gold standard for causal inference[20]. However, there is a popular but mistaken belief that they require no assumptions on causal mechanisms – as we have discussed, without an understanding of – or making assumptions on – the underlying causal mechanism, any causal effect estimate is meaningless, whether produced by RCTs or observational studies[21,22].

- In addition, there are limits to the practicality and usefulness of RCTs.

---

[20] See page 116 for a discussion of their exact virtues.

[21] Heckman and Vytlacil (2007) on the consequence of conducting RCTs without an understanding of the underlying mechanism: "Simplicity in estimation is often accompanied by obscurity in interpretation ... Blind empiricism leads nowhere."

[22] Deaton and Cartwright (2018): "There is no escape from thinking about the way things work; the why as well as the what."

The Experimental Ideal and Its Limitations

- For many causal inference problems, RCTs are impossible or impractical to run.
  - infeasibility (e.g., monetary policy)
  - ethical reasons (e.g., smoking and lung cancer)
  - cost and duration (e.g., childhood intervention and adult outcomes)
    - ⋆ RCTs require special conditions if they are to be conducted successfully – local agreements, compliant subjects, affordable administrators, multiple blinding, people competent to measure and record outcomes reliably, etc.,
    - ⋆ Long duration studies often suffer from significant (non-random) attrition.
  - high-dimensional treatment or nuisance factors

The Experimental Ideal and Its Limitations

- Many causal inference problems involve a large number of effect modifiers s, making it infeasible to conduct RCTs that control for all dimensions of s or have enough randomization such that enough values of s in each dimension are observed.

- If we do not specifically control for or randomize over s, however, then the causal effect estimate we get may be very local (conditional on fixed values of s in many dimensions). This limits the usefulness of RCTs and often means that we have to rely on observational studies, which are less subject to the constraints listed on the preceding page.

Meaningful Causal Effects

Fumigation and Yield (cont.)

- If we believe the effect of fumigation on barley yield varies by seasons, then we need to control for or randomize over seasons in order to obtain causal effect estimates that are not conditional on a particular season. This, however, increases the duration of the experiment and would significantly increase its cost.
- When many other factors also determine the effect of fumigation and we do not specifically control or randomize over them, then we may end up getting a very local effect such as
  $E\left[\tau|\text{summer}, \text{lcrop=barley}, \text{rainfall=high}, \text{altitude=low}, ...\right]$

Scaling Up

- Often, RCTs are conducted for program evaluation purposes. Because of cost and duration constraints, they are typically small-scale, but if deemed successful, the program is then a candidate for scaling-up – applying the same intervention to a much larger area and population.
- Predicting the same results at scale as in the trial can be problematic, however, as the larger target population can be very different from the study population so that causal effects are not transportable.
- But even if the trial sample is a random sample of the target population, so that the target population = the study population[23], applying the same intervention to everyone in the population could generate very different effects than in the trial due to general equilibrium effects – a particular problem that often limits the usefulness of RCTs for program evaluation.

[23] e.g., the target population being the national population and the trial sample being a nationally representative sample.

Scaling Up

Fumigation and Yield (cont.)

- Suppose a government is interested in finding ways to help farmers increase their income. Since in the fumigation study, farmers whose fields have been fumigated see significant increase in their crop production and hence income, the government believes that it is a good idea to subsidize fumigation.
- However, if the use of fumigants on barley fields is scaled up to the whole country, then the price will drop – assuming a closed domestic barley market – and if the demand for barley is price inelastic, then farmers' incomes will fall.
- In this case, the scaled-up effect is *opposite in sign* to the trial effect.

SUTVA

- The existence of general equilibrium effects can be thought of as a violation of the **stable unit treatment value assumption** (**SUTVA**), which is an assumption that an individual unit's potential outcome under a treatment does *not* depend on the treatments received by other individual units[24].
- SUTVA is implicitly assumed in the RCM: $y_i = \mathcal{Y}_i^{(x_1,\ldots,x_N)} = \mathcal{Y}_i^{x_i}$
- SUTVA can be violated if there exists interaction among individual units[25].

[24] An individual unit could be a person, a firm, a country, etc. at a given point in time.
[25] Of which general equilibrium effect is a manifestation.

SUTVA

Treatment assignment patterns			Potential outcomes	
$\begin{bmatrix} d_1=1 \\ d_2=0 \\ d_3=0 \end{bmatrix}$	or $\begin{bmatrix} d_1=0 \\ d_2=1 \\ d_3=0 \end{bmatrix}$	or $\begin{bmatrix} d_1=0 \\ d_2=0 \\ d_3=1 \end{bmatrix}$	$\begin{matrix} y_1^1=3 \\ y_2^1=3 \\ y_3^1=3 \end{matrix}$	$\begin{matrix} y_1^0=1 \\ y_2^0=1 \\ y_3^0=1 \end{matrix}$
$\begin{bmatrix} d_1=1 \\ d_2=1 \\ d_3=0 \end{bmatrix}$	or $\begin{bmatrix} d_1=0 \\ d_2=1 \\ d_3=1 \end{bmatrix}$	or $\begin{bmatrix} d_1=1 \\ d_2=0 \\ d_3=1 \end{bmatrix}$	$\begin{matrix} y_1^1=2 \\ y_2^1=2 \\ y_3^1=2 \end{matrix}$	$\begin{matrix} y_1^0=1 \\ y_2^0=1 \\ y_3^0=1 \end{matrix}$

The causal effect of d on y depends on how many individuals receive $d = 1$. SUTVA is violated if d is treatment and y is outcome, in which case there is **treatment dilution**: the more treated, the less effective the treatment.

SUTVA

When SUTVA is violated,

$$p\left(y_i|\operatorname{do}\left(x_i=a,x_j=b\right)\right)\neq p\left(y_i|\operatorname{do}\left(x_i=a\right),x_j=b\right)$$
$$\neq p\left(y_i|\operatorname{do}\left(x_i=a\right)\right)=\int p\left(y_i|\operatorname{do}\left(x_i=a\right),x_j\right)p\left(x_j\right)dx_j$$

In this case, $\operatorname{do}\left(x_i=a,x_j=b\right)$, $\operatorname{do}\left(x_i=a,x_j=c\right)$, $\operatorname{do}\left(x_i=a\right)|\,x_j=b$, and $\operatorname{do}\left(x_i=a\right)|\,x_j=c$ are effectively different interventions[26].

[26] Thus, the violation of SUTVA can also be viewed as a problem of ill-defined interventions.

SUTVA

SUTVA can be thought of as an i.i.d. assumption on causal effects. If violated, then we need to take the interaction into account. Assuming a population of $\{i, j\}$, there are two approaches to do so:

1. Learn $p\left(y_i \middle| \operatorname{do}\left(x_i\right)\right)$ or $p\left(y_i \middle| \operatorname{do}\left(x_i\right), x_j\right)$, treating x_j as an effect modifier.
    - When estimating the treatment effect on an individual unit, if SUTVA is violated, then we need to consider the treatments received by other individual units as *effect modifiers*.

SUTVA

2. Learn $p\left(y_i \middle| \operatorname{do}\left(x_i, x_j\right)\right)$.
   - This requires changing the unit of analysis from the original individual unit to a population of those units where interaction occurs and is confined in[27].

[27] For example, for the problem on page 68, we can define each individual unit i to be a "local population" where such interference occurs and is confined in, and let the underlying population be a population of such local populations. Define the treatment variable x to be

$$x = \begin{cases} 1 & \text{if 1 individual in the local population receives } d = 1 \\ 2 & \text{if 2 individuals in the local population receive } d = 1 \\ \vdots & \vdots \end{cases}$$

SUTVA would be satisfied if we look at the causal effect of x on y.

SUTVA

- Socio-economic outcomes are often the results of individual interaction[28]. Individual choices are rarely independent and each person's choice can affect other people[29].
  - micro-scale: social and strategic interaction (e.g., firm competition in oligopolistic markets)
  - macro-scale: general equilibrium effects
- Thus, SUTVA would often be violated if we do not properly take these interaction effects into account.

[28] Unlike, say, in the medical sciences.

[29] Such interaction effects are sometimes negligible, as in the case of buyers and sellers in competitive markets. Here we emphasize that they are often not.

SUTVA

College Education and Wage

- Labor economists are perpetually interested in the effect of education on labor market returns. But questions such as "what is the effect of college education on wage?" needs some clarity.
- Wage is an equilibrium outcome. How much wage a person would earn if she receives college education depends on labor demand and labor supply. Labor supply, in turn, depends on how many other people have received college education[a].
- The effect of college education on wage is clearly different if only one person receives college education and if all individuals do.
- When people ask this question, the causal effect they most likely *really* have in mind is the effect of an individual receiving college education on her wage *conditional* on current labor demand and labor supply.

[a] Disregarding the heterogeneity in college education (good college, bad college, history major, economics major) and treating it as homogeneous here.

Prior Knowledge

- Causal inference generally requires prior knowledge regarding the underlying causal mechanisms. Such knowledge can only exist as a result of previously observed information and conducted studies. Causal inference therefore builds on causal inference[30].

[30] See page 133 for a discussion on causal mechanism learning and the process of scientific progress.

Causal Model

How to represent our knowledge of a causal mechanism? Given variables x and y, we can say

$$\begin{aligned} x &\sim U(0,1) \\ y &= 2x \end{aligned} \tag{3}$$

But without giving "=" a causal reading, this is just a statistical model that gives us the joint distribution $p(x,y)$.

Causal Model

(3) becomes a causal model if we imbue "$=$" with the causal meaning that the variable on the left is determined by the variables on the right.

- Here: x causes y, so if we set $x = 1$, y will be 2, but setting the value of y will not affect x.
- Once given a causal meaning, "$y = 2x$" becomes a **structural equation** and is sometimes written as "$y \leftarrow 2x$".

Causal Model

- A **causal model**[31] for a set of random variables $\{x_1, \ldots, x_n\}$ is a model that specifies the joint distribution $p(x_1, \ldots, x_n)$, as well as the **causal structure** governing $\{x_1, \ldots, x_n\}$, which describes the causal relationships among the variables[32].

[31] Also called **scientific model**.

[32] In other words, a causal model $\mathcal{M} = (\mathcal{H}, \mathcal{G})$, where $\mathcal{H}$ is a generative statistical model, while $\mathcal{G}$ specifies the causal structure.

Causal Diagrams

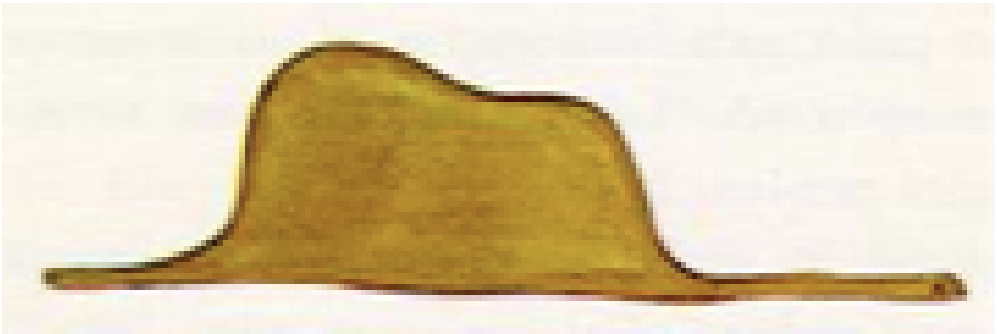
- **Causal diagrams** are graphs that can be used to represent causal structures and therefore describe our qualitative knowledge about a causal mechanism[33,34].

[33] A causal model whose causal structure is represented by a causal diagram is called a **causal graphical model** or **causal Bayesian network**. See Appendix I for an introduction to graphical models and Bayesian networks.

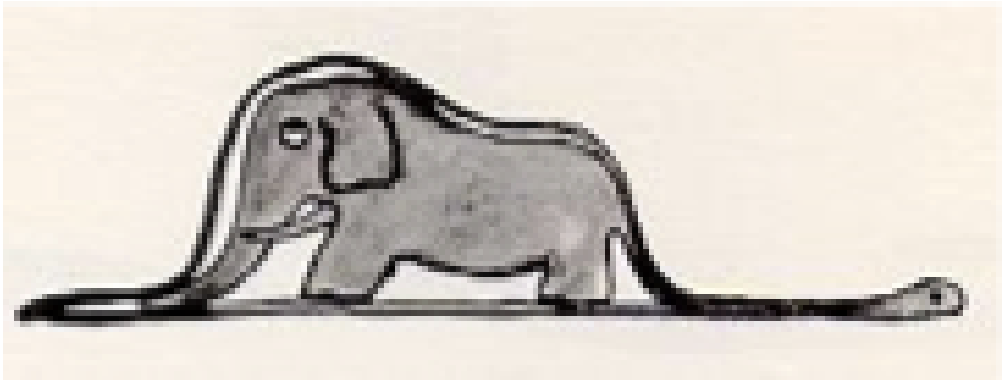
[34] While the RCM and the potential outcomes framework were developed in statistics, the modern theory of causal graphical models arose within the disciplines of computer science and artificial intelligence. See Pearl (2009).

Causal Diagrams

For example, suppose we see a boa constrictor that looks like this:



Our theory of the causal mechanism that leads to the boa constrictor looking like this is that it has just swallowed a baby elephant:



Causal Diagrams

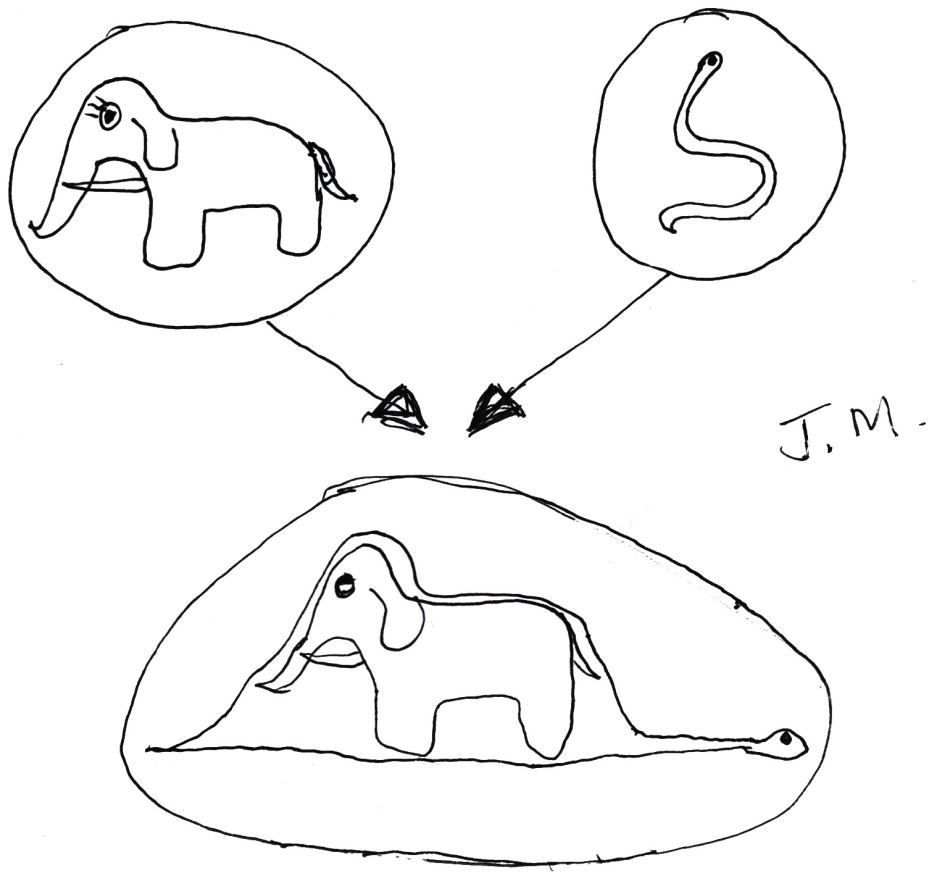
How do we represent our theory? We can write out a full causal model:

$$\begin{aligned}
\log h^e &\sim \mathcal{N}(0, 0.1) \\
\log \ell^e &\sim \mathcal{N}(0.5, 0.1) \\
\log \ell^b &\sim \mathcal{N}(1.5, 0.2) \\
a \left| \ell^e, \ell^b, \ell^b > \ell^e \right. &\sim U\left(0, \ell^b - \ell^e\right) \\
y &\leftarrow \begin{cases} h^e \mathcal{I}(a \leq x \leq a + \ell^e)\, \mathcal{I}(E = 1) & \ell^e < \ell^b \\ 0 & \ell^e \geq \ell^b \end{cases}
\end{aligned}$$

, where y is the height of the boa constrictor, x is the distance along the body of the boa constrictor from its head, (h^e, ℓ^e) are respectively the height and length of the baby elephant, ℓ^b is the length of the boa constrictor, and $E \in \{0, 1\}$ is the event that the boa constrictor has swallowed the elephant.

Causal Diagrams

Or we can draw a causal diagram:



Causal Diagrams

- In a causal diagram, the nodes (vertices) are $\{x_1, \ldots, x_n\}$, with directed edges (arrows) representing direct causation.

- The presence of an arrow that points from x_i to x_j indicates either that x_i has a direct causal effect on x_j – an effect not mediated through any other variables on the graph, or that we are unwilling to assume such a effect does not exist. The lack of an arrow from x_i to x_j then indicates the absence of a direct effect[35].

[35] The absence of an arrow therefore represents a more substantive assumption.

Causal Diagrams

- If an arrow points from x_i to x_j, then x_i is a **parent** of x_j and x_j a **child** of x_i[36].
- A **path** is a sequence of connected nodes. The path is **causal** if all its arrows point in the same direction. Otherwise it is **noncausal**.
- All nodes on a causal path that begins with x_i are **descendants** of x_i. Those on a causal path that leads to x_j are **ancestors** of x_j. A variable is a cause of all its descendants.
- Variables with no parents are said to be **exogenous** to the causal model represented by the causal diagram. Others are **endogenous**.

[36] For detailed definitions on the concepts introduced here, see Appendix I

Causal Diagrams

A causal diagram must satisfy the following properties:

- **Causal Markov Condition**: $x_i \amalg \mathrm{nd}(x_i) | \mathrm{pa}(x_i)$, where $\mathrm{pa}(x_i)$ and $\mathrm{nd}(x_i)$ denote respectively the parents and non-descendants of x_i[37].
- **Completeness**: All common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph[38].
- **Faithfulness**: The joint distribution $p(x_1, \ldots, x_N)$ has all of the conditional independence relations implied by the causal diagram, and *only* those conditional independence relations.

---

[37] i.e., a variable x_i is independent of any other variables (except its own effects) conditional on its direct causes.

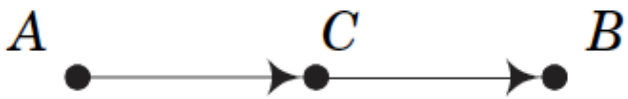
[38] As it turns out, this property is tantamount to the requirement that all relevant factors are accounted for. Our ability to extract causal information from data is predicated on this untestable assumption.

