Lecture Notes on Causal Inference

(with corrections)

Qingyuan Zhao

May 30, 2022

Website for this course: http://www.statslab.cam.ac.uk/~qz280/teaching/causal-2020/.

Please contact me if you find any mistakes or have any comments.

Contents

1 **What is causal inference?** **1**
- 1.1 Some motivating examples . . . 1
- 1.2 Languages for causality . . . 7
- 1.3 Concepts and principles . . . 8

2 **Randomised experiments** **12**
- 2.1 Assignment mechanism . . . 12
- 2.2 Potential outcomes/Counterfactuals . . . 13
- 2.3 Randomisation distribution of causal effect estimator . . . 17
- 2.4 Randomisation test of sharp null hypothesis . . . 18
- 2.5 Super-population inference and regression adjustment . . . 23
- 2.6 Comparison of different modes of inference . . . 26

3 **Path analysis** **28**
- 3.1 Graph terminology . . . 28
- 3.2 Linear structural equation models . . . 29
- 3.3 Path analysis . . . 31
- 3.4 Correlation and causation . . . 32
- 3.5 Latent variables and identifiability . . . 34
- 3.6 Factor models and measurement models . . . 35
- 3.7 Estimation in linear SEMs . . . 38
- 3.8 Strengths and weaknesses of linear SEMs . . . 39

4 **Graphical models** **41**
- 4.1 Markov properties for undirected graphs . . . 41
- 4.2 Markov properties for directed graphs . . . 42
- 4.3 Structure discovery . . . 48
- 4.4 Discussion: Using DAGs to represent causality . . . 50
- 4.A Graphical proofs . . . 51

5 **A unifying theory of causality** **54**
- 5.1 From graphs to structural equations to counterfactuals . . . 54

5.2 Markov properties for counterfactuals 56
5.3 From counterfactual to factual 60
5.4 Causal identification 63
5.5 Proofs (non-examinable) 67

6 No unmeasured confounders: Randomisation inference 69
6.1 The logic of observational studies 69
6.2 No unmeasured confounders 71
6.3 Matching algorithms 71
6.4 Covariate balance 73
6.5 Randomisation inference 74

7 No unmeasured confounders: Semiparametric inference 77
7.1 An introduction to semiparametric inference 77
7.2 Discrete covariates 79
7.3 Outcome regression and inverse probability weighting 81
7.4 Doubly robust estimator 82
7.5 A comparison of the statistical methods 85

8 Sensitivity analysis 87
8.1 A roadmap 87
8.2 Rosenbaum's sensitivity analysis 87
8.3 Sensitivity analysis in semiparametric inference 89

9 Unmeasured confounders: Leveraging specificity 91
9.1 Structural specificity 91
9.2 Instrumental variables and two-stage least squares 92
9.3 Instrumental variables: Method of moments 94
9.4 Complier average treatment effect 96

10 Mediation analysis 100
10.1 Linear SEM 100
10.2 Identification of controlled direct effect 101
10.3 Natural direct and indirect effects 102
10.4 Observed confounders 103
10.5 Extended graph interpretation 106

Chapter 1

What is causal inference?

$$\text{Causal inference} \approx \text{Causal language/model} + \text{Statistical inference}.$$

1.1 Some motivating examples

Example 1: Smoking and lung cancer

By the mid-1940s, it had been observed that lung cancer cases had tripled over the previous three decades. But the cause for the increase in lung cancer was unclear and not agreed upon. Possible explanations included

- Changes in air quality due to the introduction of automobile;
- Widespread expansion of paved roads that contained many carcinogens;
- Aging of the population;
- The advent of radiography;
- Better clinical awareness of lung cancer and better diagnostic methods;
- Smoking.

Advertisement for cigarette smoking (1950):

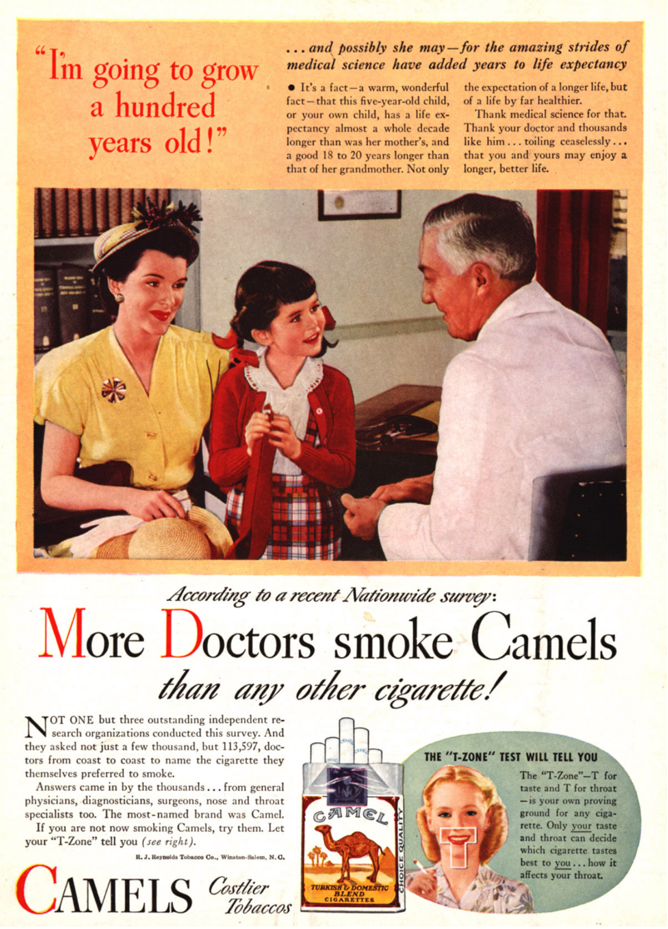
A series of observational studies reported overwhelming association between smoking and lung cancer[1]. Some data[2]: 36,975 pairs of heavy smoker and nonsmoker, matched by age, race, nativity, rural versus urban residence, occupational exposures to dust and fumes, religion, education, marital status, ...

Of the 36,975 pairs, there were 122 discordant pairs in which exactly one person died of lung cancer. Among them,

- 12 pairs in which nonsmoker died of lung cancer;
- 110 pairs in which smoker died of lung cancer.

So smoking is very strongly associated with lung cancer.

Fisher strongly objected the idea that smoking is carcinogenic:[3]



"I'm going to grow a hundred years old!"

. . . and possibly she may—for the amazing strides of medical science have added years to life expectancy

● It's a fact—a warm, wonderful fact—that this five-year-old child, or your own child, has a life expectancy almost a whole decade longer than was her mother's, and a good 18 to 20 years longer than that of her grandmother. Not only the expectation of a longer life, but of a life by far healthier.

Thank medical science for that. Thank your doctor and thousands like him . . . toiling ceaselessly . . . that you and yours may enjoy a longer, better life.

According to a recent Nationwide survey:

More Doctors smoke Camels

than any other cigarette!

NOT ONE but three outstanding independent research organizations conducted this survey. And they asked not just a few thousand, but 113,597, doctors from coast to coast to name the cigarette they themselves preferred to smoke.

Answers came in by the thousands . . . from general physicians, diagnosticians, surgeons, nose and throat specialists too. The most-named brand was Camel.

If you are not now smoking Camels, try them. Let your "T-Zone" tell you *(see right).*

R. J. Reynolds Tobacco Co., Winston-Salem, N. C.

CAMELS *Costlier Tobaccos*

THE "T-ZONE" TEST WILL TELL YOU

The "T-Zone"—T for taste and T for throat—is your own proving ground for any cigarette. Only your taste and throat can decide which cigarette tastes best to you . . . how it affects your throat.

> Such results suggest that an error has been made of an old kind, in arguing from correlation to causation.... Such differences in genetic make-up between those classes would naturally be associated with differences of disease incidence without the disease being causally connected with smoking.