Independence in Causal Diagrams

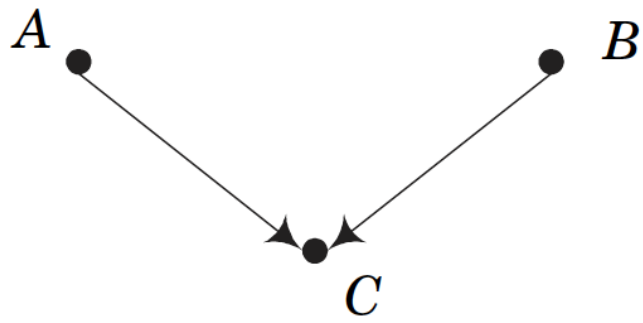
- Given a path $\mathcal{P}$, a **collider** is a node c on $\mathcal{P}$ with neighbors a and b on $\mathcal{P}$ such that $a \to c \leftarrow b$.
- A path is said to be **blocked** if it contains a noncollider that has been conditioned on, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on.
- Two variables are said to be **d-separated** if all paths between them are blocked. Otherwise they are **d-connected**.
- If two variables are d-separated (after conditioning on a set of variables), then they are (conditionally) independent. If two variables are d-connected (after conditioning on a set of variables), then they are (conditionally) dependent (associated)[39].

[39] Without the faithfulness condition, d-connection does not necessarily imply conditional dependence.

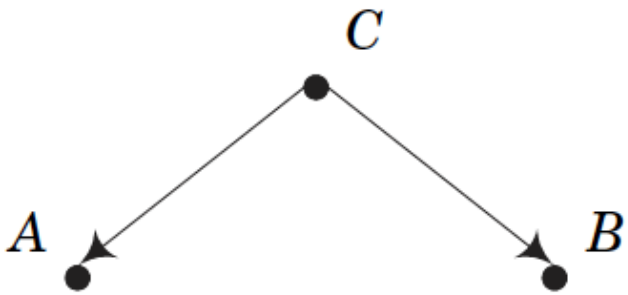
Basic Patterns of Causal Relations



(a) Mediation



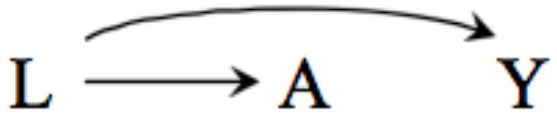
(c) Mutual causation



(b) Mutual dependence

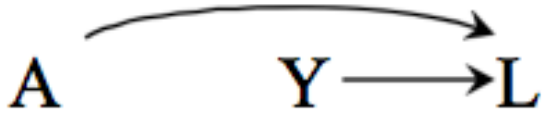
Basic patterns of causal relationships among three variables

Association and Causation



- L has a causal effect on both A and Y. A does not have a causal effect on Y. A depends on L and on *no other causes* of Y.
- L is called a **common cause** to A and Y.
- A and Y are d-connected and hence associated, because there exists an *open* path, $A \leftarrow L \rightarrow Y$, between them.
- Having information about A improves our ability to predict Y, even though A does not have a causal effect on Y.
- E.g., A : carrying a lighter; Y : lung cancer; L : smoking

Association and Causation



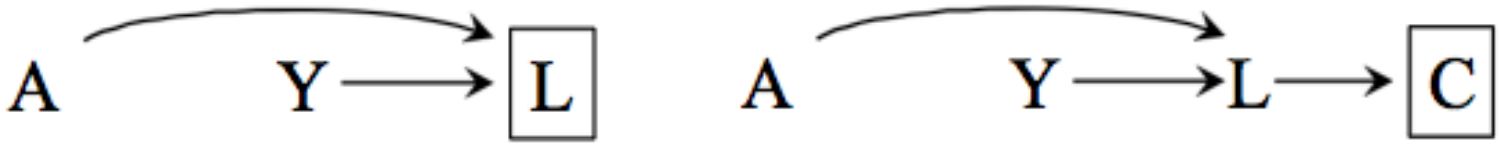
- Both A and Y have a causal effect on L. A does not have a causal effect on Y.
- L is called a **common effect** of A and Y.
- L is a collider on the path $A \rightarrow L \leftarrow Y$ that has not been conditioned on. Hence L blocks the path.
- A and Y are d-separated and hence independent, because the only path between them, $A \rightarrow L \leftarrow Y$, is blocked.
- E.g., A : family heart disease history; Y : smoking; L : heart disease

Association and Causation

Box indicates conditioning

- Conditioning on B and L block the paths $A \to B \to Y$ and $A \leftarrow L \to Y$.
- A and Y are d-separated after conditioning on B and L. Therefore, they are conditionally independent given B and L, even though they are marginally associated in both graphs.
- E.g. (left), A : smoking; B : tar deposits in lung; Y : lung cancer

Association and Causation



- Conditioning on collider L or its descendent C *opens* the path $A \to L \leftarrow Y$, which is blocked otherwise.
- A and Y are d-connected after conditioning on L and C. Therefore, they are conditionally associated given L and C, even though they are marginally independent.
- E.g. (right), A : family heart disease history; Y : smoking; L : heart disease; C : taking heart disease medication

Association and Causation

In summary, there are three structural reasons why two variables may be associated:

1. One causes the other[40]
2. They share common causes
3. The analysis is conditioned on their common effects[41]

[40] either directly or through mediating variables, i.e. there exists an open causal path that connects the two variables.

[41] or the consequences of the common effects.

Confounding

- When x and y share common causes, they are correlated even if they do not cause each other. This makes it harder for us to learn the causal effect of x on y or vice versa. We call this problem **confounding**. The common causes are called **confounders**. In the presence of confounding, $p(y|x) \neq p(y|\text{do}(x))$.
- Self-selection bias is an important type of confounding. Consider treatment x and outcome y. When x is selected based on the values of z, if z also has a causal effect on y, then z is a confounder and there is self-selection bias.
- When z is fully observed, we say there is **selection on observables** (or, there exists **no unmeasured confounding**). When z is not fully observed, there is **selection on unobservables** (or, there exists **unmeasured confounding**).

Identifiability of Causal Effects

- Given a model $\mathcal{M}$ with variables $x = (x_1, \ldots, x_n)$ and a set of *observed* random variables $v = (v_1, \ldots, v_m)$, let $\theta(\mathcal{M}) = g_{\mathcal{M}}(x)$ be a function of x according to $\mathcal{M}$, we say $\theta(\mathcal{M})$ is **identifiable** if it can be uniquely determined based on observations of v[42].

- Thus, let $\mathcal{M}$ be a causal model with variables x and let θ be the causal effect of, say x_i on x_j, according to $\mathcal{M}$. Let v be our *observed* variables – the data we observe are a random sample $\mathcal{D} \sim p(v)$ – then θ is identifiable if it can be uniquely determined from $\mathcal{D}$[43].

---

[42] Technically, θ is identifiable if the map from the space of its possible values to the space of probability distributions of observables is invertible.

[43] We have said nothing about the size of $\mathcal{D}$. In particular, the identification question can be phrased as: given infinite data – if we actually know the true distribution of the observed variables – can we learn θ?

Intervention in Causal Diagrams

Given a causal model $\mathcal{M}$ with variables $x = (x_1, \ldots, x_n)$,

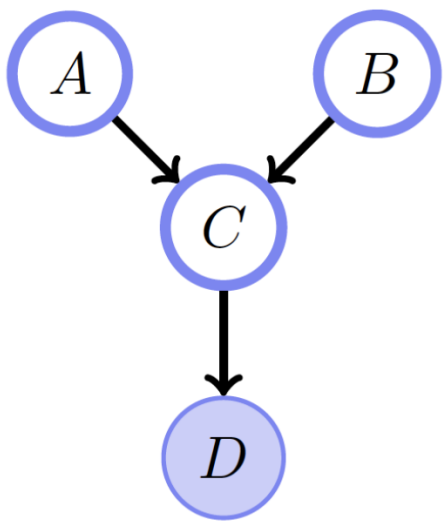
$$p(x) = \prod_{j=1}^{n} p\left(x_j \middle| \operatorname{pa}(x_j)\right) \tag{4}$$

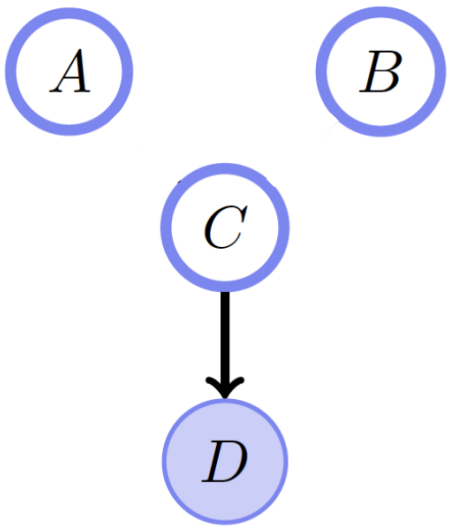
The effect of an intervention $\operatorname{do}(x_i = a)$ is to transform the **pre-intervention distribution** (4) into the **post-intervention distribution**:

$$\begin{aligned} p\left(x_{-i} \middle| \operatorname{do}(x_i = a)\right) &= \prod_{j \neq i} p\left(x_j \middle| \operatorname{pa}(x_j)\right) \\ &= \frac{p(x_{-i}, x_i = a)}{p\left(x_i = a \middle| \operatorname{pa}(x_i)\right)} \\ &= p\left(x_{-i} \middle| x_i = a, \operatorname{pa}(x_i)\right) p(\operatorname{pa}(x_i)) \end{aligned} \tag{5}$$

Intervention in Causal Diagrams

Graphically, the transformation amounts to removing all arrows entering x_i, while setting x_i equal to a.





$$p(A, B, C, D) = p(D|C)\, p(C|A, B)\, p(A)\, p(B)$$

$$p(A, B, C, D|\text{do}(C = c)) = p(D|C = c)\, p(A)\, p(B)$$

It is clear from the post-intervention graph that $p(A|\text{do}(C)) = p(A)$, though $p(A|C) \neq p(A)$.

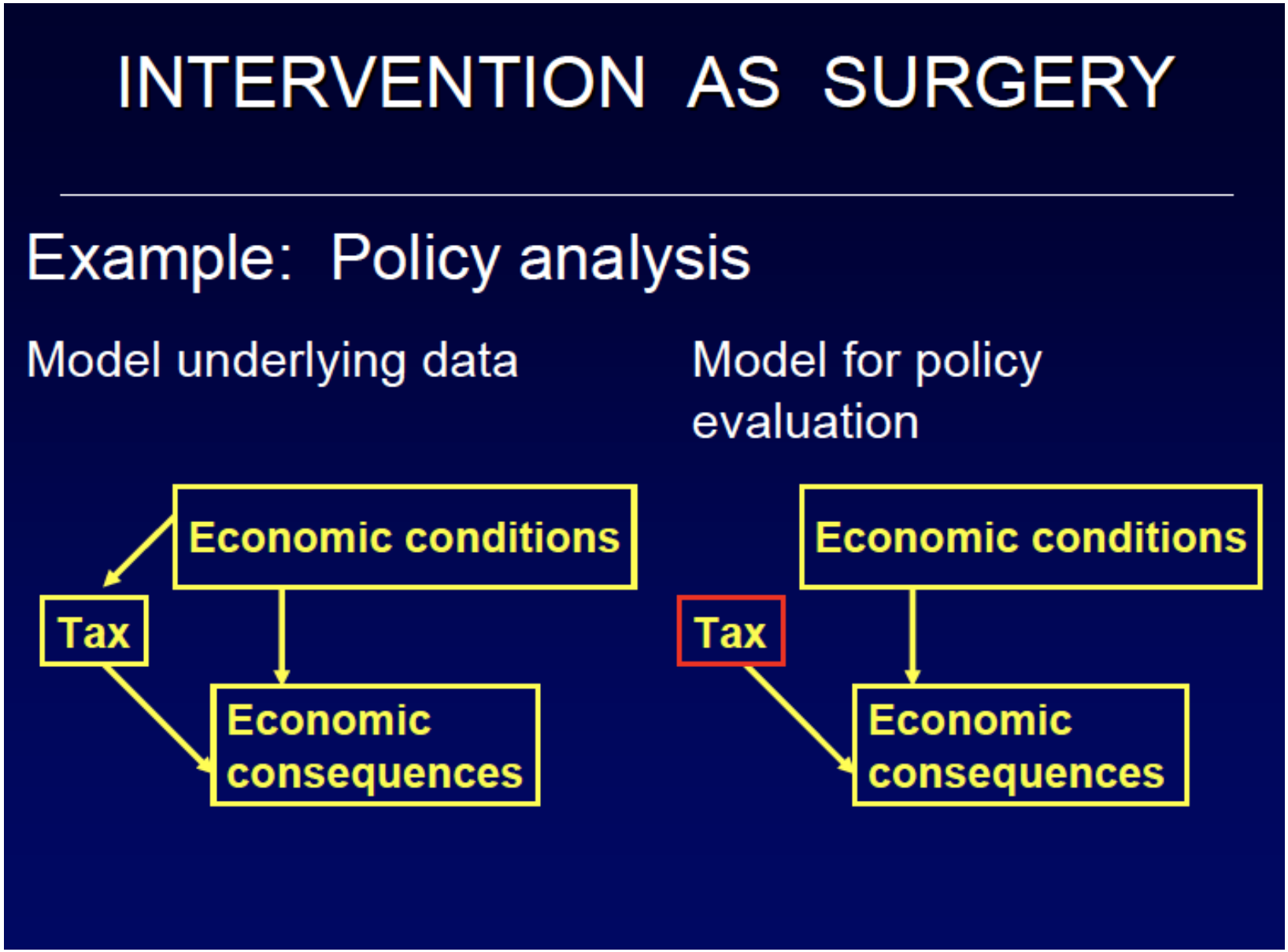
Intervention in Causal Diagrams

- The reason that $\text{do}(x_i = a)$ corresponds to removing arrows entering x_i is that before the intervention, x_i is the consequence of $\text{pa}(x_i)$. By setting $x_i = a$, we replace $\text{pa}(x_i)$ as the cause of x_i. Thus, intervention changes this local part of the causal mechanism[44].

- Since incoming arrows to x_i has been removed, the association between x_i and any other variable x_j on the graph, if exists, must be a result of x_i causing x_j, rather than of common causes. Notice also that in the post-intervention graph, since x_i no longer has parents, it has become exogenous to the model.

[44] Pearl (2009) describes intervention as "surgery on the causal diagram."

Intervention in Causal Diagrams



Causal Effect Learning

Given a causal model with variables $\{x, y, s_1 \ldots, s_n\}$, to learn the causal effect of x on y, or to make a causal prediction of y after intervening on x, it suffices to make $p(y|\text{do}(x))$ our target distribution, from which we can calculate various causal effects of interest (ATE, ATT ..) or make a causal prediction[45].

Causal effect learning = statistical learning with $E[y|\text{do}(x)]$ as the target function, or $p(y|\text{do}(x))$ as the target distribution.

[45] Pearl (2009) *defines* causal effect of x on y as $p(y|\text{do}(x))$.

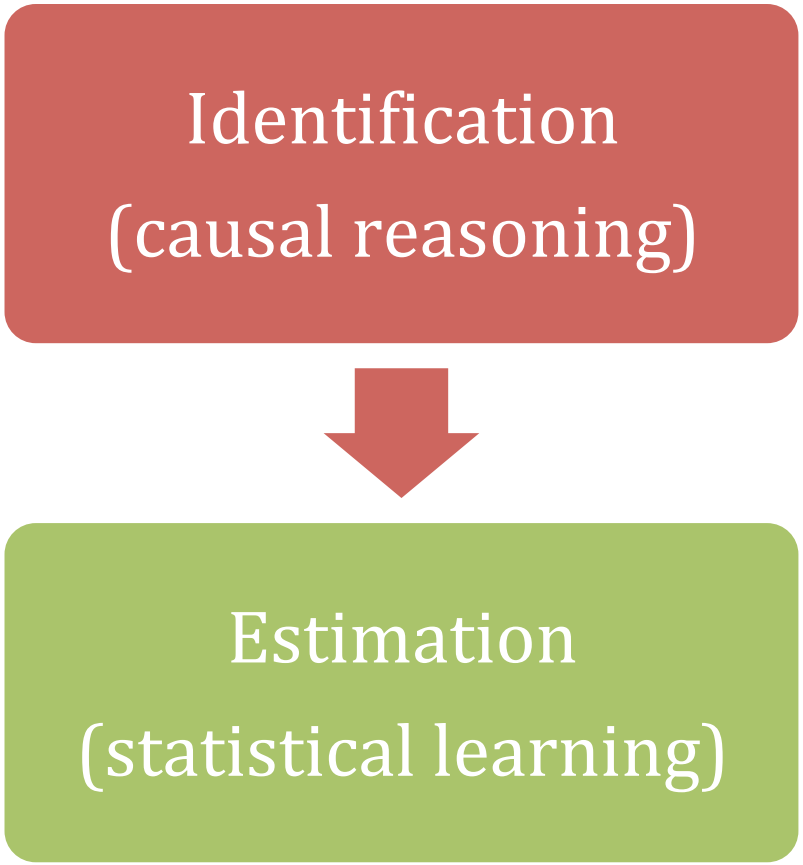
Causal Effect Learning

Note that

$$p\left(y|\text{do}\left(x\right)\right)=\int p\left(y|\text{do}\left(x\right),s^{\mathscr{M}}\right)p\left(s^{\mathscr{M}}\right)ds^{\mathscr{M}} \tag{6}$$

, where $s^{\mathscr{M}}\equiv\text{pa}\left(y\right)\backslash\left\{x\right\}$ – the causes of y other than x – are the set of *potential* effect modifiers: given any $s\in s^{\mathscr{M}}$, if the effect of s on y interacts with that of x on y, then s is an effect modifier.

Causal Effect Learning: Two Stages



Causal Effect Learning: Identification

- Suppose we are interested in learning $p(y|\text{do}(x))$. Given a causal model with variables $\{x, y, s_1 \ldots, s_n\}$, and *observed* variables v, the identification problem is whether $p(y|\text{do}(x))$ can be uniquely determined from v. Equivalently, the problem is how to express $p(y|\text{do}(x))$ in terms of $p(v)$.

- If we can express $p(y|\text{do}(x))$ in terms of $p(v)$ without making any parametric assumptions on the relationships among the variables in the causal model, then we say $p(y|\text{do}(x))$ is **nonparametrically identified** – i.e., nonparametric identification is based on knowledge of the causal structure *only* and does not rely on any statistical or functional form assumptions. On the other hand, if parametric assumptions are imposed in order to express $p(y|\text{do}(x))$ in terms of $p(v)$, then $p(y|\text{do}(x))$ is **parametrically identified**.