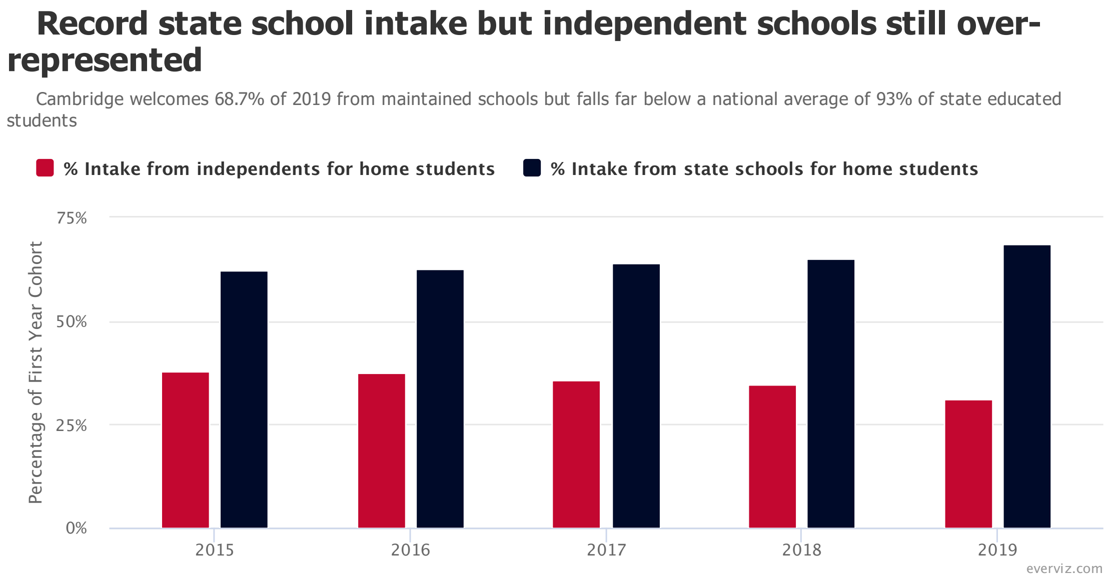
Fisher then demonstrated evidence of a gene that is associated with both smoking and lung cancer.

We now know Fisher was wrong. His criticism was logical, but the association between smoking and lung cancer is simply too strong to be explained away by different genetic make-ups[4]. Some believe that his views may have been influenced by personal and professional conflicts, by his work as a consultant to the tobacco industry, and by the fact that he was himself a smoker.

Example 2: Undergraduate admissions

Some great visualisations of Cambridge's 2019–2020 admission data[5]:

Chart 1 State vs. independent schools.

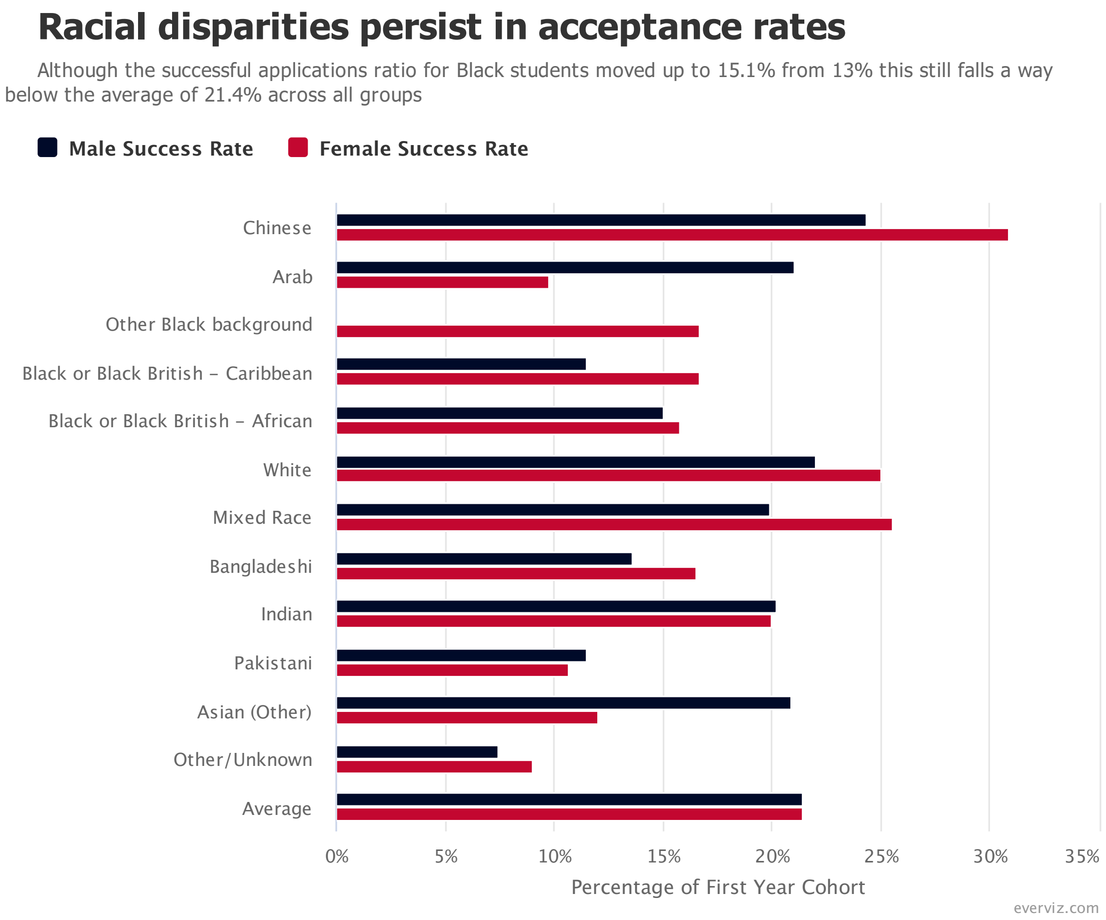


Some interesting quotes: “Considering 93% of pupils in England are taught in state schools, a figure of 68.7% means that state school students are still vastly under-represented in the University.... Cambridge’s acceptance of state school applicants continues to be amongst the lowest in the UK, with 90% of university students on average hailing from state schools across the country. ”

Does this mean Cambridge’s admission is biased against state schools? Not necessarily. For example, applicants from independent schools may have better A-level results.

Causal inference can be used to understand fairness in decisions made by human and computer algorithms[6].

Chart 3 Racial disparities.

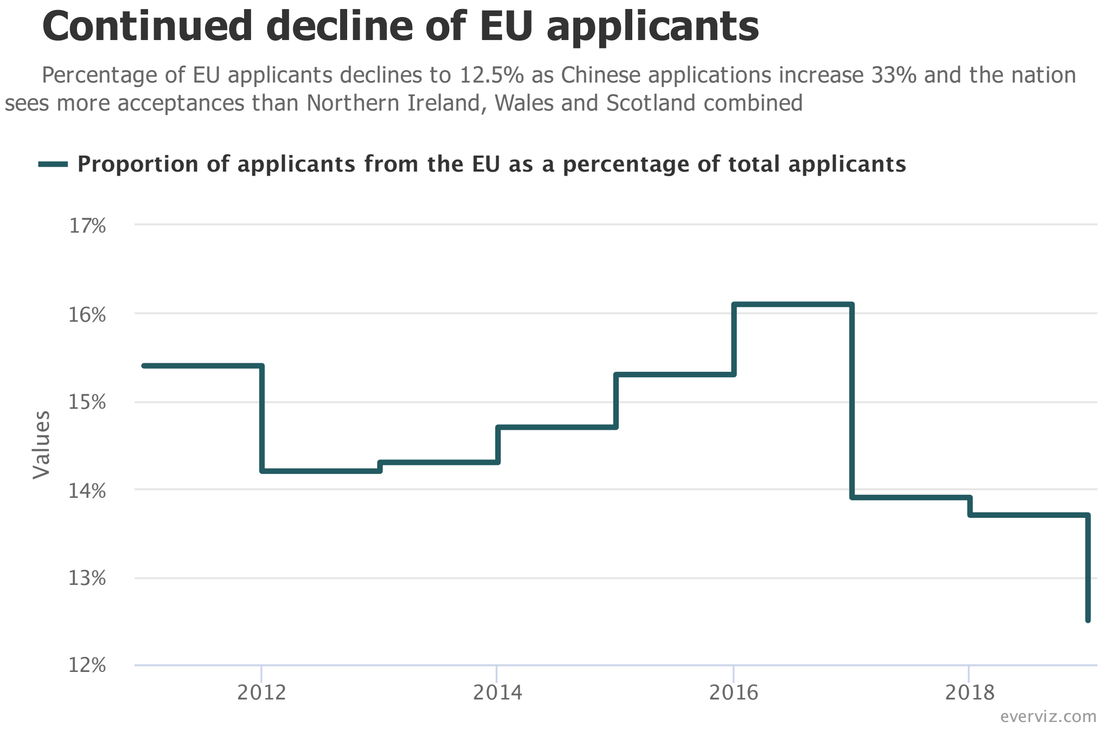


Racial disparities persist in acceptance rates

Although the successful applications ratio for Black students moved up to 15.1% from 13% this still falls a way below the average of 21.4% across all groups