Causal Effect Learning: Estimation

- Once we have expressed $p(y|\text{do}(x))$ in terms of $p(v)$, say $p(y|\text{do}(x)) = g(p(v))$, then we can estimate $g(p(v))$ from the observed data using any appropriate statistical models – parametric or nonparametric[46].

- Herein lies the connection between statistical learning and causal effect learning: once we have established identification using causal reasoning based on causal models, we are left with a pure statistical learning problem: how to learn $g(p(v))$ from the observed data $\mathcal{D} \sim p(v)$.

[46] Hence, we could have nonparametrically identified–nonparametrically estimated causal effect, nonparametrically identified–parametrically estimated causal effect, etc.

Causal Effect Learning from Randomized Experiments

- Perhaps the simplest way to learn $p(y|\text{do}(x))$ is to directly sample from $p(y|\text{do}(x))$ – this is what an RCT does[47].
- An RCT that randomly assigns values to treatment x and records the outcomes y generates data $\mathcal{D} \sim p(x, y) = p(y|\text{do}(x))\, p(x)$.
- Hence $p(y|\text{do}(x))$ is nonparametrically identifiable and can be estimated from $\mathcal{D}$ using any appropriate statistical models.

[47] Randomization helps ensure that the links from $\text{pa}(x)$ to x are broken.

Causal Effect Learning from Randomized Experiments

Causal Effect Learning from Observational Studies

For observational studies, given a causal model with variables $\{x, y, s_1 \ldots, s_n\}$, if all variables are observed, then (5) provides the following way for computing $p(y|\text{do}(x))$:

$$p(y|\text{do}(x = a)) = \int p(y|x = a, \text{pa}(x)) \, p(\text{pa}(x)) \, d\text{pa}(x) \tag{7}$$

Hence $p(y|\text{do}(x))$ is nonparametrically identifiable and can be estimated by estimating $p(y|x, \text{pa}(x))$ and $p(\text{pa}(x))$ respectively from data.

Causal Effect Learning from Observational Studies

Fumigation and Yield (cont.)

Suppose we cannot do an RCT to investigate the effect of fumigation on crop yield, so instead we rely on observational data. Suppose the mechanism that generates our observed data works as follows: fumigation (F) helps control eelworms (E), which affects crop yield (Y). Farmers' use of fumigation is affected by weather (W) and the price of fumigants (C). Weather also affects yield independently. Finally, we observe the equilibrium crop price (P), which is affected by both the price of fumigants and the realized crop yield.

Fumigation and Yield (cont.)

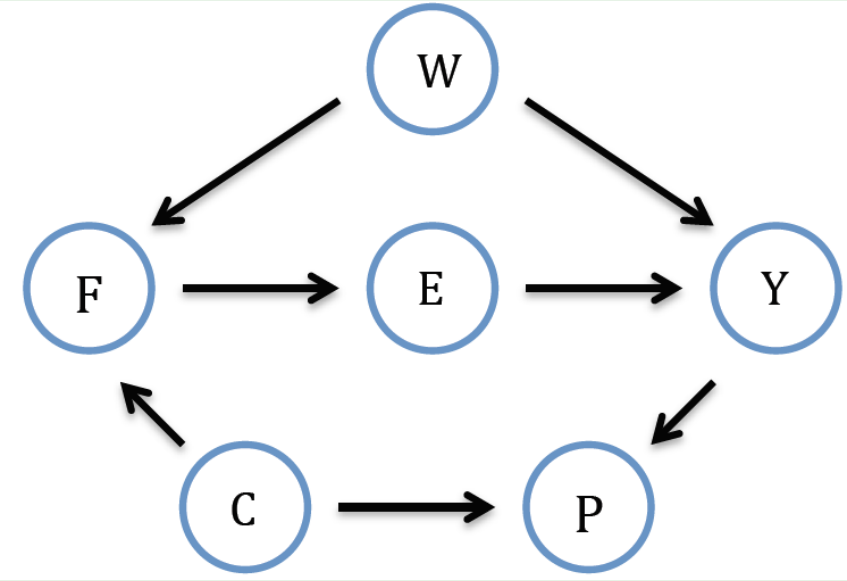


Figure 1:
Fumigation and Yield

Fumigation and Yield (cont.)

Assuming all variables are discrete, then (7) $\Rightarrow$

$$p\left(Y|\text{do}\left(F=1\right)\right)=\sum_{w,c}p\left(Y|F=1,W=w,C=c\right)\times p\left(W=w\right)p\left(C=c\right) \tag{8}$$

, where $F\in\{0,1\}$ is treated as a binary variable.

If we observe $\{F,W,C,Y\}$, then $p\left(Y|F,W,C\right),p\left(W\right),p\left(C\right)$ can all be estimated from data, from which we can calculate $p\left(Y|\text{do}\left(F=1\right)\right)$.

But what if we do not observe W or C?

Causal Effect Learning from Observational Studies

When $\text{pa}(x)$ are fully observed, $p(y|\text{do}(x))$ is *always* nonparametrically identifiable. This is no longer true when $\text{pa}(x)$ are not fully observed, in which case we have the following **identification criteria**:

1. The back-door criterion
2. The front-door criterion

These criteria provide *sufficient* conditions for the nonparametric identification of causal effects.

The Back-Door Criterion

- A **back-door path** is a path between treatment x and outcome y that has an arrow *into* x.
  - These are the paths that, if left open, induce association between x and y that is not a result of x causing y.
- A set of variables $s^{\mathscr{B}}$ satisfies the **back-door criterion** if (i) conditioning on $s^{\mathscr{B}}$ blocks every back-door path from x to y, and (ii) no node in $s^{\mathscr{B}}$ is a descendant of x.
- When $s^{\mathscr{B}}$ meets the back-door criterion and is observed[48], $p\left(y|\text{do}\left(x\right)\right)$ is nonparametrically identifiable:

$$p\left(y|\text{do}\left(x\right), s^{\mathscr{B}}\right) = p\left(y|x, s^{\mathscr{B}}\right) \tag{9}$$

$$p\left(y|\text{do}\left(x\right)\right) = \int p\left(y|x, s^{\mathscr{B}}\right) p\left(s^{\mathscr{B}}\right) ds^{\mathscr{B}}$$

[48] along with x and y, i.e., the observed variables include $\left\{x, y, s^{\mathscr{B}}, \ldots\right\}$

The Back-Door Criterion

- x is said to be **exogenous** to y if there is no open back-door path from x to y, in which case $p\left(y|\text{do}\left(x\right)\right) = p\left(y|x\right)$.
- x is **conditionally exogenous** to y if x is exogenous to y after conditioning on s, in which case $p\left(y|\text{do}\left(x\right), s\right) = p\left(y|x, s\right)$.
- When x is (conditionally) exogenous to y, individual units with different values of x are (conditionally) exchangeable with respect to y, in which case an observational study resembles a (conditionally) randomized experiment[49].
- Exchangeability and conditional exchangeability are therefore coded in graph language as the lack of open back-door paths between the treatment and outcome.

[49] in which case we also say that the treatment assignment mechanism is **ignorable**.

The Back-Door Criterion

Fumigation and Yield (cont.)

The back-door path $F \leftarrow C \rightarrow P \leftarrow Y$ is already blocked by the collider P. Therefore, by the back-door criterion, we only need to condition on W[a]:

$$p\left(Y|\text{do}\left(F=1\right)\right)=\sum_{w} p\left(Y|F=1,W=w\right)p\left(W=w\right) \tag{10}$$

Estimation using this strategy requires only the observation of $\{F, Y, W\}$.

---

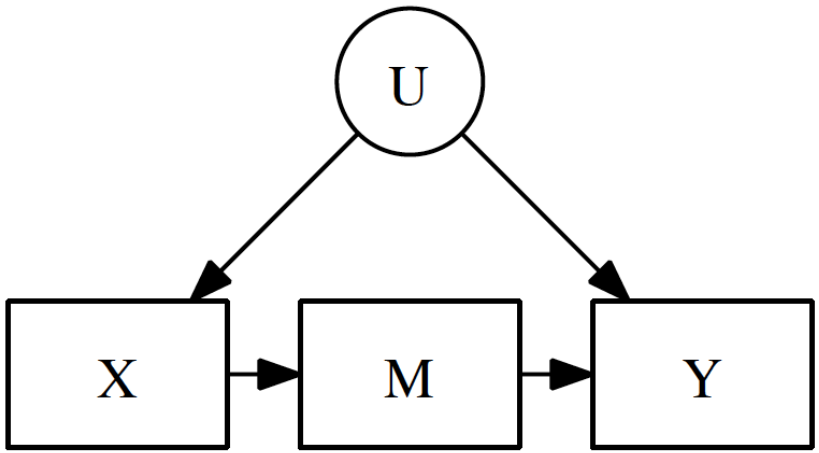
[a] (10) can be derived from (8) by noticing that conditioning on F blocks the back-door path from C to Y. Hence $p\left(Y|F,W,C\right)=p\left(Y|F,W\right)$.

The Front-door Criterion

- If we can establish an isolated and exhaustive mechanism that relates x to y, then the causal effect of x on y can be calculated as it propagates through the mechanism.
- A set of variables $s^{\mathscr{F}}$ satisfies the **front-door criterion** when (i) conditioning on $s^{\mathscr{F}}$ blocks all causal paths from x to y; (ii) no open back-door paths exist from x to $s^{\mathscr{F}}$, i.e. x is exogenous to $s^{\mathscr{F}}$; (iii) conditioning on x blocks all back-door paths from $s^{\mathscr{F}}$ to y, i.e. $s^{\mathscr{F}}$ is exogenous to y conditional on x.
- When $s^{\mathscr{F}}$ meets the front-door criterion and is observed, $p\left(y|\text{do}\left(x\right)\right)$ is nonparametrically identifiable:

$$\begin{aligned} p\left(y|\text{do}\left(x=a\right)\right) &= \int p\left(s^{\mathscr{F}}\middle|\,\text{do}\left(x=a\right)\right) p\left(y|\text{do}\left(s^{\mathscr{F}}\right)\right) ds^{\mathscr{F}} \qquad (11) \\ &= \int p\left(s^{\mathscr{F}}\middle|\,x=a\right)\left(\int p\left(y\middle|s^{\mathscr{F}},x\right) p\left(x\right) dx\right) ds^{\mathscr{F}} \end{aligned}$$

The Front-door Criterion



The causal path $X \to M \to Y$ represents an *isolated* (not affected by U) and *exhaustive* (only causal path from X to Y) mechanism. All of the effect of X on Y is mediated through X's effect on M. M's effect on Y is confounded by the back-door path $M \leftarrow X \leftarrow U \to Y$, but X blocks this path. So we can use back-door adjustment to find $p\left(Y|\text{do}\left(M\right)\right)$ and directly find $p\left(M|\text{do}\left(X\right)\right) = p\left(M|X\right)$. Putting these together gives $p\left(Y|\text{do}\left(X\right)\right)$.

The Front-door Criterion

Fumigation and Yield (cont.)

The causal path $F \to E \to Y$ constitutes an isolated and exhaustive mechanism. Hence E satisfies the front-door criterion. (11)$\Rightarrow$

$$p\left(Y|\text{do}\left(F=1\right)\right)=\sum_{e}p\left(E=e|F=1\right)\times\left[p\left(Y|E=e,F=0\right)p\left(F=0\right)\right.$$
$$\left.\times\ p\left(Y|E=e,F=1\right)p\left(F=1\right)\right]$$

Estimation using this strategy requires only the observation of $\{F, Y, E\}$.

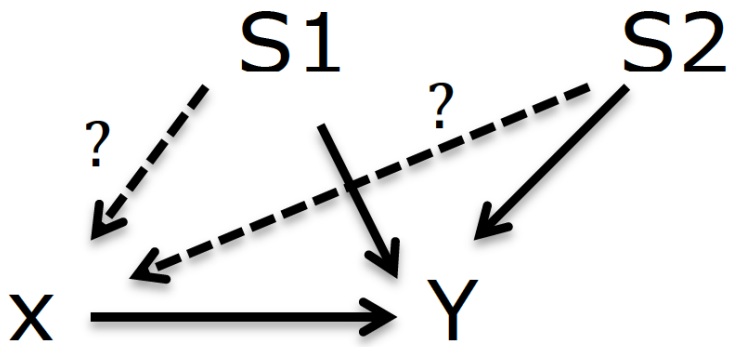
The Experimental Ideal, Revisited

- The adjustment formula (7) as well as the back-door and front-door criteria allow nonparametric identification of causal effects from observational data.

- Similarly, RCTs allow nonparametric identification of causal effects, but require even *fewer* assumptions. Specifically, we do not need to know the treatment assignment mechanism. In graph terms, only knowledge of the *post-intervention* causal diagram is needed.

- A major limitation of RCTs, as already discussed, is that it is hard – often impossible – to generate data from a $p(y|\text{do}(x))$ that is the result of a desired distribution of effect modifiers.

The Experimental Ideal, Revisited



For RCTs, whether $S1$ or $S2$ is a cause of X *pre-intervention* is immaterial to learning the causal effect of X on Y. As an example, suppose X is a worker training program and Y is employment outcome, $S1$ is ability and $S2$ is motivation. An RCT investigator who randomly assigns workers to training programs needs not worry about whether in non-experimental settings, workers who enroll in such programs have higher motivation or higher ability or both. However, knowledge about $S1$ and $S2$ as causes of Y is still needed and a rough idea about how they are distributed in the trial sample, even if unobserved, is necessary for defining the scope of the resulting causal effect estimate.

Instrumental Variables

- If we use observational data, but neither the back-door nor the front-door criterion is satisfied by the variables that we observe, then we may need additional parametric assumptions in order to identify the causal effect of interest.

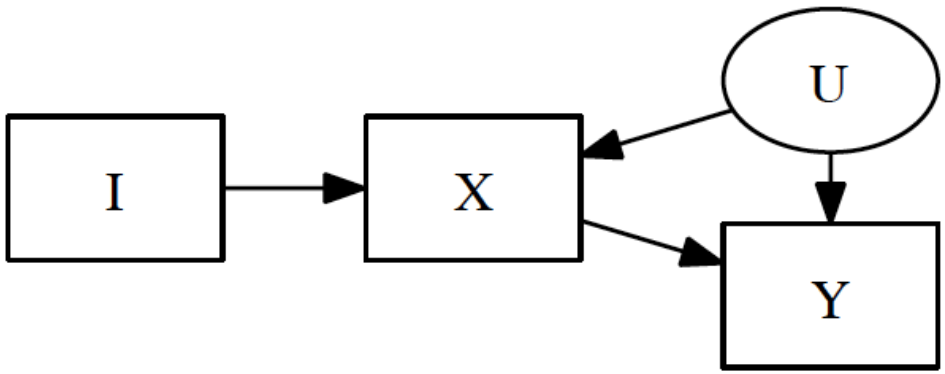


Figure 2: The causal effect of x on y is not identified if U is not observed and we assume only the causal relationships encoded in this diagram. If in addition we assume $Y = \beta X + \alpha U$, then β can be identified using I as an instrument.

Instrumental Variables

- One strategy that can be used to identify causal effects under additional parametric assumptions is the instrumental variable strategy.
- A variable $s^{\mathscr{J}}$ can serve an **instrumental variable** (**IV**) for identifying the causal effect of x on y if (i) $s^{\mathscr{J}}$ is associated with x, (ii) every open path connecting $s^{\mathscr{J}}$ and y has an arrow pointing *into* x.
  - (ii) $\Rightarrow$ $s^{\mathscr{J}}$ is d-separated and independent from any common causes of x and y, since x is a collider on their paths.
  - (i) and (ii) imply that any association between $s^{\mathscr{J}}$ and y could only be a result of the association between $s^{\mathscr{J}}$ and x and the causal effect of x on y – a condition the IV strategy relies on to identify the causal effect.

Instrumental Variables

- One of the simplest parametric assumptions under which an IV strategy works is that of a linear model: $y = \beta' x + \alpha' u$, where u are the causes of y other than x, including those that are x and y's common causes.
- Consider figure 2: assume $Y = \beta X + \alpha U + \varepsilon$, where U and ϵ are unobserved. Then $\beta \equiv \partial E\left[Y|\text{do}\left(X = a\right)\right]/\partial a$ represents the causal effect of X on Y and can be estimated by using I as an instrumental variable. By assumption, we have:

$$Cov\left(Y, I\right) = Cov\left(\beta X + \alpha U + \varepsilon, I\right) = \beta Cov\left(X, I\right)$$

$$\Rightarrow$$

$$\beta = \frac{Cov\left(Y, I\right)}{Cov\left(X, I\right)}$$

Instrumental Variables

Fumigation and Yield (cont.)

The variable C can serve as an instrument for F: it affects Y only through its effect on F[a]. Assume $Y = \alpha W + \gamma E + \epsilon$ and $E = \beta F + \varepsilon$, then $Y = \alpha W + \beta\gamma F + \gamma\varepsilon + \epsilon$. $\beta\gamma$ represents the causal effect of F on Y and can be estimated by:

$$\beta\gamma = \frac{Cov(Y, C)}{Cov(F, C)}$$

Estimation using this strategy requires only the observation of $\{F, Y, C\}$.

[a] The collider P blocks the path $C \to P \leftarrow Y$.

Non-random Sampling

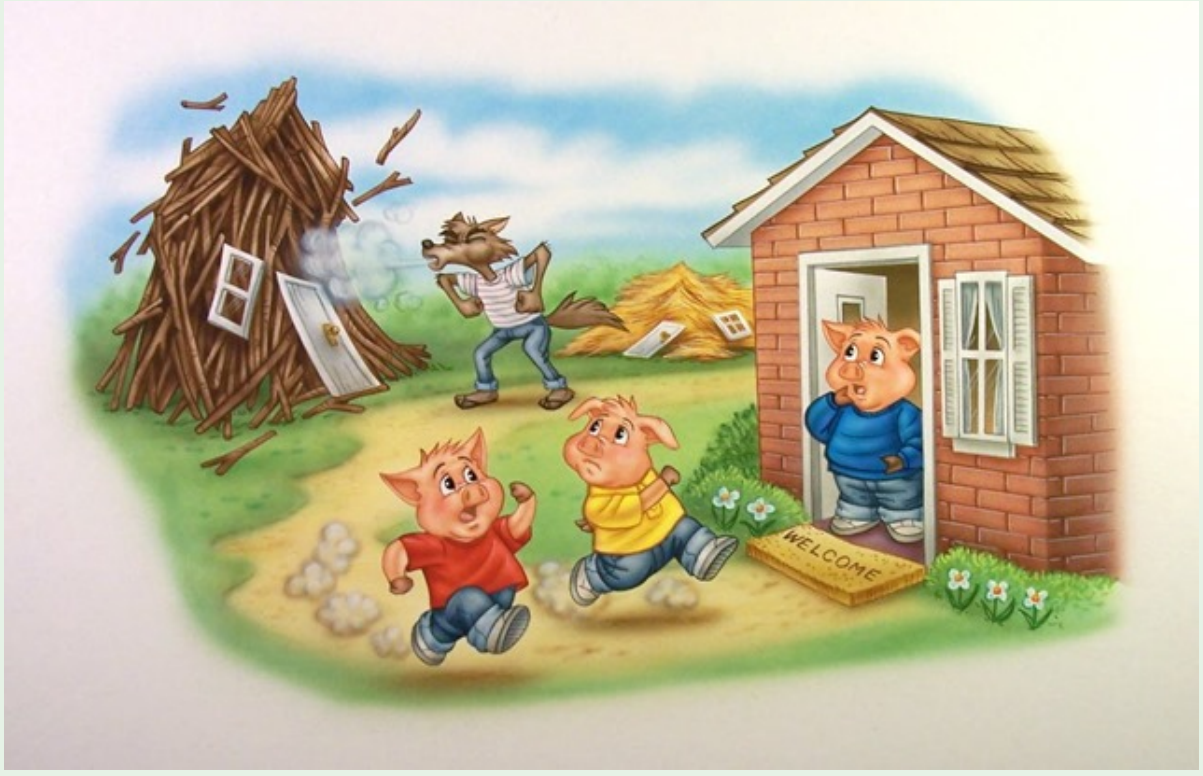
So far we have assumed that we can at least observe a random sample of $\{x, y\}$. But what if we don't?

Mutual Fund Performance

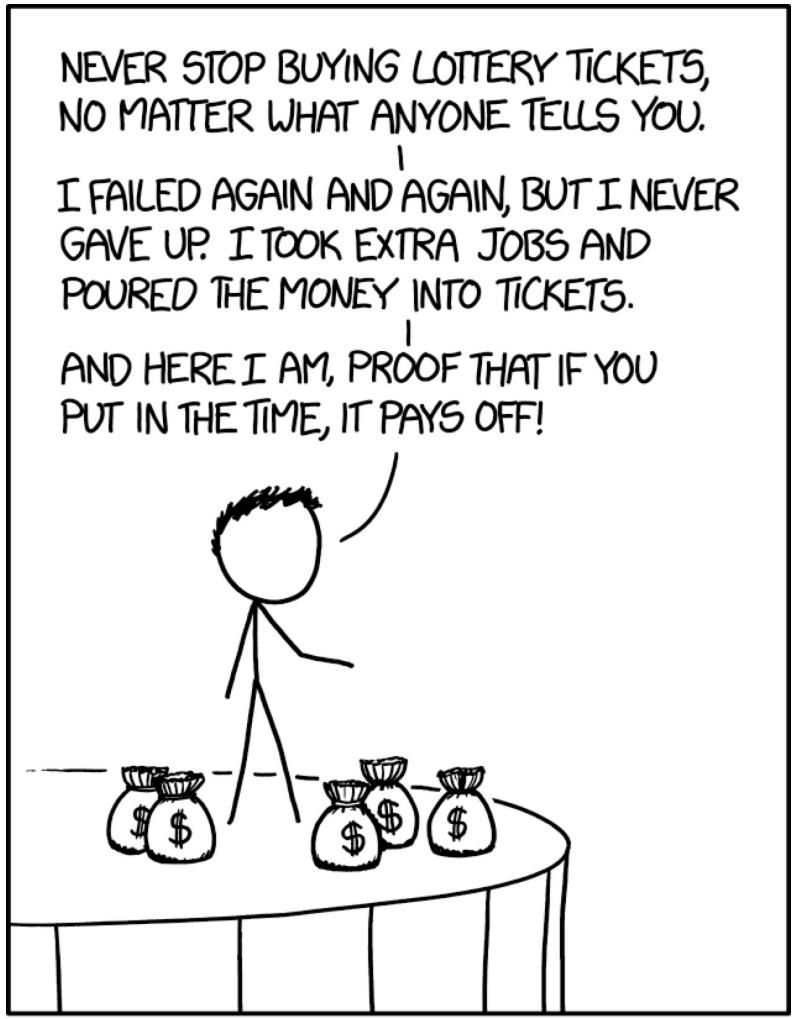
Suppose we are interested in how the size of assets under management affects a fund's performance. If we simply look at the relationship between fund size and returns among existing funds, however, there will be what is referred to as a **survival bias**: we do not observe funds that have closed due to bad performance. So if fund size negatively affects performance, we may end up *under-estimating* the magnitude of the effect.

Survival Bias

Similarly, a study of how intelligence and work ethic are related among pigs will generate biased results if we only look at pigs that survive wolf attacks (the ones that built brick houses).



Survival Bias



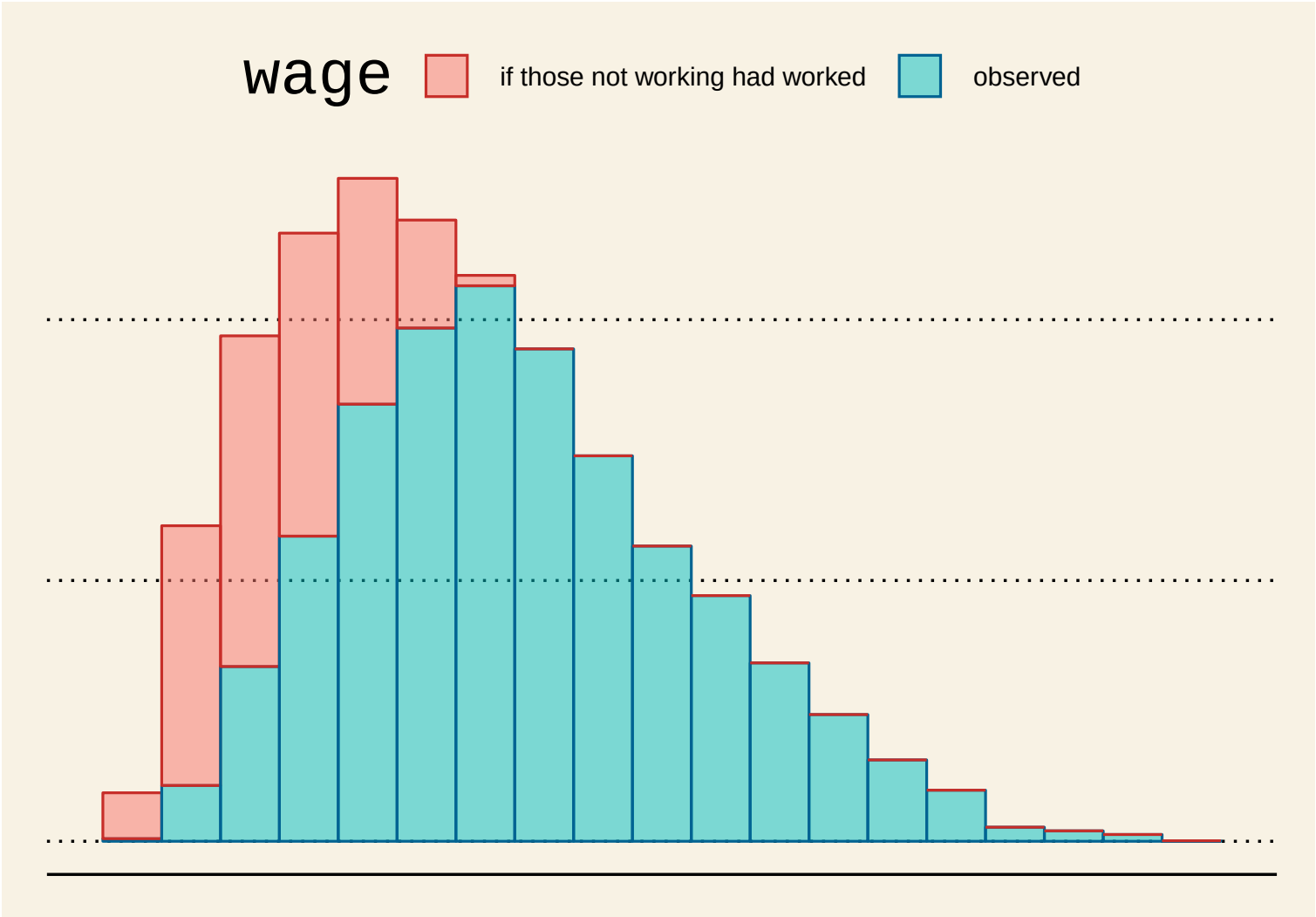
Non-random Sampling

- When the data we observe is not a random sample of the population we are interested in[50] – if we have a non-representative sample – then **sampling bias** may arise.
- Sampling bias is a general problem that can arise when we try to make inference – whether *statistical* or *causal* – about a population using data collected from another population.
- **Sample-selection bias** is a type of sampling bias that arises when we try to make inference about a *larger* population from a sample that is drawn from a distinct subpopulation.
  - Survival bias is a special type of sample-selection bias.
  - More generally, sample-selection bias can be thought of as a **missing data problem**, where data are *not missing at random* (*NMAR*)[51].

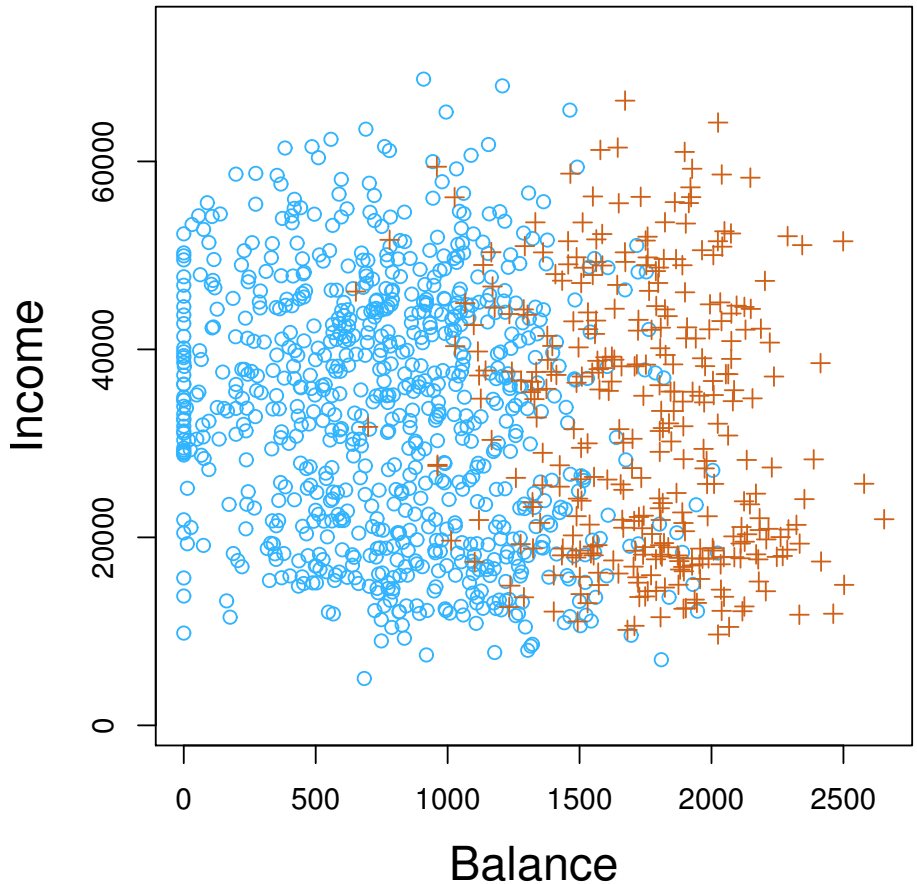
[50] Equivalently, the study population is different from the target population.

[51] As opposed to *missing at random* (*MAR*)

Non-random Sampling



Non-random Sampling



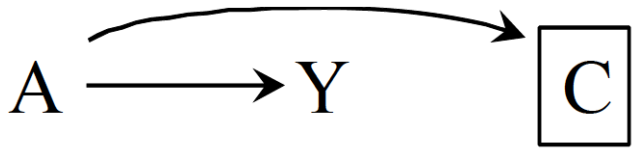
Income, balance, and default status for credit card holders

Can the relationship learned from this data set be used to predict default rate for a random credit card applicant?

Sample Selection Bias

- Let $c \in \{0, 1\}$ indicate whether an individual unit belongs to the subpopulation. Analyses based on samples drawn from the subpopulation can be thought of as analyses using the whole population, but *conditional on* $c = 1$.
- The reason that doing so may produce biased learning about the whole population is that the subpopulation with $c = 0$ and the subpopulation with $c = 1$ are not exchangeable.
- Suppose we want to learn $p(y)$ in the whole population, then c can be considered as a treatment itself: $p(y) = p(y|\text{do}(c = 1))$, i.e. the distribution of y when everyone is "assigned" to the "observed group."
- Hence *nonparametric* identification of $p(y)$ relies on whether we can make c exogenous to y, which can be done by blocking all backdoor paths between c and y.

Sample Selection Bias



A: income; $Y \in \{0, 1\}$: default; $C \in \{0, 1\}$: credit card holder

---

Suppose credit card companies determine whether to accept (C) credit card applications based solely on income (A). Once a person is issued a credit card, income determines the probability of her default (Y).

Assume we observe A for both $C = 0$ and $C = 1$, but only observe Y for $C = 1$[52].

[52] This is a case of **censoring**, where, given a random sample of individuals drawn from the population of interest, some variables – in particular the outcome variable – are observed only on individuals belonging to a subpopulation, while other variables are observed on all individuals in the sample. If *all* variables are observed only on individuals belonging to a subpopulation, then we have **truncation**. Truncation entails greater information loss than censoring.

Sample Selection Bias

- If the goal is to learn $p(Y)$, then since $p(Y) = p(Y|\text{do}(C=1)) \neq p(Y|C=1)$, there is sample selection bias[53]. To correct the bias, notice that A satisfies the back-door criterion. Hence,

$$p(Y) = p(Y|\text{do}(C=1)) = \int p(Y|C=1,A)\,p(A)\,dA$$

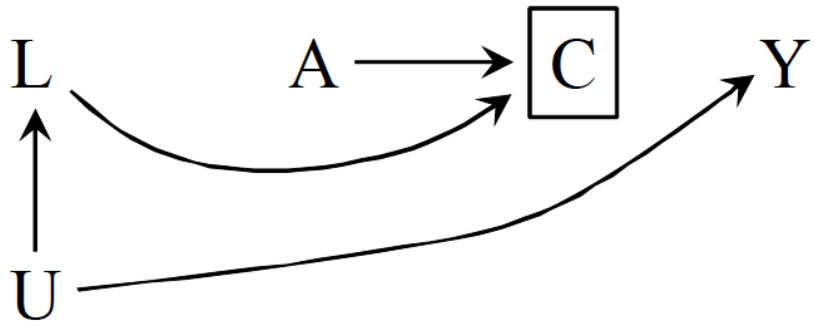
- If the goal is to learn $p(Y|\text{do}(A))$, then

$$p(Y|\text{do}(A)) = p(Y|\text{do}(A, C=1)) = p(Y|\text{do}(A), C=1)$$

, i.e. there is no sample selection bias here in learning the causal effect of A on Y.

[53] As this example demonstrates, an understanding the underlying causal mechanism is needed for non-causal statistical inferences as well if we want to extrapolate the results from one population to another.

Sample Selection Bias



Suppose the goal is to learn $p(Y|\text{do}(A))$. Notice that C is a collider on the path $A \to C \leftarrow L \leftarrow U \to Y$. Without conditioning on C, the path is blocked. Conditioning on $C = 1$ *opens* the path, making A and Y associated even though they have no causal effect on each other, leading to sample selection bias in learning $p(Y|\text{do}(A))$.

Sample Selection Bias

Assume we observe $\{A, L\}$ for both $C = 0$ and $C = 1$, but only observe Y for $C = 1$. Since conditioning on L blocks all back-door paths from A to Y and from C to Y, we have:

$$\begin{aligned} p\left(Y|\text{do}\left(A\right)\right) &= p\left(Y|\text{do}\left(A, C = 1\right)\right) \\ &= \int p\left(Y|A, L, C = 1\right) p\left(L\right) dL \end{aligned}$$

, i.e., $p\left(Y|\text{do}\left(A\right)\right)$ is nonparametrically identifiable and can be estimated by estimating $p\left(Y|A, L, C = 1\right)$ and $p\left(L\right)$ from data respectively.

Causal Mechanism Learning

- Our discussion so far focuses on learning a causal effect based on our prior knowledge of a causal mechanism. But how do we learn causal mechanisms[54]?
- Conceptually, we begin with a hypothesis set containing causal models. The learning problem is to choose the one that best fits our observed data.
- Recall that a causal model contains a specification of the joint distribution and the causal structure. A causal model is identifiable if it can be uniquely determined (out of the hypothesis set of causal models) once we observe the entire population.

[54] The distinction is between understanding the "effects of causes" (causal effect learning) and understanding the "causes of effects" (causal mechanism learning).

Causal Mechanism Learning

- Often, our interest lies mainly in learning the causal structure[55]. Since each causal structure implies, statistically, a set of conditional independence relations, their identifiability depends on whether they can be uniquely determined by the set of conditional independence relations observed in the population.

[55] As our discussion on meaningful causal effects shows, causal effects are often as unstable as statistical relationships. Both can change from population to population even though the underlying causal relationships are the same. Causal relationships are more stable than causal effects and statistical relationships. That's why our knowledge about the physical world is largely encoded and transmitted in the qualitative language of causal relationships ("pushing the glass off the table will cause it to break"), rather than the quantitative language of causal effects and statistical relationships ("pushing the glass off the table will result in a 95% probability of breakage.").

Causal Mechanism Learning

- If two causal structures are *observationally equivalent*[56], then they cannot be distinguished without resorting to manipulative experimentation or temporal information. Hence, experiments play an important role in identifying observationally equivalent causal structures.

[56] See Appendix I

Causal Mechanism Learning

- In practice, we (human beings) have been learning causal mechanisms by formulating models, then conducting experiments or observational studies, and based on the results of which, updating our belief about each model's probability of being true. This, in essence, is the process of scientific progress[57,58].

[57] The big question for AI is whether this process can be automated.

[58] This view of scientific progress as continuous Bayesian updating based on evidence has been challenged by historians like Thomas Kuhn, who pointed out that the sociological nature of the scientific community leads to periodic paradigm shifts rather than continuous progress.

Structual Estimation

- Econometrics began as a discipline that seeks to link *economic theory* to data.
- Today in the econometrics literature, causal models based on *economic theory* are referred to as **structural models**. Their estimation is called **structural estimation**.

The Birth of Econometrics

- The Econometric Society was founded on 29th December 1930 at a gathering during the annual joint meeting of the American Economic Association and the American Statistical Association.
- Ragnar Frisch, a founding member of the Econometric Society and the first editor-in-chief of *Econometrica*, wrote in 1923[59]:

  Intermediate between mathematics, statistics, and economics, we find a new discipline which for lack of a better name, may be called econometrics. Econometrics has as its aim to subject abstract laws of theoretical political economy or 'pure' economics to experimental and numerical verification, and thus to turn pure economics, as far as possible, into a science in the strict sense of the word.