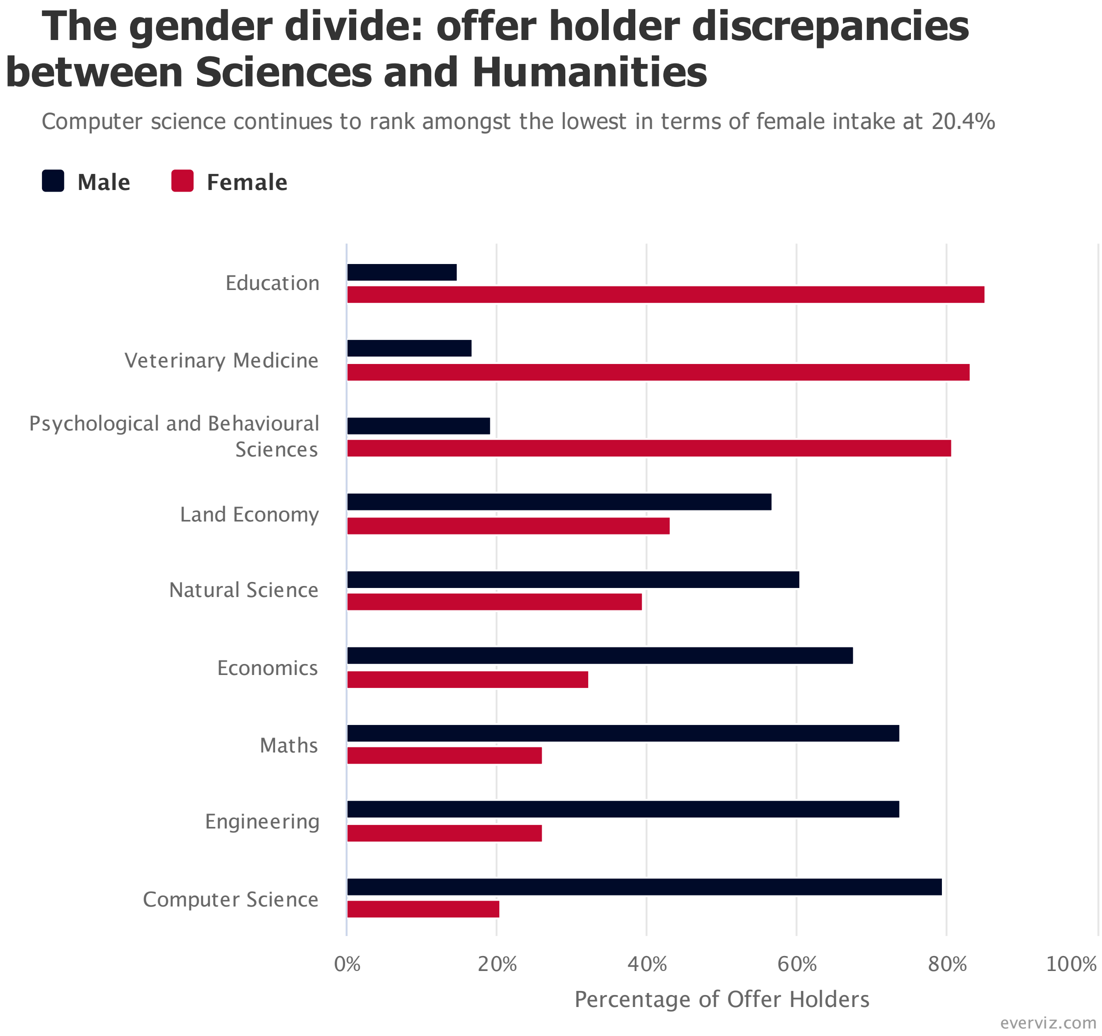
Again this is showing associations. But in general it is not straightforward to discuss the causal effect of race, because it is hard to conceptualise "manipulation" of race at birth. One possibility is to consider "perceived race" instead.

Chart 4 Impact of Brexit.

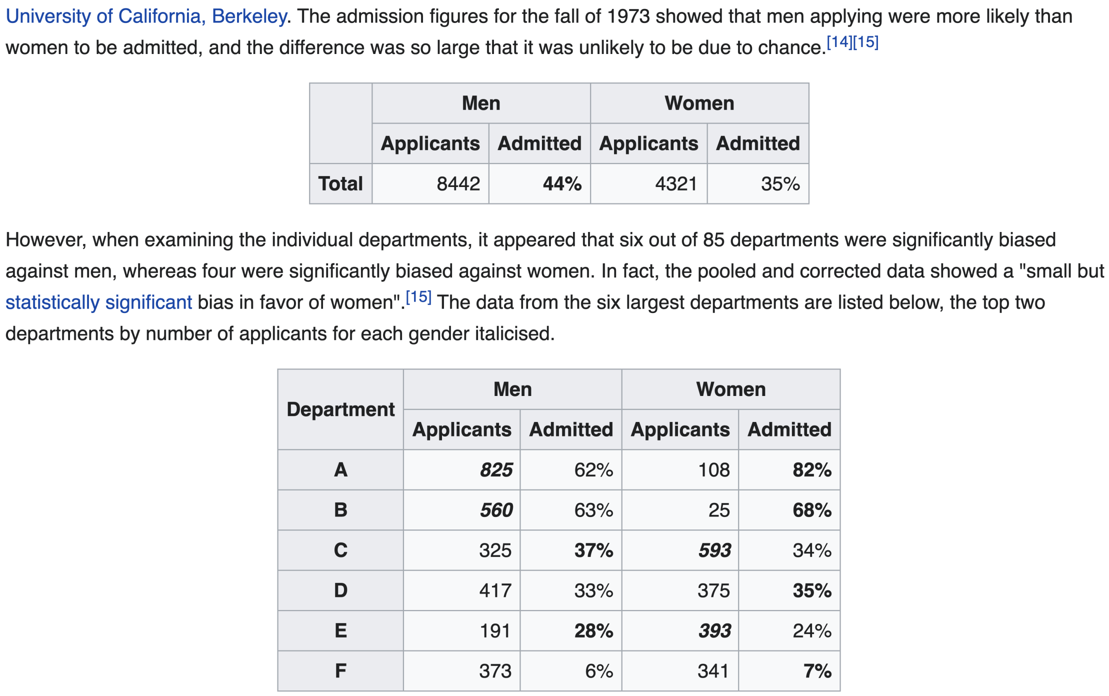


Is the steep decline in EU applications caused by Brexit (or Brexit dubiety)? It is possible to answer this question by using a concept in causal inference called probability of causation[7], which is quite useful in law[8].

Chart 5 Gender difference.



Related: Simpson's (or Yule-Simpson) paradox. UC Berkeley 1973 admission data[9]:



University of California, Berkeley. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance.[14][15]

	Men		**Women**	
	Applicants	**Admitted**	**Applicants**	**Admitted**
Total	8442	**44%**	4321	35%

However, when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas four were significantly biased against women. In fact, the pooled and corrected data showed a "small but statistically significant bias in favor of women".[15] The data from the six largest departments are listed below, the top two departments by number of applicants for each gender italicised.