[59] Frisch (1926). Quoted is an English translation of the French original. Underlining mine.

The Birth of Econometrics

- Regarding the name "Econometrics", Frisch later wrote[60]:

  So far, we have been unable to find any better word than "econometrics". We are aware of the fact that in the beginning somebody might misinterpret this word to mean economic statistics only. But ... we believe that it will soon become clear to everybody that the society is interested in economic theory just as much as in anything else.

[60] Bjerkholt (1998).

The Birth of Econometrics

- The Cowles Commission for Research in Economics was founded in 1932 and moved to the University of Chicago from 1939 to 1955[61].
  - Motto: "Theory and Measurement"[62]
- The Cowles Commission made foundational contributions to the early development of Econometrics during its Chicago years[63].
  1. Introducing the probabilistic framework as well as the methods of modern statistical inference into Econometrics
  2. Estimation of structural simultaneous equations models (SEMs).

---

[61] The commission moved to Yale in 1955 and has been renamed the Cowles Foundation.

[62] According to Cowles' own statement: "This motto replaced the original Cowles Commission motto 'Science is Measurement,' reflecting the importance of theory that became clear early in the history of Cowles."

[63] See Haavelmo (1943, 1944), Koopmans (1950), Hood and Koopmans (1953).

The Birth of Econometrics

- Thus at its inception, Econometrics was conceived as "a branch of economics in which *economic theory* and *statistical method* are fused in the analysis of numerical and institutional data" (Hood and Koopmans 1953)[64].

[64] Emphasis mine

Structual Estimation

- A complete structural model may specify preferences, technology, the information available to agents, the constraints under which they operate, and the rules of interaction among agents in market and social settings.

- More generally, we refer to any causal models that use economic theory to specify the functional form of causal relationships as structural models.

Structual Estimation

- Given a structural model $\mathcal{M}$ with variables $X = \{x_1, \ldots, x_n\}$ and causal structure $\mathcal{G}$, let $X^E \subset X$ be the variables that are exogenous to $\mathcal{G}$. The structural model specifies the statistical distributions of X^E as well as the functional forms governing $\mathcal{G}$, which together determine the joint distribution $p(X)$.

- Let $\theta \in \Theta$ be the set of parameters that parametrize the statistical distributions and functional forms in $\mathcal{M}$. The hypothesis set corresponding to $\mathcal{M}$ is therefore $\{p(X;\theta), \theta \in \Theta\}$, which we denote as $p(X;\mathcal{M})$. The goal of structural estimation is to learn θ.

Structual Estimation

- Because structural estimation learns the entire causal model, once a model is learned, we can use it to derive $p\left(x_j \middle| \operatorname{do}\left(x_i\right)\right)$ for any $\left\{x_i, x_j\right\} \subset X$.

- In contrast, the two-stage causal effect learning procedure introduced on page 100, whose goal is to learn a single causal effect, has been called **reduced-form analysis** in the econometrics literature[65].

- Difference between the two: structural estimation is a generative approach to causal inference, while reduced-form is a discriminative approach.

[65] Historically, given a structural model $g\left(x, y\right) = 0$ that specifies the relationship governing exogenous variable x and endogenous variable y, if y is solved as a function of x, i.e. $y = f\left(x\right)$, then f is referred to as the **reduced form** of g.

Structual Estimation

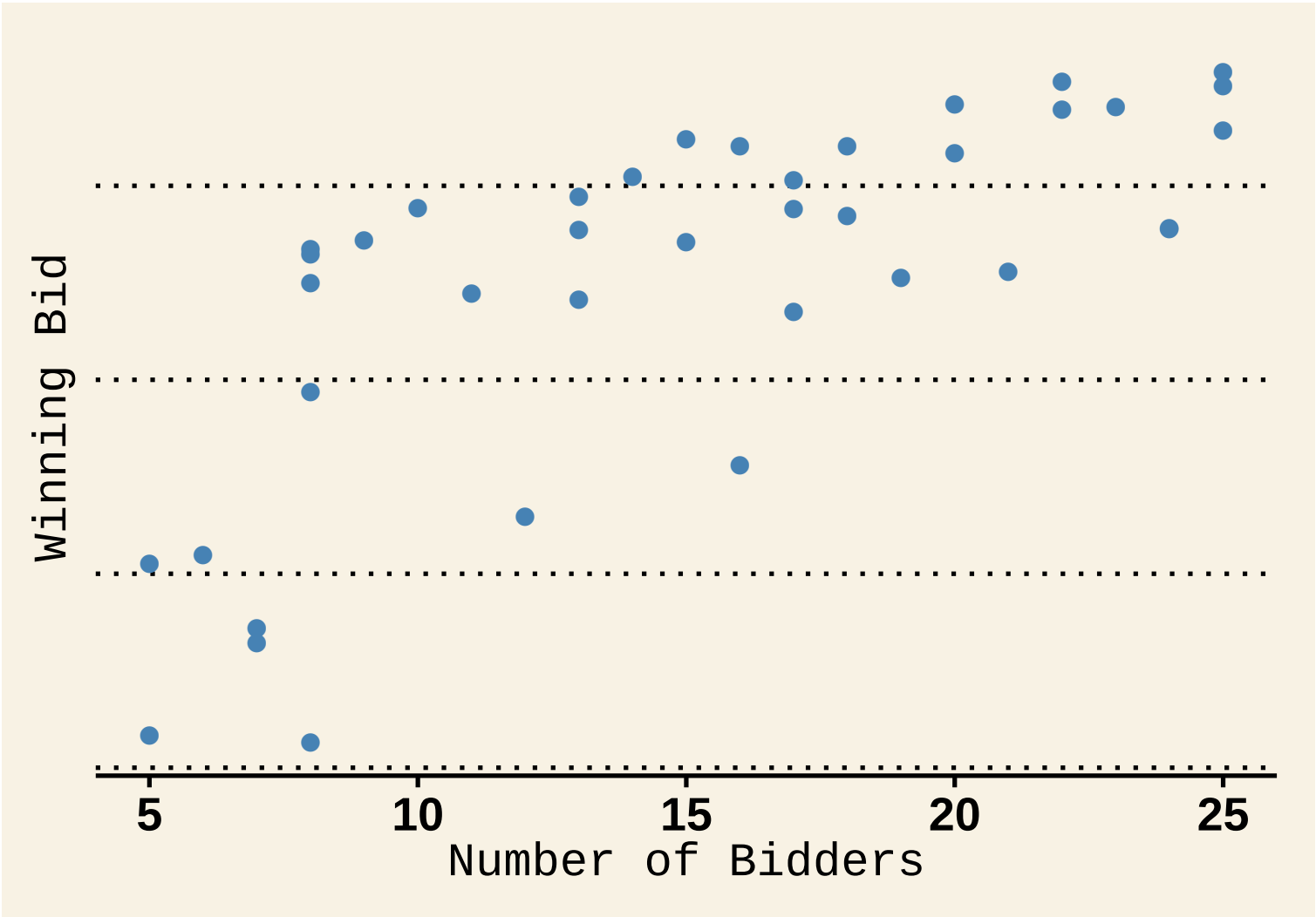
Discriminative Model

Discriminative statistical learning	$p(x_j \mid x_i)$
Causal effect learning	$p(x_j \mid \mathrm{do}(x_i))$

Generative Model

Generative statistical learning	$p(x_1, \ldots x_N; \mathcal{M})$; $\mathcal{M}$: statistical model
Structural estimation	$p(x_1, \ldots x_N; \mathcal{M})$; $\mathcal{M}$: causal model

Auction



First-price Sealed-bid Auctions for Identical Goods

Auction

Model

- N risk-neutral bidders
- Independent private value $v_i \sim^{i.i.d.} F(.)$
- Each bidder knows her own v_i and the distribution F, but not the v_i of others
- Observed bids are the Bayesian Nash equilibrium outcome of the game

$\Rightarrow$ Equilibrium bidding strategy:

$$b_i = v_i - \frac{1}{F(v_i)^{N-1}} \int_0^{v_i} F(x)^{N-1} dx$$
$$= v_i - \frac{1}{N-1} \frac{G_N(b_i)}{g_N(b_i)}$$

, where $G_N(.)$ and $g_N(.)$ are the c.d.f. and p.d.f. of the bid distribution.

Auction

Structural Estimation

1. For each auction[a], nonparametrically estimate $G_N(.)$ and $g_N(.)$ from observed bids $\{b_1, \ldots, b_N\}$.
2. For each bidder, calculate

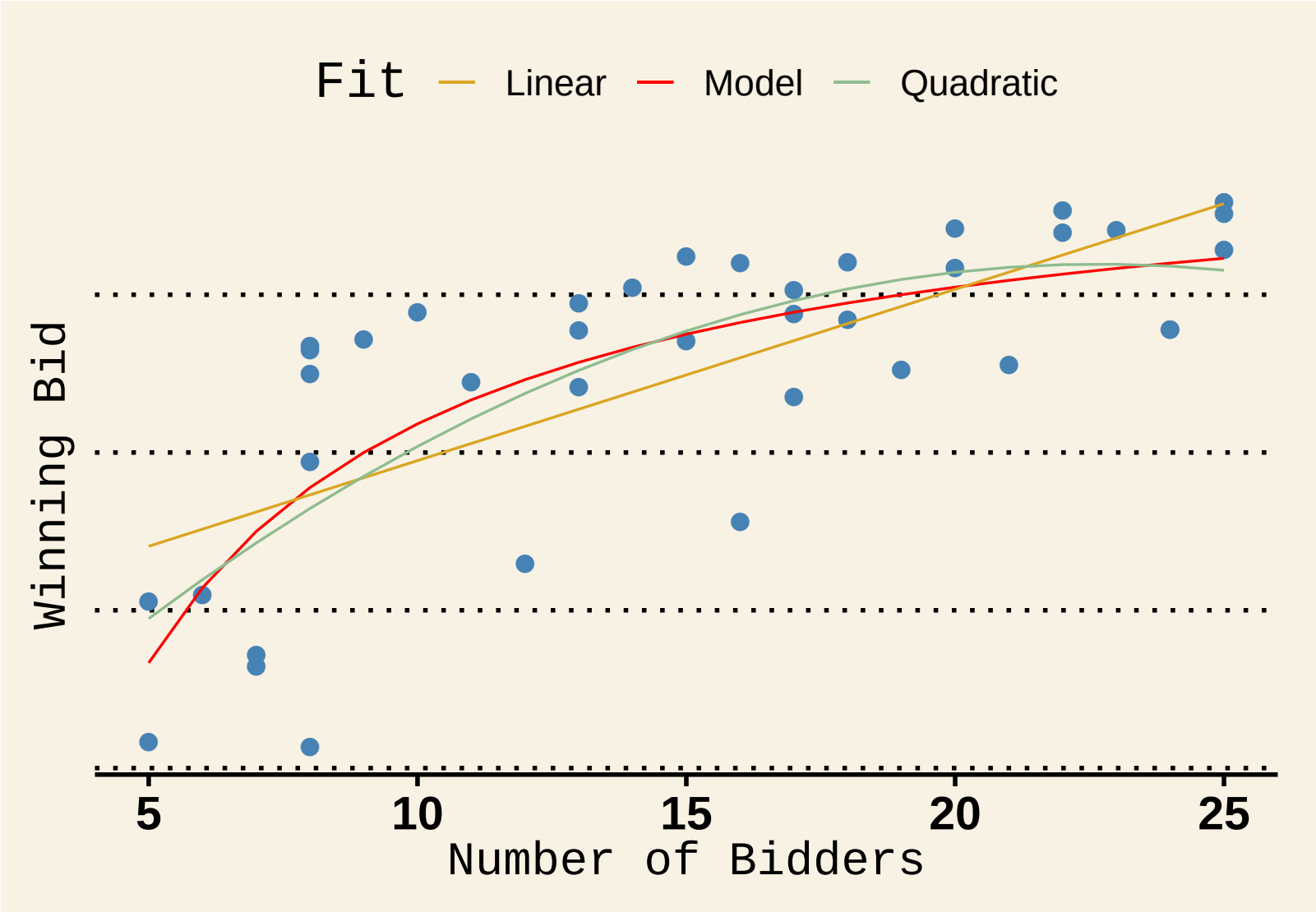
$$\widehat{v}_i = b_i + \frac{1}{N-1} \frac{\widehat{G}_N(b_i)}{\widehat{g}_N(b_i)} \tag{12}$$

3. Use $\widehat{v}_i$ to nonparametrically estimate $F(.)$
4. $\widehat{F}(.)$ can be used to predict the winning bid in an $N-$bidder auction:

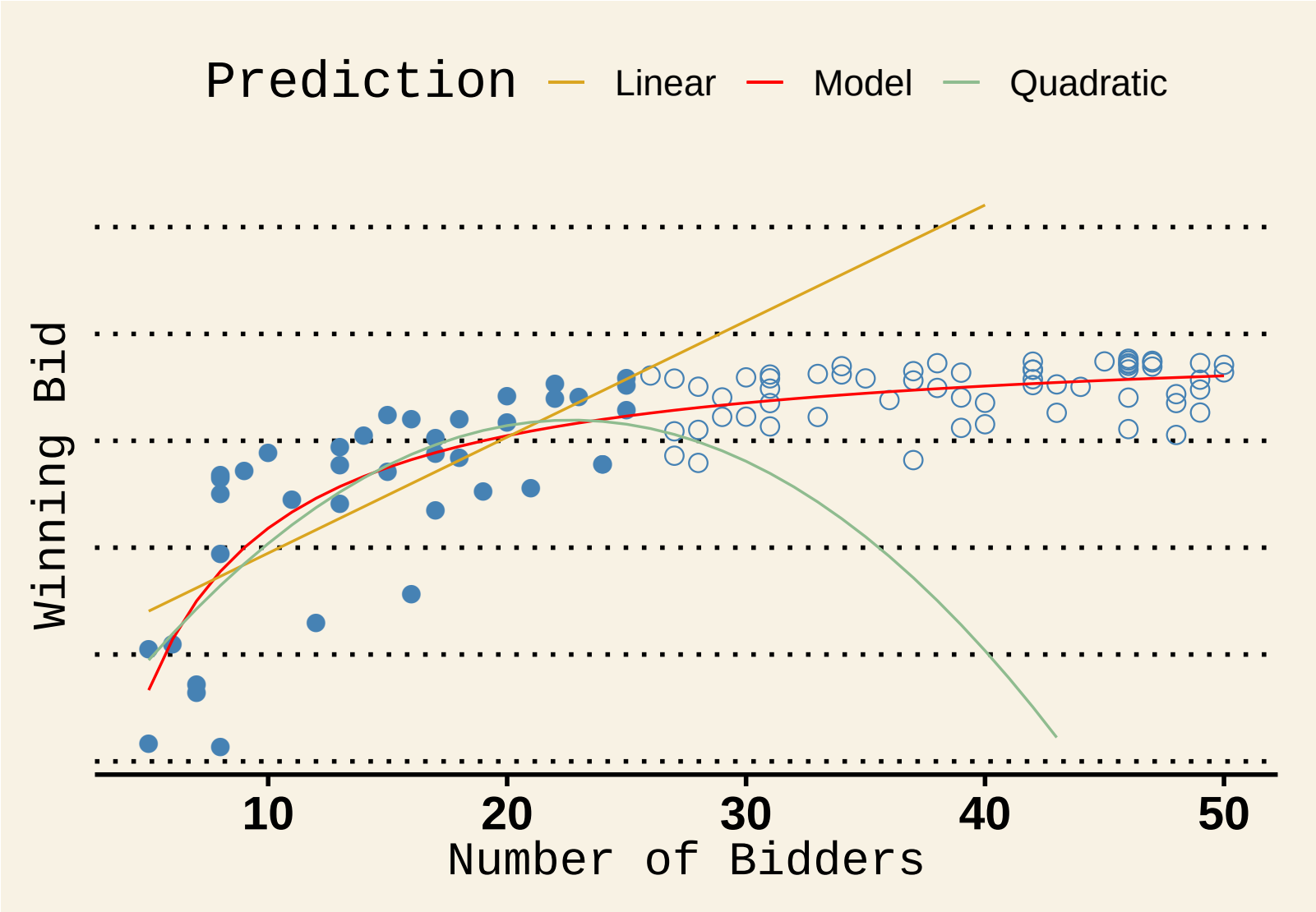
$$E[\max\{b_i\}] = E\left[\max\left\{v_i - \frac{1}{\widehat{F}(v_i)^{N-1}} \int_0^{v_i} \widehat{F}(x)^{N-1} dx\right\}\right]$$

[a] See Guerre et al. (2000).

Auction



Auction



Auction

- Here, there is no confounding between N (the number of bidders) and $b_{\max}$ (the winning bid). Hence $f(N) \equiv E[b_{\max} | N]$ *nonparametrically identifies* the effect of N on $b_{\max}$.

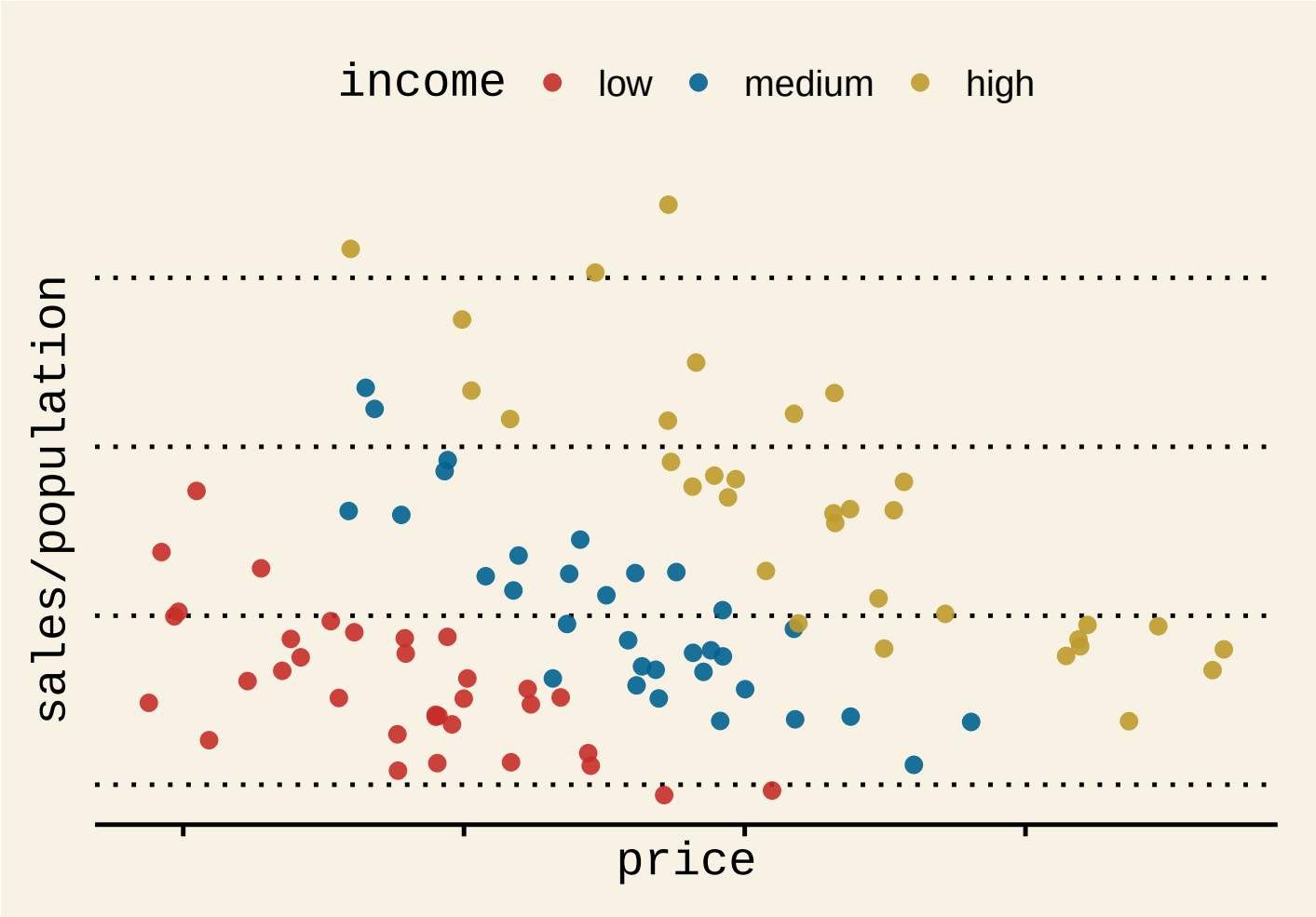
- The estimation problem is to learn $f(N)$ from data. Here, theory helps specify the functional form of $f(N)$ and therefore serves as a model selection mechanism.

- Theory also helps us to learn the values of the bidders (equation (12)) – which cannot be identified nonparametrically – by specifying the functional form of the mapping from $\{v_i\}$ to $\{b_i\}$.

Monopoly

A monopoly firm's pricing and sales in different geographical markets

Data: price, sales, average income, population for each market



Monopoly

Model: Demand

In each market m with population N_m and mean income I_m, consumers choose between the monopoly product and an outside good. Individual utilities are given by:

$$\begin{aligned} U_{i0}^m &= \epsilon_{i0}^m \\ U_{i1}^m &= \beta_0 + \beta_1 I_m - \beta_2 p_m + \epsilon_{i1}^m \end{aligned} \tag{13}$$

, where (U_{i0}^m, U_{i1}^m) are respectively the indirect utilities of the outside good and the monopoly product, and $\epsilon_{ij}^m \sim \text{Gumbel}(0, 1)$.

(13) $\Rightarrow q_m \sim \text{Binomial}(N_m, \pi_m)$, where

$$\pi_m = \frac{\exp(\beta_0 + \beta_1 I_m - \beta_2 p_m)}{1 + \exp(\beta_0 + \beta_1 I_m - \beta_2 p_m)}$$

Monopoly

Model: Supply

For each market m, given demand $q_m(p)$, the monopoly firm chooses p to maximize:

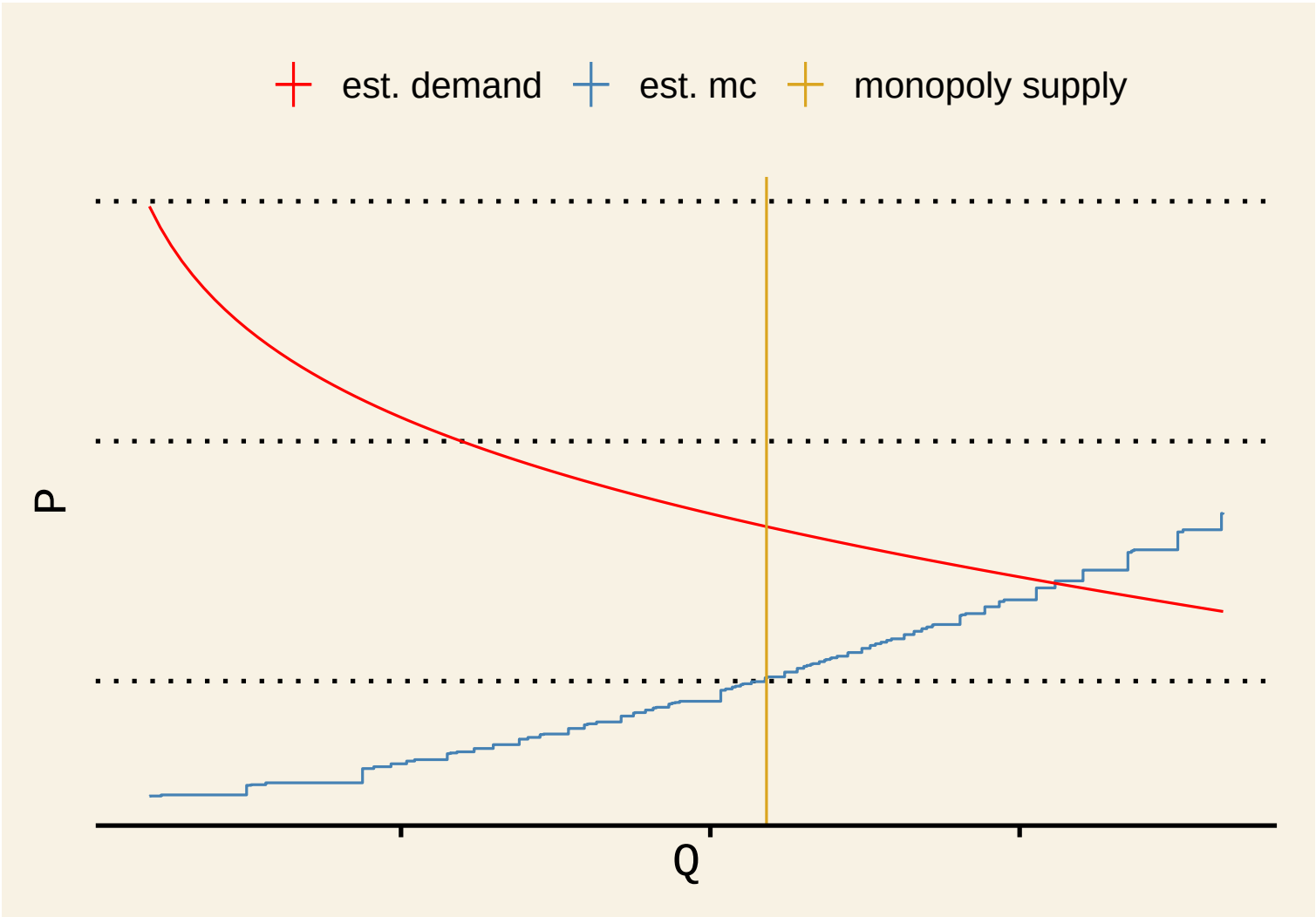
$$\max_p \{p \times q_m(p) - c(q_m(p))\} \tag{14}$$

, where $c(q)$ is the firm's cost function.

(14) $\Rightarrow$

$$c'(q_m) = p_m + [q'_m(p_m)]^{-1} q_m \tag{15}$$

Monopoly



Estimated marginal cost and demand curves
for a market with median income and population

Monopoly

- Here, theory helps us to learn the marginal cost function of the monopoly firm as well as the consumer utility function – neither of which is observed and neither can be nonparametrically identified.

- Using the estimation results, we can conduct welfare analysis and make normative statements.
  - For example, calculating the total deadweight loss due to monopoly.

Counterfactual Simulation

- One of the main benefits of learning a structural model is that it allows us to predict the effect of a completely new treatment – a treatment that has never been observed before.
  - If in the observed data, x_j is always equal to 0, what would be the effect of $\text{do}\,(x_j = 1)$?
- Because structural models are generative models, once we have learned a model, we can use it to generate synthetic data $\mathcal{D} = \{(x_{i,1}, \ldots, x_{i,j-1}, x_{i,j} = a, x_{i,j+1}, x_{i,n})\}$ from $p\left(x_1, \ldots, x_n | \,\text{do}\,(x_j = a)\right)$. This is called **counterfactual simulation**.

Counterfactual Simulation

Counterfactual Simulation

What if Chuji Qu never visited Liu's Village?

Counterfactual Simulation

Or Caesar never crossed the Rubicon?

Monopoly

What happens if the government imposes a 20% sales tax on the company?

After tax:

Δ `Consumer Surplus:`	-27.83%
Δ `Total Surplus:`	-27.95%

Tax incidence:

`Consumer:`	26.65%

Dynamic Structural Model

- In a changing environment, with new information arriving each period, individual are forward-looking when making decisions: choices are made partly based on expectations of the future.
- Decisions are also often influenced by the past. Since it can be costly to transition from one state to another, payoffs to different choices are often history-dependent: our past partly shapes our future.
- In dynamic models, treatment effects can be time-varying and it's often useful to distinguish between short-run and long-run effects.

Dynamic Structural Model

Negative Political Advertising

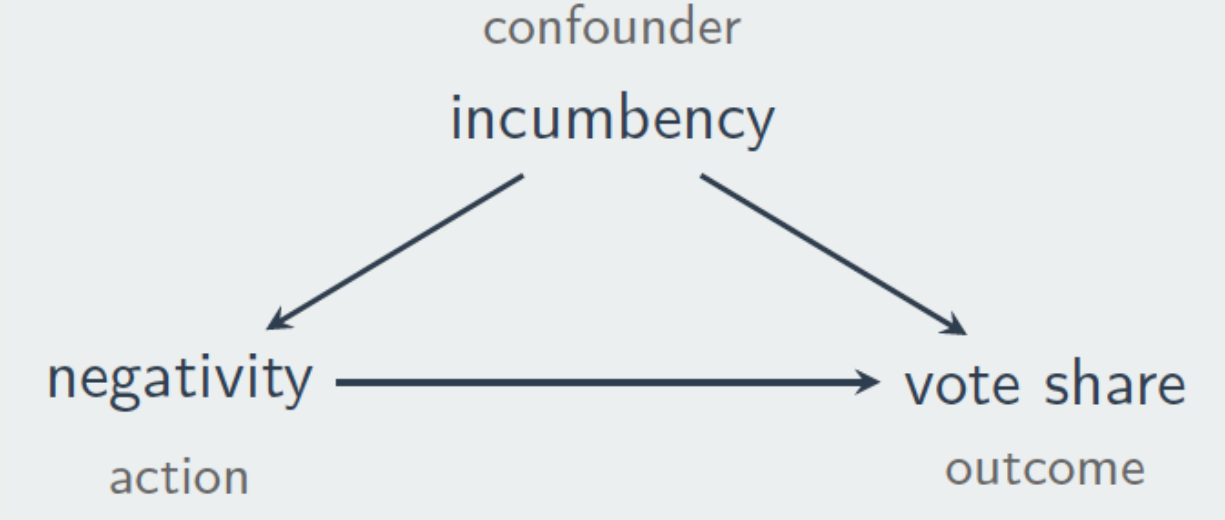
- Candidate decides whether to go negative based on polling[a].
- Going negative affects future polling, which in turn, affects future negative advertising decisions.
- Outcome: final vote share.

[a] This example is taken from Blackwell (2015).

Dynamic Structural Model

Negative Political Advertising (cont.)

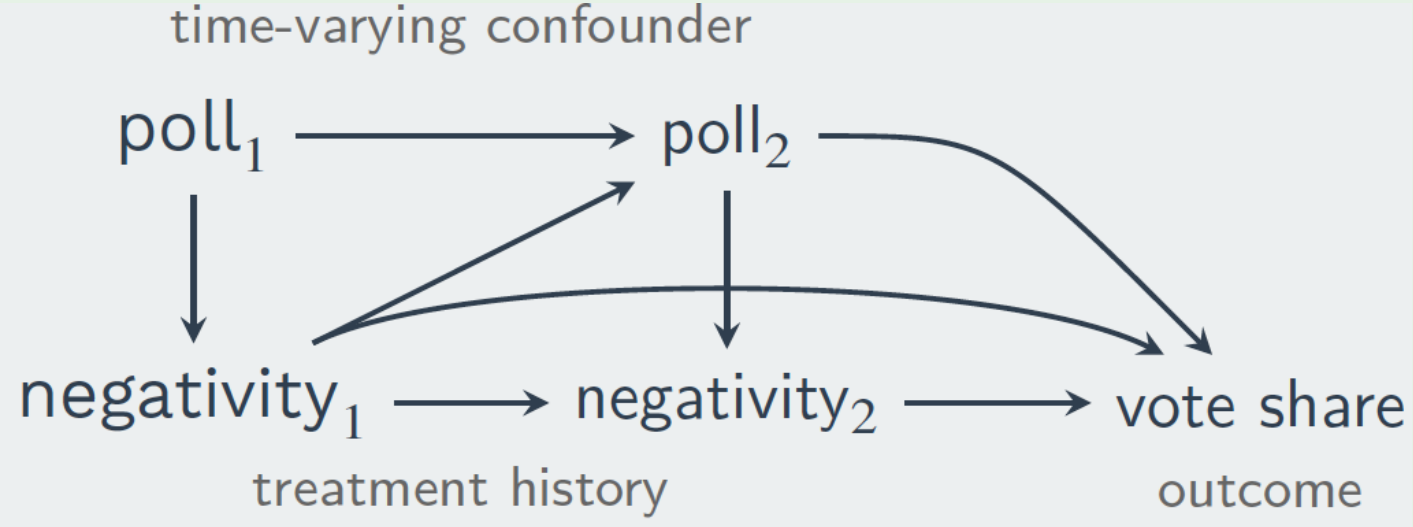
Static (single-shot) causal inference:



Dynamic Structural Model

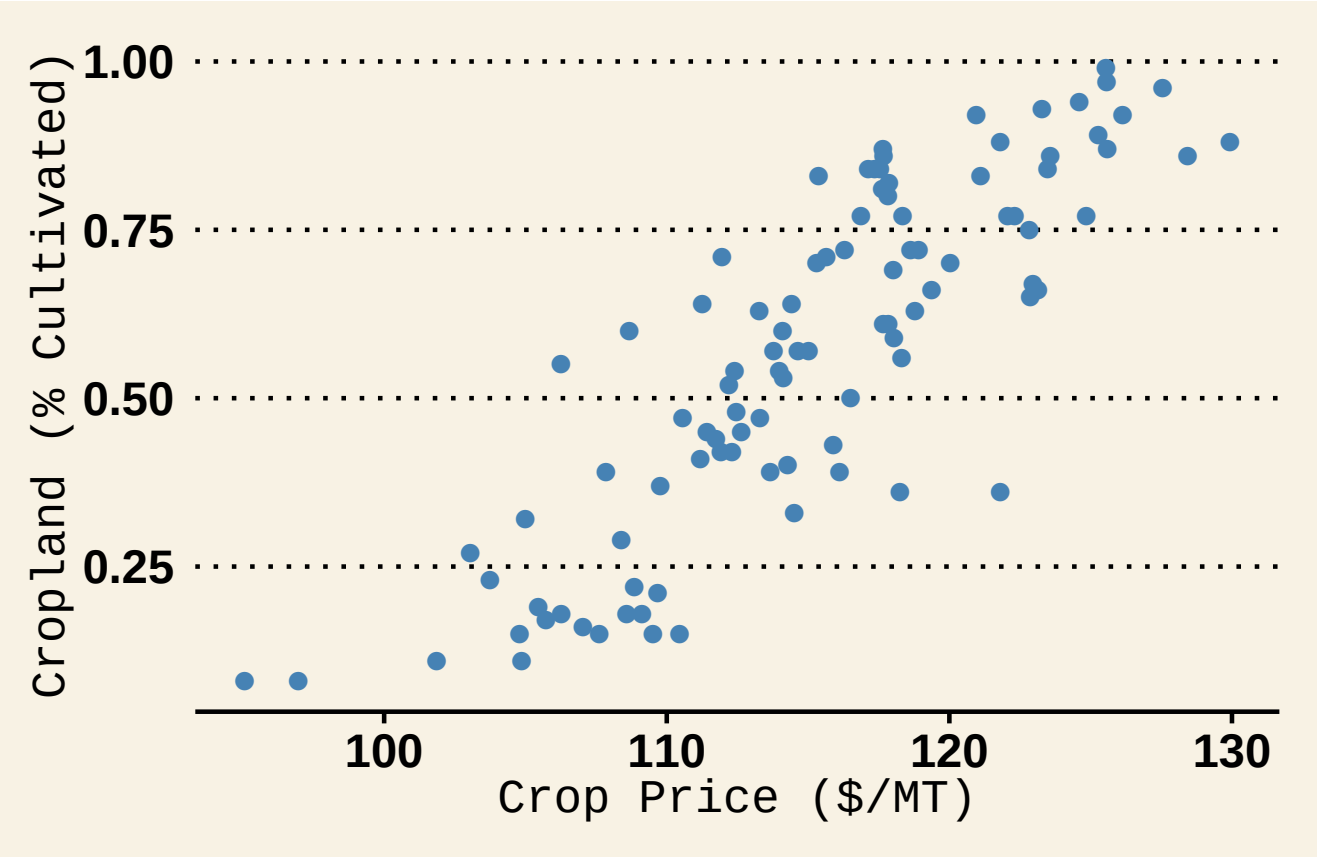
Negative Political Advertising (cont.)