Department	**Men**		**Women**	
	Applicants	**Admitted**	**Applicants**	**Admitted**
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

This paradox is first discovered by Pearson (1899), who offered a causal explanation: "To those who persist on looking upon all correlation as cause and effect, the fact that correlation can be produced between two quite uncorrelated characters A and B by taking an artificial mixture of the two closely allied races, must come as rather a shock."[10]

1.2 Languages for causality

(0) **"Implicit" in randomisation**:

1925 Fisher (statistics, genetics, agricultural experiments).

(i) **Using potential outcomes/counterfactuals**:

1923 Neyman (statistics);

1973 Lewis (philosophy);

1974 Rubin (statistics);

1986 Robins (epidemiology);

(ii) **Using structural equations**:

1921 Wright (genetics);

1943 Haavelmo (econometrics);

1975 Duncan (social sciences);

2000 Pearl (computer science).

(iii) **Using graphs**:

1921 Wright (genetics);

1988 Pearl (computer science "AI");

1993 Spirtes, Glymour, Scheines (philosophy).

Some remarks:

- The multi-disciplinary origin and development.
- Theorisation driven by demand from applications.
- Applications are often related to humans (biology, public health, economics, political sciences...). Why? Open-system with external interactions, difficult or nearly impossible for manipulation in experiments.
- State-of-the-art: The three languages are basically equivalent and advantageous for different purposes.
  - Graphs: Easy to visualise the causal assumptions; Difficult for statistical inference because model is nonparametric.
  - Structural equations: Bridge between graphs and counterfactuals; Easy to operationalise; Danger to be confused with regressions.
  - Counterfactuals: Easy to incorporate additional assumptions; Elucidation of the meaning of statistical inference; Not as convenient if system is complex.

1.3 Concepts and principles

(i) **Observation vs. intervention.**

- Non-experimental vs. experimental data.
- Seeing vs. doing (J. Pearl).

(ii) **(Controlled) Randomised experiment is the gold standard of causal inference.**

(iii) **Different types of inference** (C. S. Peirce, late 1800s):

Deduction Necessary inference following logic.

- Example: Euclidean geometry.

Induction Probable or non-necessary inference (purely) based on statistical data.

- Example: Survey sampling; Correlation between cigarette smoking and lung cancer.

Abduction Inference with implicit or explicit appeal to explanatory considerations.

    - Example: Investigation of aircraft crash; Cigarette smoking causes lung cancer.
- Question: What type of inference is mathematical induction?
- The boundary between induction and abduction is not always clear.
- Very very roughly speaking, deduction ≈ mathematics; induction ≈ statistics; abduction ≈ causal inference.

(iv) **Causal identification.**

- Identifiability of statistical models: $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Longrightarrow \theta_1 = \theta_2,\ \forall \theta_1, \theta_2$.
- Causal identifiability: think about θ as counterfactuals or unobserved variables and $\mathbb{P}_\theta$ as the distribution or model of the factuals or observed data.
- Causal identification with non-experimental data always requires **non-verifiable assumptions**.
    - Example: No unmeasured confounding; Instrumental variable is exogenous.

(v) **Design trumps analysis** (D. Rubin).

- Design: Choose an identification strategy and collect relevant data.
- Analysis: Apply an appropriate statistical method to analyse the data.
- Success of a research study: 99% (maybe exaggerating...) depends on the design and how data are collected.
- Unfortunately, we usually learn much more about analysis in statistics courses.

(vi) **All models are wrong, but some are useful** (G. Box).

- Historically, causality is often defined as certain parameters in a statistical model (e.g., a linear model) not equal to 0.
- A strong emphasis of modern causal inference is on robustness to the statistical models, this includes
    - Nonparametric identification of causal effect that is free of modelling assumptions.
    - Nonparametric/semiparametric inference for causal effect.

(vii) **Causal mechanism and causal mediation.**

- Asking why and how.
- Example: How does (or which substance in) cigarettes cause lung cancer?
- Often more important than “is”.

(viii) **Specificity.**

- One of Hill's 9 criteria for causality[11]: "If as here, the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favor of causation."
- Original definition now considered weak or obsolete. Counterexample: smoking.
- In Hill's era, exposure = an occupational setting or a residential location (proxies for true exposures).
- Nowadays, exposure is much more precise (for example, a specific gene expression).
- Specificity is still useful. Examples: Instrumental variables, negative controls, sparsity.

(ix) **Corroboration of evidence.**

- Famous quote from Fisher[12]:

  > About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: **"Make your theories elaborate."** The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold... this multi-phasic attack is one of the most potent weapons in observational studies.

- Falsifiability of scientific theory (K. Popper, 1959).
- Evolution of scientific paradigms—the social and collaborative nature of scientific progress (T. Kuhn, 1962).

Notes

[1] Doll, R., & Hill, A. B. (1950). Smoking and carcinoma of the lung. *BMJ*, *2*(4682), 739–748. doi:10.1136/bmj.2.4682.739.

[2] Hammond, E. C. (1964). Smoking in relation to mortality and morbidity. findings in first thirty-four months of follow-up in a prospective study started in 1959. *Journal of the National Cancer Institute*, *32*(5), 1161–1188.

[3] Fisher, R. A. (1958). Cancer and smoking. *Nature*, *182*(4635), 596–596. doi:10.1038/182596a0.

[4] This is pointed out by Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, *22*(1), 173–203, which is widely regarded as the first sensitivity analysis to observational studies.

[5] Vides, G., & Powell, J. (2020, June 16). The eight charts that explain the university's 2019-2020 undergraduate admissions data. *Varsity.*

[6] See e.g. Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, *578*(7793), 34–36. doi:10.1038/d41586-020-00274-3.

[7] Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, *121*, 93–149. doi:10.1023/a:1005233831499.

[8] Dawid, A. P., Musio, M., & Murtas, R. (2017). The probability of causation. *Law, Probability and Risk*, *16*(4), 163–179. doi:10.1093/lpr/mgx012.

[9] https://en.wikipedia.org/wiki/Simpson's_paradox#UC_Berkeley_gender_bias

[10] See Sec. 6.1 of Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

[11] Hill, A. B. (2015). The environment and disease: Association or causation? *Journal of the Royal Society of Medicine*, *108*(1), 32–37. doi:10.1177/0141076814562718.

[12] Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, *128*(2), 234–266.

Chapter 2

Randomised experiments

- Randomised experiment (or randomised controlled trial) is the gold standard for establishing causality.
- This Chapter: basic concepts and techniques in designing and analysing a randomised experiment.
- We will focus on the case of a binary treatment.

2.1 Assignment mechanism

- Suppose there are n units in this experiment.
- For the i-th unit, observed some *covariates* $\boldsymbol{X}_i$ prior to treatment assignment.
- *Treatment* $A_i \in \mathcal{A} = \{0, 1\}$. Example: a new drug.
  - Convention: $A_i = 1$ is the *treatment* or *treated* and $A_i = 0$ is the *control.*
  - Some also refer to A_i as the *exposure.* Example: toxic chemicals in a factory.
- A_i is assigned according to a randomisation scheme (below).
- Notation: $[n] = \{1, 2, \ldots, n\}$; $\boldsymbol{A}_{[n]} = (A_1, A_2, \ldots, A_n)^T$; $\boldsymbol{X}_{[n]} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)^T$. All vectors are column vectors. We often suppress the subscript i to refer to a generic variable; for example, A refers to a generic A_i.

The *assignment mechanism* is the conditional distribution:

$$\mathbb{P}(\boldsymbol{A}_{[n]} = \boldsymbol{a}_{[n]} \mid \boldsymbol{X}_{[n]} = \boldsymbol{x}_{[n]}) = \pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{x}_{[n]}),$$

where the function $\pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{x}_{[n]})$ is prespecified.

Next: some commonly used assignment mechanisms. What is the corresponding $\pi(\cdot \mid \cdot)$?

2.1 Example (Bernoulli trial)**.** The treatment assignments are independent and the probability of being treated is a constant $0 < \pi < 1$. $\pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{x}_{[n]}) = \prod_{i=1}^{n} \pi^{a_i}(1-\pi)^{1-a_i}$.