Dynamic causal inference:

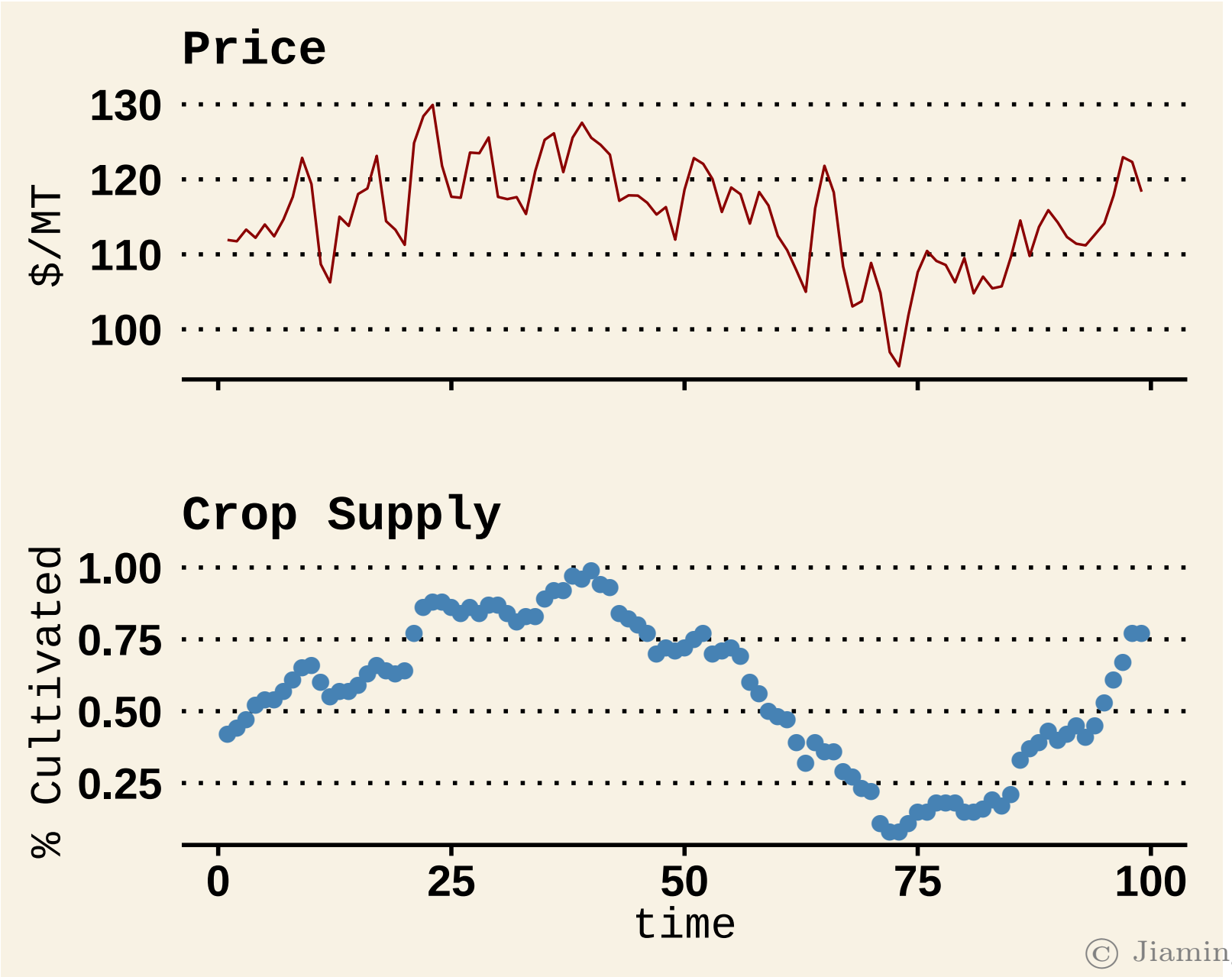


Crop Supply

Data: crop price and percentage of cropland cultivated in a county for $t = 1, \ldots, T$



Crop Supply



Crop Supply

Model

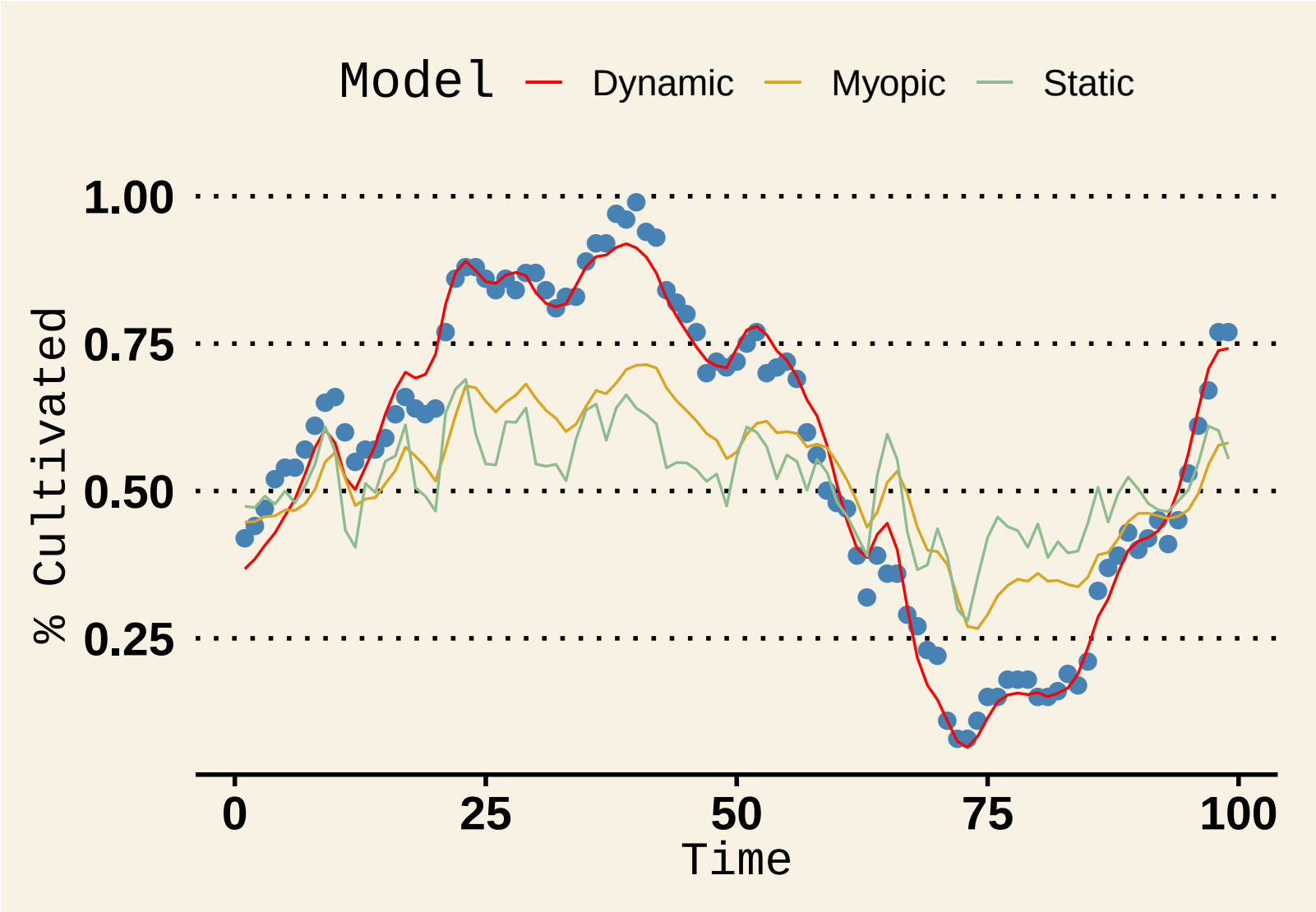
- At the beginning of each period t, each field owner decides whether or not to plant the crop in the current period.
- The decision is based on observed period$-t$ price as well as expectations of future prices.
- If a field has not been cultivated for k periods, then in order to (re)-cultivate it, the farmer needs to pay a one time cost $c(k)$.
- Farmers have **rational expectation**: their expectations of future prices are unbiased conditional on the information they have.
  - Here we assume that crop prices follow an AR(1) process, which is known to the farmers.

Crop Supply

Counterfactual Simulation:

- How would crop supply change in response to changes in crop prices if farmers are myoptic: if they are not forward-looking?
- How would crop supply change in response to changes in crop prices if farmers are static: if they are neither forward-looking, nor subject to any re-cultivation costs, so that planting decisions are made entirely based on current prices?

Crop Supply



Crop Supply

- In general, if we are interested in the effect of x on y, but x is self-selected based on expectations of y, then without any measures of such expectations, the causal effect cannot be nonparametrically identified and we need to rely on theory to specify how expectations are formed.
- Here, farmers take crop prices as given and decide whether or not to plant[66]. Models like this are called **dynamic discrete choice models**.
- If prices are endogenous – if farmers' planting decisions affect equilibrium prices[67] – then we need to model both crop supply and crop demand. Such models are called **dynamic general equilibrium models**.

[66] This is a reasonable assumption since we are looking at a single county.

[67] For example, if we look at all the farmers in a country or in the world.

Why Structural Estimation

S-0. Structural models, by *explicitly* modeling the underlying causal mechanisms, make clear what prior knowledge (assumptions) are relied upon to draw causal inference[68].

[68] There is a mistaken belief among some practitioners that structural estimation *is* causal mechanism learning. This is incorrect: structural estimation is learning based on an assumed causal mechanism.

Why Structural Estimation

By using theory to specify the functional forms of causal relationships, structural models can be used to:

S-1. identify causal effects or the values of unobserved variables that *cannot* be nonparametrically identified[69].

S-2. serve as a model selection mechanism[70] for causal effects that *can* be identified nonparametrically.

[69] For this reason, structural estimation is sometimes described as *identification by functional form*.

[70] i.e., determine the functional form of $E[y|\text{do}(x)]$.

Why Structural Estimation

S-3.1. Using structural models, what we learn from one set of data $\mathcal{D} \sim p(x, y)$ can be potentially used to explain and predict data drawn from another distribution, say $p(u, v)$, if $\{x, y\}$ and $\{u, v\}$ are generated from a similar causal mechanism.

- In other words, what we learn from one observed phenomenon can be used to explain and predict other related phenomena.
- For example, we can learn individuals' risk aversion from their investment behavior, which in turn, can help explain and predict their career choices.

Why Structural Estimation

S-3.2. Structural models make it possible to predict the effects of existing treatments in a new population/environment, or the effects of completely new treatments.

- To do so, a structural model must be "deep" enough so that its parameters remain invariant in the new population/environment, or when new treatments are applied.
- The concept of invariance is closely related to the concept of stability for causal relationships. The need for invariant parameters is key to causal analysis and policy evaluation.

S-4. Once we have learned a structural model, we can use it to generate synthetic data and perform counterfactual simulations.

Three Types of Program Evaluation Problems

P-1 Evaluating the impacts of historical programs on outcomes.

P-2 Forecasting the impacts of programs implemented in one population/environment in other populations/environments.

P-3 Forecasting the impacts of programs never historically experienced.

- P-1 is the problem of internal validity.
- P-2 is the problem of external validity.
- P-2 and P-3 often require the use of structural models.
- For all three types of problems, if we want to evaluate welfare impact, we need a structural model.

Why Structural Estimation

S-5. Structural models, by linking economic theories based on individual preferences to data, allow the economist to make welfare calculations and normative statements.

- Individual choices reveal information about their preferences and the potential outcomes they face[71].

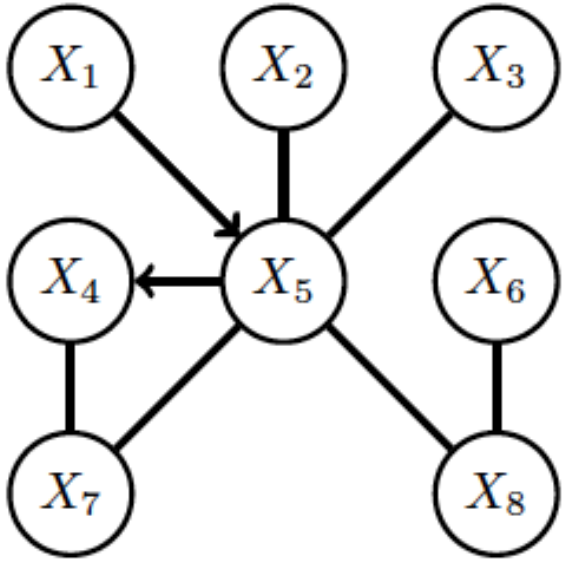
[71] Heckman and Vytlacil (2007): "Incorporating choice into the analysis of treatment effects is an essential and distinctive ingredient of the econometric approach .. An assignment in [the RCM] is an assignment to treatment, not an assignment of *incentives* and *eligibility* for treatment with the agent making treatment choices. [The statistical treatment effect approach] has only one assignment mechanism and treats noncompliance with it as a problem rather than as a source of information on agent preferences, as in the econometric approach ... Accounting for uncertainty and subjective valuations of outcomes ... is a major contribution of the econometric approach."

Why Structural Estimation

S-6. Like all scientific models, structural models can potentially deliver better predictive performance than statistical models trained on single data sets, because their parameters can be learned from a combination of data from various sources that share the same underlying causal mechanism[72].

[72] This point is related to S-3.1 and S-3.2: different observed phenomena can all help inform the values of the same set of "deep" parameters.

Appendix I: Graphs

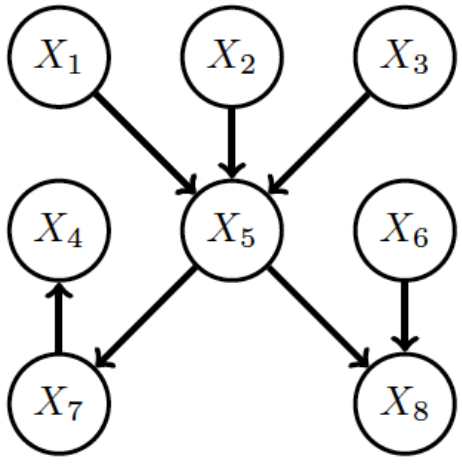


A **graph** consists of nodes (**vertices**) and undirected or directed links (**edges**) between nodes.

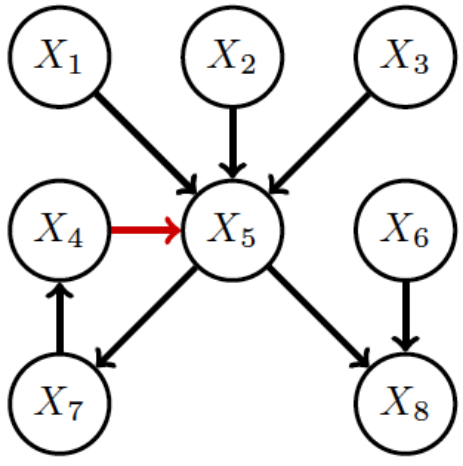
A **path** from X_i to X_j is a sequence of connected nodes starting at X_i and ending at X_j.

Appendix I: Directed Graphs

Directed Acyclic Graph



Directed Cyclic Graph

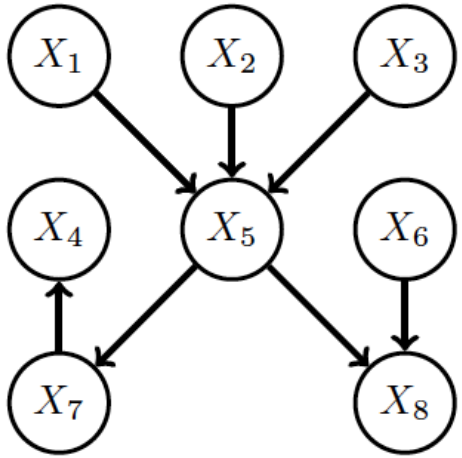


Directed Graphs are graphs in which all the edges are directed.

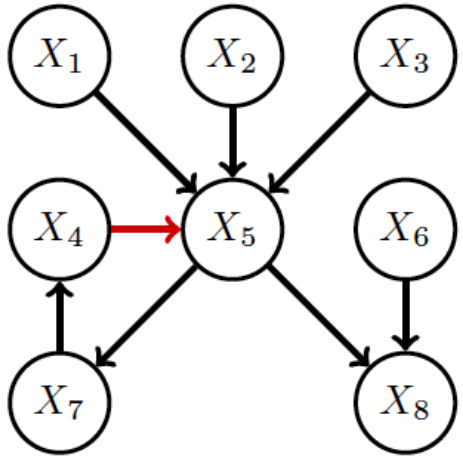
Directed Acyclic Graph (**DAG**): Graph in which by following the direction of the arrows a node will never be visited more than once.

Appendix I: Directed Graphs

Directed Acyclic Graph



Directed Cyclic Graph



If there is a link from X_i to X_j, then X_i is called a **parent** of X_j and X_j is a **child** of X_i.

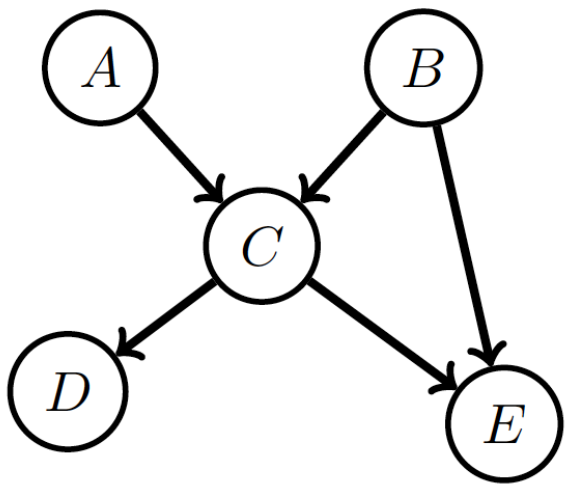
The **ancestors** of a node X_i are the nodes with a directed path ending at X_i. The **descendants** of X_i are the nodes with a directed path beginning at X_i.

Appendix I: Bayesian Networks

A **Bayesian network** (**BN**) is a DAG in which each node has associated the conditional probability of the node given its parents.

The joint distribution is obtained by taking the product of the conditional probabilities:

$$p(A, B, C, D, E) = P(A)\, P(B)\, P(C|A, B)\, P(D|C)\, P(E|B, C)$$



$p(E|B, C)$

Appendix I: Bayesian Networks

Formally, a Bayesian network[73] is a distribution of the form:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | \text{pa}(x_i))$$

, where $\text{pa}(x_i)$ denotes the parental variables of x_i. The BN is represented as a DAG with the i^{th} node in the graph corresponding to the factor $p(x_i | \text{pa}(x_i))$.

A BN encodes the following set of conditional independence assumptions, called the **local independencies**:

$$x_i \amalg \text{nd}(x_i) | \text{pa}(x_i)$$

, where $\text{nd}(x_i)$ denotes variables that are not descendants of x_i.

[73] Also called **belief network**.

Appendix I: Bayesian Networks

- The possible ways variables can interact is extremely large. Given N binary variables $\{x_1, \ldots, x_N\}$, specifying all the joint probabilities $p(x_1, \ldots, x_N)$ would take $\mathcal{O}\left(2^N\right)$ space – impractical for large N.
- To render specification and inference tractable, we need to constrain the nature of the variable interactions. The key idea is to specify which variables are independent of others, leading to a structured factorization of the joint probability distribution.
- Bayesian networks provide a convenient framework for representing such independence assumptions via graphs.

Appendix I: Bayesian Networks

Sally's burglar Alarm (A) is sounding. Has she been Burgled (B), or was the alarm triggered by an Earthquake (E)? She turns the car Radio (R) on for news of earthquakes.