2.2 Example (Sample without replacement)**.** The treatment assignments are "completely randomised" with the only restriction that the number of treated units is $0 < n_1 < n$.

$$\pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{x}_{[n]}) = \begin{cases} \binom{n}{n_1}^{-1}, & \text{if } \sum_{i=1}^{n} a_i = n_1, \\ 0, & \text{otherwise.} \end{cases}$$

2.3 Example (Bernoulli trial with covariates)**.** Bernoulli trial with π replaced by a function $0 < \pi(\boldsymbol{x}) < 1$. $\pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{x}_{[n]}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{a_i}\{1 - \pi(\boldsymbol{x}_i)\}^{1-a_i}$.

2.4 Example (Pairwise experiment)**.** Suppose n is even. The units are divided into $n/2$ pairs based on the covariates. Within each pair, one unit is randomly assigned to treatment.

Let $B_i = B_i(\boldsymbol{x}_{[n]})$ be the pair that unit i is assigned to. Then

$$\pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{x}_{[n]}) = \begin{cases} 2^{-n/2}, & \text{if } \sum_{i=1}^{n} a_i \cdot I(B_i = j) = 1, \ \forall j = 1, \dots, n/2, \\ 0, & \text{otherwise.} \end{cases}$$

2.5 Exercise (Stratified experiment)**.** Generalise the pairwise experiment to allow more than 2 units in each group. Suppose there are m groups and group j has n_j units and n_{1j} treated units. What is the assignment mechanism?

More complicated assignment mechanisms attempt to make the distributions of $\boldsymbol{X}$ in the treated and the control as close as possible. Examples: covariate adaptive randomisation[1]; re-randomisation[2].

2.2 Potential outcomes/Counterfactuals

After treatment assignment, we follow up the units and measure an outcome variable Y_i for unit i.

"Implicit" causal inference

To estimate the causal effect of A on Y, one may

- Compare the conditional expectations $\mathbb{E}[Y \mid A = 0]$ with $\mathbb{E}[Y \mid A = 1]$.
- Compare the conditional distributions $\mathbb{P}(Y \leq y \mid A = 0)$ with $\mathbb{P}(Y \leq y \mid A = 1)$.
- Further condition on $\boldsymbol{X}$ and compare $\mathbb{E}[Y \mid A = 0, \boldsymbol{X} = \boldsymbol{x}]$ with $\mathbb{E}[Y \mid A = 1, \boldsymbol{X} = \boldsymbol{x}]$ or the conditional distributions.

This approach allows us to apply familiar statistical methodologies, but it has several limitations:

(i) Causal inference is only implicit and informal, as it seems that any difference can only be reasonably attributed to the different treatment assignments.

(ii) Difficult to extend to non-iid treatment assignments.

(iii) Cannot distinguish *internal validity* from *external validity.*

Internal validity: Inference for the *finite population* consisting of the n units.

External validity: Inference for the *super-population* from which the n units are sampled from.

Potential outcomes

The potential outcome model avoids the above problems and provides a flexible basis for causal inference. It is first introduced by Neyman in his 1923 Master's thesis[3] to study randomised experiments and later brought to observational studies by Rubin[4].

This approach posits a *potential outcome* (or *counterfactual*), $Y_i(\boldsymbol{a}_{[n]})$, for unit i under treatment assignment $\boldsymbol{a}_{[n]}$. The potential outcomes (or counterfactuals) are linked to the observed outcome (or factuals) via the following assumption.

2.6 Assumption (Consistency)**.** $Y_i = Y_i(\boldsymbol{A}_{[n]})$ for all $i \in [n]$.

This should not to be confused with *consistency of statistical estimators*, which says the estimator converges to its targeting parameter as sample size grows.

Question: How many potential outcomes are there in an experiment with n units? $|\mathcal{A}^n| = 2^n$.

To reduce the unknowns in the problem, a common assumption is

2.7 Assumption (No interference)**.** $Y_i(\boldsymbol{a}_{[n]}) = Y_i(a_i)$ for all $i \in [n]$ and $\boldsymbol{a}_{[n]} \in \mathcal{A}^n$.

Note the abuse of notation.

Due to historical reasons, people often use the jargon SUTVA (stable unit treatment value assumption) to refer to Assumptions 2.6 and 2.7.

Unless noted otherwise, we will maintain these assumptions in this Chapter. Because A_i is binary, we are only dealing with two potential outcomes, $Y_i(0)$ and $Y_i(1)$, for each unit i.

Fundamental problem of causal inference

The potential outcome framework allows as to view causal inference as a *missing data* problem.

Consider a hypothetical experiment in Table 2.1. Using the observed (A_i, Y_i) and the consistency assumption, we can impute one of potential outcomes $Y_i(A_i)$. However, the other potential outcome $Y_i(1 - A_i)$ is unknown.

The difference $Y_i(1) - Y_i(0)$ is called the *individual treatment effect* for unit i, which can never be observed. This is often referred to as the "fundamental problem of causal inference"[5].

However, it should be possible to estimate the treatment effect *at the population level* if the treatment is randomised. Two populations can be considered:

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	?	-3.7	1	-3.7
2	2.3	?	0	2.3
3	?	7.4	1	7.4
4	0.8	?	0	0.8
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 2.1: Illustration of potential outcome and the consistency assumption.

2.8 Definition. The population average treatment effect (SATE) is defined as

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0).$$

The population average treatment effect (PATE, or simply ATE) is defined as

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

2.9 Remark. The latter implicitly assumes that the n units are sampled from a *super-population*, so $Y_i(0)$ and $Y_i(1)$ follow an unknown bivariate probability distribution.

The role of randomisation

Intuitively, it should be possible to estimate the treatment effect *at the population level* if the treatment is randomised. This can be formalised by the following assumption:

2.10 Assumption (Randomisation)**.** $\boldsymbol{A}_{[n]} \perp\!\!\!\perp \boldsymbol{Y}_{[n]}(\boldsymbol{a}_{[n]}) \mid \boldsymbol{X}_{[n]}$ for $\boldsymbol{a}_{[n]} \in \mathcal{A}^n$.

The conditioning on $\boldsymbol{X}_{[n]}$ can be removed if $\boldsymbol{X}$ is not used in the treatment assignment (such as in Examples 2.1 and 2.2).

2.11 Remark (Fatalism). To better understand Assumption 2.10, it is often helpful to view $\boldsymbol{Y}_{[n]}(0)$ and $\boldsymbol{Y}_{[n]}(1)$ as determined prior to treatment assignment. The randomness of $\boldsymbol{A}_{[n]}$ given $\boldsymbol{X}_{[n]}$ (e.g. picking balls from an urn or a computer pseudo-random number generator) should then be independent of the potential outcomes. From a statistical point of view, this fatalism interpretation is unncessary. One may regard the statistical inference as being conditional on the potential outcomes.

Note that Assumption 2.10 is different from $\boldsymbol{A}_{[n]} \perp\!\!\!\perp \boldsymbol{Y}_{[n]} \mid \boldsymbol{X}_{[n]}$, as $Y_i = Y_i(A_i)$ generally depends on A_i.

Recall that we are using $\boldsymbol{X}$, A, and Y to refer to a generic $\boldsymbol{X}_i$, A_i, and Y_i when they are iid.

2.12 Theorem (Causal identification in randomised experiments)**.** *Consider a Bernoulli trial with covariates (Example 2.3), where $\{\boldsymbol{X}_i, A_i, Y_i(a), a \in \mathcal{A}\}$ are iid. Suppose the above assumptions are given and*

$$\mathbb{P}(A = a \mid \boldsymbol{X} = \boldsymbol{x}) > 0,\ \forall a \in \mathcal{A}, \boldsymbol{x}, \tag{2.1}$$

then we have, for all $a \in \mathcal{A}$ and $\boldsymbol{x}$,

$$\Big(Y(a) \mid \boldsymbol{X} = \boldsymbol{x}\Big) \overset{d}{=} \Big(Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}\Big), \tag{2.2}$$

where $\overset{d}{=}$ means the random variables have the same distribution.