Without loss of generality, we can write

$$p(A, R, E, B) = p(A|R, E, B)\, p(R|E, B)\, p(E|B)\, p(B)$$

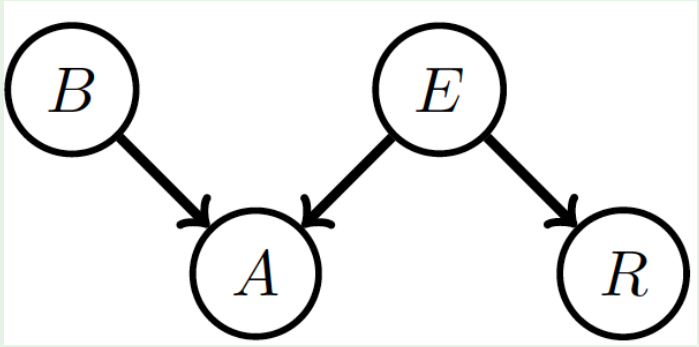
Independence Assumptions:

- $p(A|R, E, B) = p(A|E, B)$
- $p(R|E, B) = p(R|E)$
- $p(E|B) = P(E)$

Given these assumptions:

$$p(A, R, E, B) = p(A|E, B)\, p(R|E)\, p(E)\, p(B)$$

Appendix I: Bayesian Networks



Specifying the tables:

$p(B = 1) = .01$

$p(E = 1) = .000001$

$p(A|B, E)$

Alarm = 1	Burglar	Earthquake
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

$p(R|E)$

Radio = 1	Earthquake
1	1
0	0

Appendix I: Bayesian Networks

Initial Evidence: The alarm is sounding

$$
\begin{aligned}
p(B=1|A=1) &= \frac{\sum_{E,R} p(B=1,E,A=1,R)}{\sum_{B,E,R} p(B,E,A=1,R)} \\
&= \frac{\sum_{E,R} p(A=1|B=1,E)\, p(B=1)\, p(E)\, p(R|E)}{\sum_{B,E,R} p(A=1|B,E)\, p(B)\, p(E)\, p(R|E)} \\
&\approx 0.99
\end{aligned}
$$

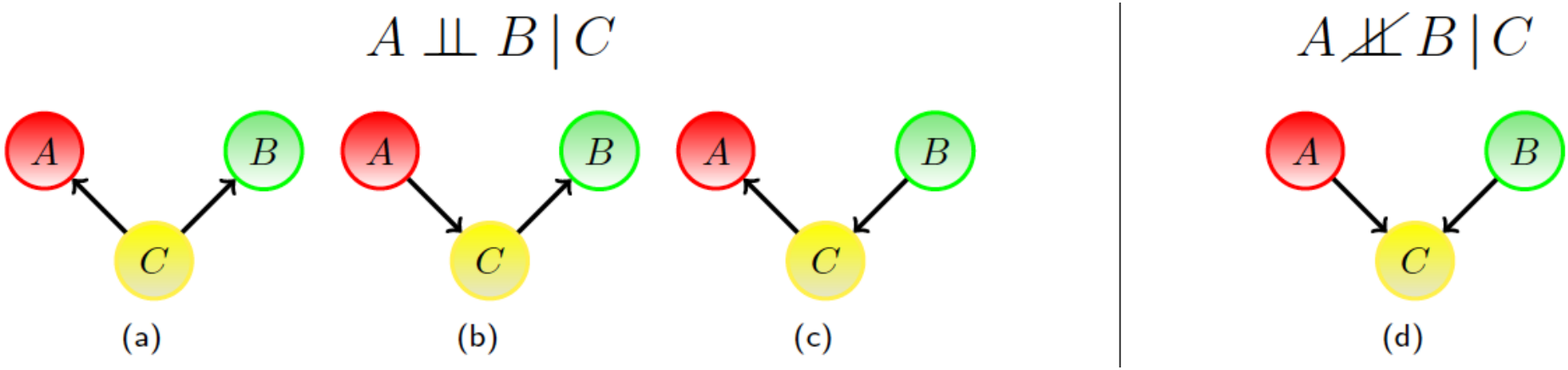
Additional Evidence: The radio broadcasts an earthquake warning

A similar calculation gives $p(B=1|A=1) \approx 0.01$

Initially, because the alarm sounds, Sally thinks that she's been burgled. However, this probability drops dramatically when she hears there's been an earthquake. The earthquake "explains away" the alarm.

Appendix I: Independence in Bayesian Networks

All BNs with three nodes and two links:



In (a), (b) and (c), A, B are conditionally independent given C.

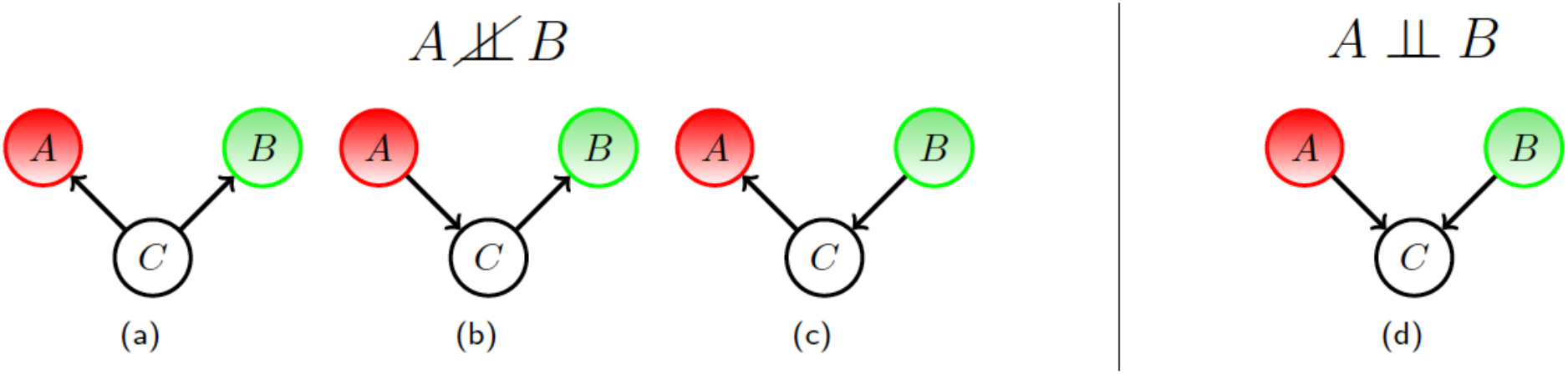
(a) $p(A, B|C) = \frac{p(A,B,C)}{p(C)} = \frac{p(A|C)p(B|C)p(C)}{o(C)} = p(A|C)\,p(B|C)$

(b) $p(A, B|C) = \frac{p(A,C)p(B|A,C)}{p(C)} = \frac{p(A,C)p(B|C)}{o(C)} = p(A|C)\,p(B|C)$

(c) is equivalent to (b). In (d), A, B are conditionally dependent given C: $p(A, B|C) \propto p(C|A, B)\,p(A)\,p(B)$.

Appendix I: Independence in Bayesian Networks

All BNs with three nodes and two links:



In (a), (b) and (c), A, B are marginally dependent.

In (d), A, B are marginally independent:

$$p(A,B) = \sum_C p(A,B,C) = \sum_C p(C|A,B)\, p(A)\, p(B) = p(A)\, p(B)$$

Appendix I: Collider

- Given a path $\mathcal{P}$, a **collider** is a node c on $\mathcal{P}$ with neighbors a and b on $\mathcal{P}$ such that $a \rightarrow c \leftarrow b$[74].
- Given a set of nodes $\mathcal{C}$, a path $\mathcal{P}$ is **blocked** by $\mathcal{C}$ if at least one of the following conditions is satisfied:
  1. $\exists$ a non-collider on $\mathcal{P}$ that $\in \mathcal{C}$.
  2. $\exists$ a collider on $\mathcal{P}$ such that neither the collider nor any of its descendants $\in \mathcal{C}$.

[74] Note that a collider is *path specific*.

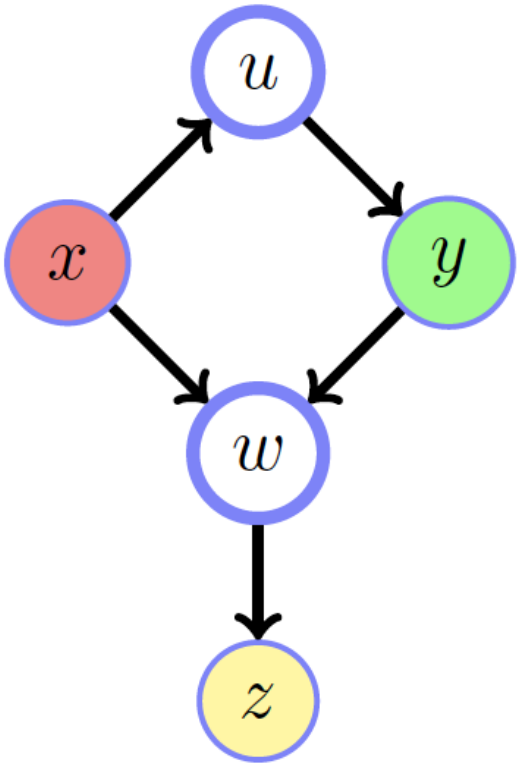
Appendix I: D-Separation

- Given three disjoint sets of nodes $\{\mathcal{X}, \mathcal{Y}, \mathcal{C}\}$, $\mathcal{X}$ and $\mathcal{Y}$ are said to be **d-separated** by $\mathcal{C}$ if all paths from any element of $\mathcal{X}$ to any element of $\mathcal{Y}$ are blocked by $\mathcal{C}$. Otherwise they are **d-connected** by $\mathcal{C}$.

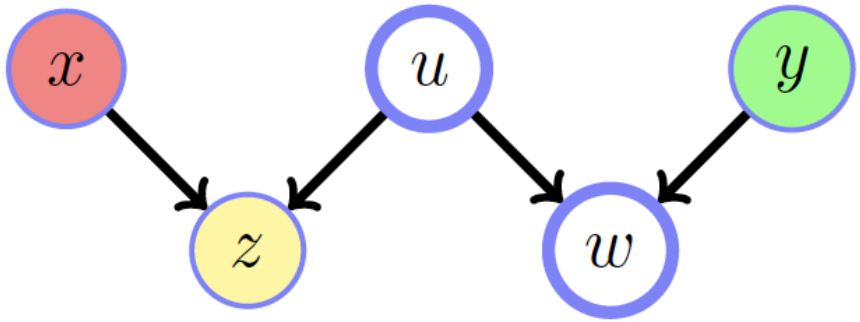
- $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{C} \Rightarrow \mathcal{X} \amalg \mathcal{Y} | \mathcal{C}$[75].

[75] D-connection, however, does not necessarily imply conditional dependence. More precisely, if $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{C}$, then $\mathcal{X} \amalg \mathcal{Y} | \mathcal{C}$ in *every* distribution compatible with the causal diagram. If $\mathcal{X}$ and $\mathcal{Y}$ are d-connected by $\mathcal{C}$, then $\mathcal{X} \not\amalg \mathcal{Y} | \mathcal{C}$ in *at least one* distribution compatible with the causal diagram.

Appendix I: Independence in Bayesian Networks



$x \not\perp\!\!\!\perp y$, $x \perp\!\!\!\perp y \mid u$, $x \not\perp\!\!\!\perp y \mid w$, $x \not\perp\!\!\!\perp y \mid z$,
$x \not\perp\!\!\!\perp y \mid \{u, w\}$, $x \not\perp\!\!\!\perp y \mid \{u, z\}$,
$x \not\perp\!\!\!\perp y \mid \{w, z\}$, $x \not\perp\!\!\!\perp y \mid \{u, w, z\}$



$x \perp\!\!\!\perp y$, $x \perp\!\!\!\perp y \mid u$, $x \perp\!\!\!\perp y \mid w$, $x \perp\!\!\!\perp y \mid z$,
$x \perp\!\!\!\perp y \mid \{u, w\}$, $x \perp\!\!\!\perp y \mid \{u, z\}$,
$x \not\perp\!\!\!\perp y \mid \{w, z\}$, $x \perp\!\!\!\perp y \mid \{u, w, z\}$

Appendix I: Observational Equivalence

Skeleton

Formed from a directed graph by removing the arrows

Immorality

An immorality in a DAG is a configuration of three nodes, A, B, C such that C is a child of both A and B, with A and B not directly connected[76].

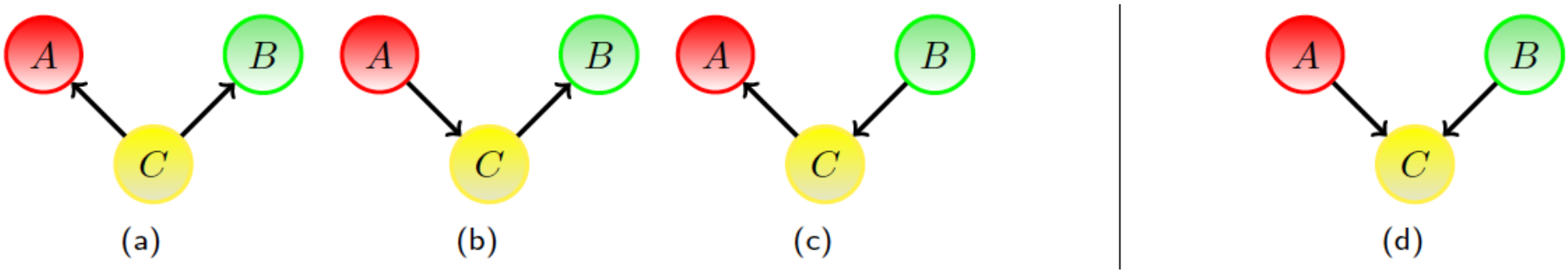
Observational Equivalence (Markov Equivalence)

Two DAGs are observationally equivalent, in the sense that they represent the same set of independence assumptions, if and only if they have the same skeleton and the same set of immoralities.

---

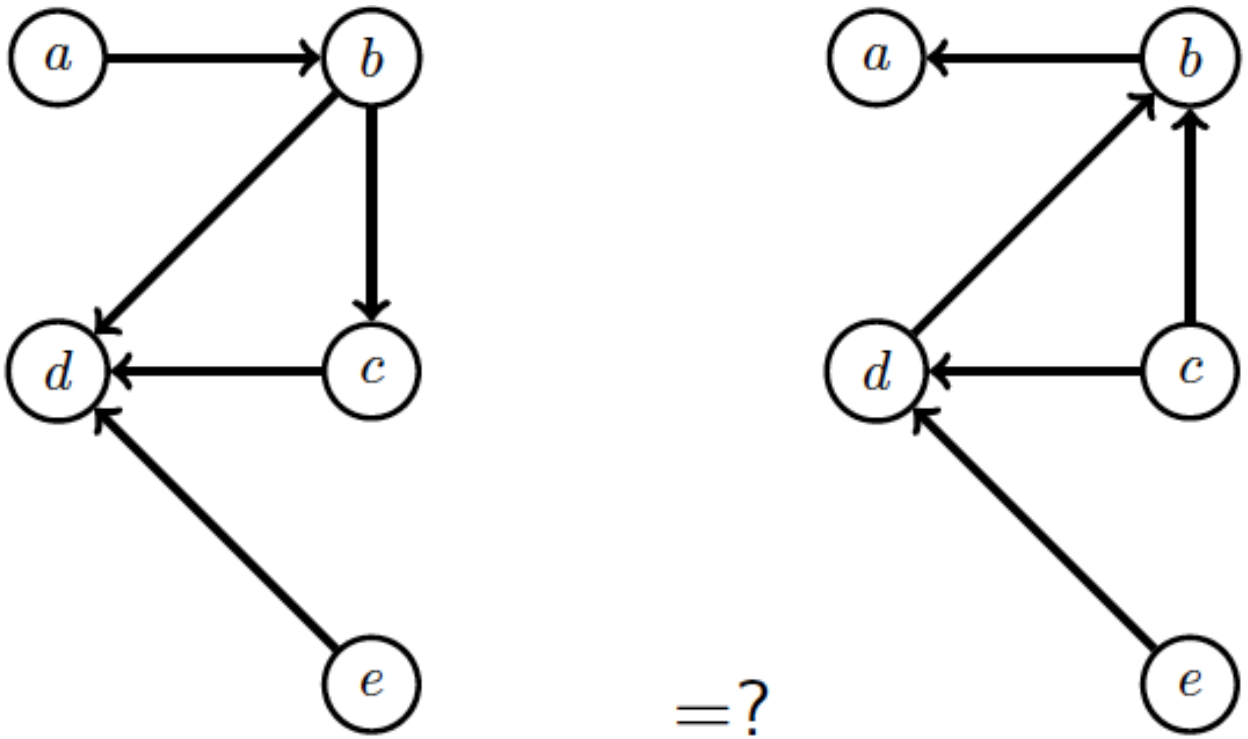
[76] i.e., not "married."

Appendix I: Observational Equivalence



(a), (b) and (c) are observationally equivalent, while (d) is observationally different from (a) - (c).

Appendix I: Observational Equivalence



Acknowledgement I

Part of this lecture is adapted from the following sources:

- Antoine de Saint-Exupéry. 2001. *Le Petit Prince*. Harcourt, Inc. (Original work published 1943.)
- Barber, D. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.
- Blackwell, M. 2015. *Causal Inference*. Lecture at Harvard University, retrieved on 2017.01.01. [link]
- Hernán, M. A. and J. M. Robins. 2018. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.
- James, G., D. Witten, T. Hastie, R. Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.

Acknowledgement II

- Morgan, S. L. and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.
- Pearl, J. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.
- Pearl, J., M. Glymour, and N. P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley.
- Silva, R. *Causal Inference in Machine Learning*. Talk at Imperial College London, retrieved on 2017.01.01. [link]
- Varian, H. R., *Machine Learning and Econometrics.* Talk at Google, retrieved on 2017.01.01. [link]

Reference I

- Angrist, J. D. and J. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Bjerkholt, O. 1998. "Ragnar Frisch and the Foundation of the Econometric Society," In Steinar Strøm (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, Cambridge University Press.
- Dawid, A. P. 2000. "Causal inference without counterfactuals," *Journal of the American Statistical Association*, 95(450).
- Deaton, A. and N. Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, 210.
- Frisch, R. 1926. "Sur un problème d'économie pure," *Norsk Matematisk Forenings Skrifter*, Series I, No. 16.
- Guerre, E., I. Perrigne, and Q. Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica,* 68(3).

Reference II

- Haavelmo, T. 1943. "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11(1).
- Haavelmo, T. 1944. "The Probability Approach in Econometrics," *Econometrica*, 12(Supplement).
- Heckman, J. J. and E. J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6, Elsevier.
- Holland, P. W. 1986. "Statistics and causal inference," *Journal of the American Statistical Association*, 81(396).
- Hood, W. C. and T. C. Koopmans (eds.) 1953. *Studies in Econometric Method*, Cowles Commission Monograph 14, Wiley.
- Jensen, R. T. and N. H. Miller. 2008. "Giffen Behavior and Subsistence Consumption," *American Economic Review*, 98(4).

Reference III

- Koopmans, T. C. (ed.) 1950. *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, Wiley.
- Krueger, Alan. 1999. "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114.
- Neyman, J. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science*, 5(4), (Nov., 1990).
- Rubin, D. B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 56.
- Rubin, D. B. 1978. "Bayesian inference for causal effects: The role of randomization," *Annals of Statistics*, 6.
- Russell, B., 1912. *The Problems of Philosophy*. Arc Manor, Rockville, MD (2008).
- Scott, P. T. 2013. "Dynamic Discrete Choice Estimation of Agricultural Land Use," Working Paper.