Proof. For any $y \in \mathbb{R}$, $a \in \mathcal{A}$, and $\boldsymbol{x}$,

$$\begin{aligned}\mathbb{P}(Y(a) \le y \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{P}(Y_i(a) \le y \mid \boldsymbol{X} = \boldsymbol{x}, A = a)\\ &= \mathbb{P}(Y \le y \mid \boldsymbol{X} = \boldsymbol{x}, A = a),\end{aligned}$$

where the first equality uses Assumption 2.10 and the second uses Assumption 2.6. □

2.13 Remark. Equation (2.1) is called the *positivity* assumption. It is also called the *overlap* assumption because (2.1) implies that $\boldsymbol{X} \mid A = a$ has the same support for all a. It makes sure the right hand side of (2.2) is well defined.

2.14 Corollary. *Under the assumptions in Theorem 2.12,*

$$ATE = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\Big\{\mathbb{E}[Y \mid A = 1, \boldsymbol{X}] - \mathbb{E}[Y \mid A = 0, \boldsymbol{X}]\Big\}. \tag{2.3}$$

If $\mathbb{P}(A = 1 \mid \boldsymbol{X})$ does not depend on $\boldsymbol{X}$ (Example 2.1), then

$$ATE = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]. \tag{2.4}$$

Proof. Equation (2.3) follows from taking the expectation for (2.2) and then averaging over $\boldsymbol{X}$. For (2.4), we prove it in the case of discrete $\boldsymbol{X}$. Since $A \perp\!\!\!\perp \boldsymbol{X}$, we have $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid A = 0) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid A = 1)$. By using Theorem 2.12 and the law of total expectation,

$$\begin{aligned}\mathbb{E}[Y(1)] &= \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}[Y \mid A = 1, \boldsymbol{X} = \boldsymbol{x}]\, \mathbb{P}(\boldsymbol{X} = \boldsymbol{x})\\ &= \sum_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E}[Y \mid A = 1, \boldsymbol{X} = \boldsymbol{x}]\, \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid A = 1)\\ &= \mathbb{E}[Y \mid A = 1].\end{aligned}$$

Similarly, $\mathbb{E}[Y(0)] = \mathbb{E}[Y \mid A = 0]$. □

2.15 Remark. Results like (2.2), (2.3), (2.4) are called *causal identification.*because they equate a counterfactual quantity on the left hand side with a factual (so estimable) quantity on the right hand side.

2.3 Randomisation distribution of causal effect estimator

Neyman considered the following difference-in-means estimator:

$$\hat{\beta} = \bar{Y}_1 - \bar{Y}_0, \text{ where } \bar{Y}_1 = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i}, \bar{Y}_0 = \frac{\sum_{i=1}^n (1 - A_i) Y_i}{\sum_{i=1}^n 1 - A_i}. \tag{2.5}$$

Denote $\boldsymbol{Y}(a) = (Y_1(a), Y_2(a), \dots, Y_n(a))^T$ for $a \in \mathcal{A}$. Neyman studied the conditional distribution of $\hat{\beta}$ given the potential outcomes $\boldsymbol{Y}(0), \boldsymbol{Y}(1)$. We may refer to this as the *randomization distribution*, because the only randomness left in $\hat{\beta}$ comes from the randomization of the treatment $\boldsymbol{A}_{[n]}$.

2.16 Theorem. *Let Assumptions 2.6, 2.7 and 2.10 be given and suppose the treatment assignments A_i are sampled without replacement according to Example 2.2. Then*

$$\mathbb{E}[\hat{\beta} \mid \boldsymbol{Y}(0), \boldsymbol{Y}(1)] = SATE = \frac{1}{n} \sum_{i=1}^n Y_i(1) - Y_i(0), \tag{2.6}$$

$$\operatorname{Var}\left(\hat{\beta} \mid \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) = \frac{1}{n_0} S_0^2 + \frac{1}{n_1} S_1^2 - \frac{S_{01}^2}{n}, \tag{2.7}$$

where $n_0 = n - n_1$, $S_a^2 = \sum_{i=1}^n (Y_i(a) - \bar{Y}(a))^2/(n-1)$, $\bar{Y}(a) = \sum_{i=1}^n Y_i(a)/n$ for $a = 0, 1$, and $S_{01}^2 = \sum_{i=1}^n (Y_i(1) - Y_i(0) - SATE)^2/(n-1)$.

The expectation and variance are computed under the *randomisation distribution* distribution of $\hat{\beta}$, in which the potential outcomes $\boldsymbol{Y}(1)$ and $\boldsymbol{Y}(0)$ are treated as fixed and the randomness comes from the randomisation of $\boldsymbol{A}_{[n]}$. As a consequence, the right hand side of (2.6) and (2.7) depend on the unobserved potential outcomes $\boldsymbol{Y}(1)$ and $\boldsymbol{Y}(0)$.

Proof of Equation (2.6). For simplicity of exposition, we omit the conditioning on $\boldsymbol{Y}(0), \boldsymbol{Y}(1)$ below. By using $\mathbb{E}[A_i] = n_1/n$, the consistency assumption and the linearity of expectations,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}\left[\frac{1}{n_1} \sum_{i=1}^n A_i Y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - A_i) Y_i\right] \\
&= \mathbb{E}\left[\frac{1}{n_1} \sum_{i=1}^n A_i Y_i(1) - \frac{1}{n_0} \sum_{i=1}^n (1 - A_i) Y_i(0)\right] \\
&= \frac{1}{n_1} \sum_{i=1}^n \frac{n_1}{n} Y_i(1) - \frac{1}{n_0} \sum_{i=1}^n \frac{n_0}{n} Y_i(0) \\
&= \bar{Y}(1) - \bar{Y}(0).
\end{aligned}$$

□

2.17 Exercise. Prove (2.7). *Hint: Let* $Y_i^*(a) = Y_i(a) - \bar{Y}(a), a = 0, 1$. *Show that*

$$\mathrm{Var}\left(\hat{\beta} \mid \boldsymbol{Y}(0), \boldsymbol{Y}(1)\right) = \mathbb{E}\left[\Big(\sum_{i=1}^{n} \frac{A_i}{n_1} Y_i^*(1) - \frac{1 - A_i}{n_0} Y_i^*(0)\Big)^2\right].$$

Then expand the sum of squares and use

$$\mathbb{E}[A_i A_{i'}] = \frac{n_1}{n}\frac{n_1 - 1}{n - 1}, i \neq i' \textit{ and } \sum_{i=1}^{n} Y_i^*(a) = 0.$$

To get interval estimators, it is to estimate the variance in (2.7). However, S_{01}^2 is non-estimable. Why is that? Notice that S_{01}^2 is the sample variance of the individual treatment effect and depends on the covariance of $Y_i(1)$ and $Y_i(0)$, which can never be observed together (the "fundamental problem of causal inference").

Instead, it is common to estimate the variance (2.7) by $\hat{S}_0^2/n_0 + \hat{S}_1^2/n_1$, where

$$\hat{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n} A_i (Y_i - \bar{Y}_1)^2,\ \hat{S}_0^2 = \frac{1}{n_0 - 1}\sum_{i=1}^{n} (1 - A_i)(Y_i - \bar{Y}_0)^2.$$

This is an unbiased estimator of $S_0^2/n_0 + S_1^2/n_1$ (the proof is similar to that of (2.6) and is left as an exercise). Thus we get a conservative (on average) estimator of the variance of $\hat{\beta}$.

Distributional results are further needed to form confidence intervals. Central limit theorems can be established by assuming that the potential outcomes in $\boldsymbol{Y}(0)$ and $\boldsymbol{Y}(1)$ are not too volatile.[6].

2.18 Remark. One drawback of Neyman's randomisation inference is that it is difficult to extend it to settings with covariates (unless the covariates are discrete). The main obstacle is that the randomisation distribution necessarily depends on unobserved potential outcomes.

2.4 Randomisation test of sharp null hypothesis

Fisher[7] appears to be the first to grasp fully the importance of randomisation for credible causal inference.[8]

Testing no causal effect

Fisher considered testing the *sharp null hypothesis* (or *exact null hypothesis*) $H_0 : Y_i(0) = Y_i(1), \forall i \in [n]$.

Using H_0, we can impute all the missing potential outcomes (Table 2.2).

Suppose $\boldsymbol{A}_{[n]}$ in Table 2.2 is randomised by sampling without replacement (Example 2.2). Consider Neyman's difference-in-means estimator $\hat{\beta} = \bar{Y}_1 - \bar{Y}_0$. Because the way $\boldsymbol{A}_{[n]}$ is randomised, the following 6 scenarios are equally likely:

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	**-3.7**	-3.7	1	-3.7
2	2.3	**2.3**	0	2.3
3	**7.4**	7.4	1	7.4
4	0.8	**0.8**	0	0.8

Table 2.2: Illustration of Fisher's randomisation test.

(i) $\boldsymbol{A}_{[4]} = (1,1,0,0)$, $\hat{\beta} = (-3.7 + 2.3 - 7.4 - 0.8)/2 = -4.8$.

(ii) $\boldsymbol{A}_{[4]} = (1,0,1,0)$, $\hat{\beta} = (-3.7 - 2.3 + 7.4 - 0.8)/2 = 0.3$.

(iii) $\boldsymbol{A}_{[4]} = (1,0,0,1)$, $\hat{\beta} = (-3.7 - 2.3 - 7.4 + 0.8)/2 = -6.3$.

(iv) $\boldsymbol{A}_{[4]} = (0,1,1,0)$, $\hat{\beta} = (3.7 + 2.3 + 7.4 - 0.8)/2 = 6.3$.

(v) $\boldsymbol{A}_{[4]} = (0,1,0,1)$, $\hat{\beta} = (3.7 + 2.3 - 7.4 + 0.8)/2 = -0.3$.

(vi) $\boldsymbol{A}_{[4]} = (0,0,1,1)$, $\hat{\beta} = (3.7 - 2.3 + 7.4 + 0.8)/2 = 4.8$.

The observed realisation is the second.

Fisher proposed to test H_0 based on how extreme the observed $\hat{\beta}$ is compared to other potential values of $\hat{\beta}$.

A more general setup

Next we formalise the idea above. We first consider a more general class of null hypotheses (let β be a fixed value):

$$H_0 : Y_i(1) - Y_i(0) = \beta,\ \forall i \in [n], \tag{2.8}$$

This is still a very strong hypothesis: it says the individual treatment effect is always a fixed value β.

Using the consistency assumption, (2.8) allow us to impute the potential outcomes as

$$Y_i(a) = \begin{cases} Y_i, & \text{if } a = A_i, \\ Y_i + \beta, & \text{if } a > A_i, \\ Y_i - \beta, & \text{if } a < A_i. \end{cases} \tag{2.9}$$

A more compact form is $\boldsymbol{Y}_{[n]}(\boldsymbol{a}_{[n]}) = \boldsymbol{Y}_{[n]} + \beta(\boldsymbol{a}_{[n]} - \boldsymbol{A}_{[n]})$.

In randomisation inference, it is also necessary to choose a test statistic $T(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]})$. An example is the difference-in-means estimator $\hat{\beta}$.

Randomisation distribution

The key step is to derive the *randomisation distribution* of T. There are two ways to do this:

(i) Consider the distribution of $T_1(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(0))$ given $\boldsymbol{X}_{[n]}$ and $\boldsymbol{Y}_{[n]}(0)$;

(ii) Consider the distribution of $T_2(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(\boldsymbol{A}_{[n]}))$ given $\boldsymbol{X}_{[n]}$, $\boldsymbol{Y}_{[n]}(0)$, and $\boldsymbol{Y}_{[n]}(1)$;

In both cases, the randomness comes from the randomisation of $\boldsymbol{A}_{[n]}$. The first approach tries to test the conditional independence $\boldsymbol{A}_{[n]} \perp\!\!\!\perp \boldsymbol{Y}_{[n]}(0) \mid \boldsymbol{X}$. The second approach tries to directly obtain the randomisation distribution of $T(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]})$ and bears a resemblance to Neyman's inference.

It is easy to see that the two approaches are exactly the same if $\beta = 0$. Exercise 2.21 below shows that they are still equivalent if $\beta \neq 0$. For more complex hypotheses, however, one approach can be more convenient than the other.

Let $\mathcal{F} = (\boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(0), \boldsymbol{Y}_{[n]}(1))$. The randomisation distributions in the two approaches above are given by

$$F_1(t) = \mathbb{P}\left(T_1(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(0)) \leq t \,\middle|\, \mathcal{F}\right),$$

and

$$F_2(t) = \mathbb{P}\left(T_2(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(\boldsymbol{A}_{[n]})) \leq t \,\middle|\, \mathcal{F}\right).$$

The observed test statistics are

$$T_1 = T_1(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]} - \beta \boldsymbol{A}_{[n]}),\ T_2 = T_2(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}).$$

The one-sided p-value is the probability of observing the same or a more extreme test statistic than the observed statistic T,

$$P_m = F_m(T_m),\ m = 1, 2.$$

An equivalent and perhaps more informative representation is

$$P_1 = \mathbb{P}^*\left(T_1(\boldsymbol{A}^*_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(0)) \leq T_1 \mid \mathcal{F}\right),$$

where $\boldsymbol{A}^*_{[n]}$ is an independent copy of $\boldsymbol{A}$, so $\boldsymbol{A}^*_{[n]} \mid \boldsymbol{X}_{[n]} \sim \pi$ but $\boldsymbol{A}^* \perp\!\!\!\perp \boldsymbol{A}$, and $\mathbb{P}^*$ means that the probability is with respect to the distribution of $\boldsymbol{A}^*$. The other p-value P_2 can be similarly defined. Note that the dependence of P_1 and P_2 on $\mathcal{F}$ are omitted.

A level-α randomisation test then rejects H_0 if $P_m \leq \alpha$.

2.19 Theorem. *Under SUTVA (Assumptions 2.6 and 2.7) and* H_0, $\mathbb{P}(P_m \leq \alpha) \leq \alpha,\ \forall\, 0 < \alpha < 1, m = 1, 2.$

Proof. This follows from the property of the distribution function: If $F(t)$ is the distribution function of a random variable T, then $F(T)$ stochastically dominates the uniform distribution on $[0,1]$. To show this, let $F^{-1}(\alpha) = \sup\{t \mid F(t) \le \alpha\}$. By using the fact that $F(t)$ is non-decreasing and right-continuous (see Figure 2.1),

$$\mathbb{P}(F(T) \le \alpha) = \mathbb{P}(T < F^{-1}(\alpha)) = \lim_{t \uparrow F^{-1}(\alpha)} \mathbb{P}(T \le t) \le \alpha.$$

This shows that $\mathbb{P}\left(P_m \le \alpha \mid \mathcal{F}\right) \le \alpha$. To get the unconditional result, take the expectation over the random variables in $\mathcal{F}$. □

2.20 Remark. The *probability integral transform* says that if T is a continuous random variable and $F(t)$ is its distribution function, then $F(T)$ is uniformly distributed on $[0,1]$. However, we cannot directly use this well known result here because our T has a discrete (conditional) distribution.

The role of randomisation

Theorem 2.19 essentially just restates a basic fact in probability theory. Notice that the randomisation assumption (Assumption 2.10) is not required in Theorem 2.19. So what is the role of randomisation?

The randomisation assumption (and basically all other assumptions in Theorem 2.19) make the p-values possible to compute. To see this, by definition,

$$\begin{aligned} F_1(t) &= \mathbb{P}\left(T_1(\boldsymbol{A}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(0)) \le t \,\middle|\, \mathcal{F}\right) \\ &= \sum_{\boldsymbol{a}_{[n]} \in \mathcal{A}^n} \mathbb{P}(\boldsymbol{A}_{[n]} = \boldsymbol{a}_{[n]} \mid \mathcal{F}) \cdot I\Big(T(\boldsymbol{a}_{[n]}, \boldsymbol{X}_{[n]}, \boldsymbol{Y}_{[n]}(0)) \le t\Big) \end{aligned} \tag{2.10}$$

The conditional independence in Assumption 2.10 and H_0 (so there is no further randomness in $\boldsymbol{Y}(1)$ after conditioning on $\boldsymbol{Y}(0)$) allow us to replace the first term by

$$\mathbb{P}(\boldsymbol{A}_{[n]} = \boldsymbol{a}_{[n]} \mid \mathcal{F}) = \mathbb{P}(\boldsymbol{A}_{[n]} = \boldsymbol{a}_{[n]} \mid \boldsymbol{X}_{[n]}) = \pi(\boldsymbol{a}_{[n]} \mid \boldsymbol{X}_{[n]}),$$

which is the known randomisation scheme.

2.21 Exercise. Show that the two tests for H_0 above (based on different randomisation distributions) are equivalent. *Hint: Construct a one-to-one mapping between the functions T_1 and T_2.*

2.22 Example. Some commonly used test statistics in randomisation inference include (to simplify the exposition, we consider $\beta = 0$):

(i) t-statistic: $T = |\hat{\beta}| / \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}$, where $\hat{\beta}$ is the difference-in-means estimator (2.5).

(ii) The *Mann-Whitney U statistic* (equivalent to *Wilcoxon's rank-sum statistic*): $T = \sum_{i,j=1}^{n} I(A_i > A_j) \cdot I(Y_i > Y_j)$.

(iii) Regression-adjusted statistic: T is the least squares coefficient of A in the regression $Y \sim A + X + A : X$ (R formula notation, see next section).

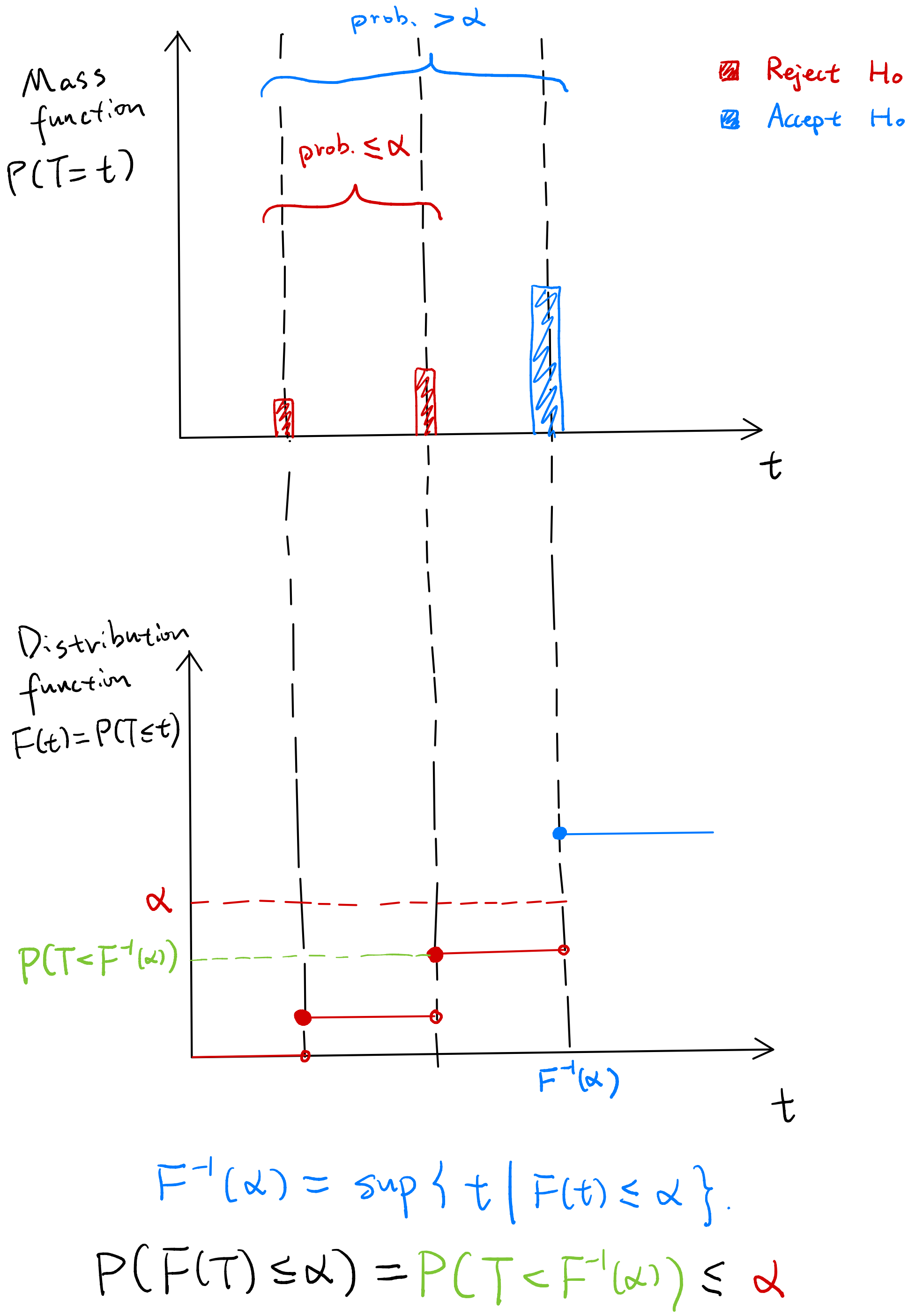

prob. > α
prob. ≤ α
Mass function P(T=t)
t
🟥 Reject H₀
🟦 Accept H₀
Distribution function F(t)=P(T≤t)
α
P(T<F⁻¹(α))
F⁻¹(α)
t
F⁻¹(α) = sup{t | F(t) ≤ α}.
P(F(T) ≤ α) = P(T < F⁻¹(α)) ≤ α

Figure 2.1: Illustration of the randomisation test.

Practical issues

Computing the p-value exactly via its definition (2.10) can be computationally intensive because it requires summing over $\mathcal{A}^n$. In practice, $F(T)$ is often computed by Monte-Carlo simulation.

In the example sheet, you will learn how to obtain an estimator of β (suppose H_0 is true for some unknown β) by using the Hodges-Lehmann estimator.[9] You will also explore how to obtain a confidence interval for β.

2.5 Super-population inference and regression adjustment

Next we consider a different inference paradigm. Rather than conditioning or assuming a hypothesis on the potential outcomes, we will assume the potential outcomes are drawn from a "super-population" and are independent and identically distributed (i.i.d.). We will discuss asymptotic super-population inference by considering the simple Bernoulli trial (Example 2.1), so $A_i \perp\!\!\!\perp \boldsymbol{X}_i$. Further, suppose $(A_i, \boldsymbol{X}_i, Y_i(0), Y_i(1))$ are i.i.d.. To focus on the main ideas, we assume that $\boldsymbol{X}$ is centred, i.e. $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{0}$.

Denote $\pi = \mathbb{P}(A = 1)$, $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]$, and $\beta = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$. Corollary 2.14 shows that β in a randomised experiment is equal to the (population) average treatment effect.

We shall consider three regression estimators of β:

$$(\hat{\alpha}_1, \hat{\beta}_1) = \arg\min_{\alpha,\beta} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta A_i)^2, \tag{2.11}$$

$$(\hat{\alpha}_2, \hat{\beta}_2, \hat{\boldsymbol{\gamma}}_2) = \arg\min_{(\alpha,\beta,\boldsymbol{\gamma})} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta A_i - \boldsymbol{\gamma}^T \boldsymbol{X}_i)^2, \tag{2.12}$$

$$(\hat{\alpha}_3, \hat{\beta}_3, \hat{\boldsymbol{\gamma}}_3, \hat{\boldsymbol{\delta}}_3) = \arg\min_{(\alpha,\beta,\boldsymbol{\gamma},\boldsymbol{\delta})} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \alpha - \beta A_i - \boldsymbol{\gamma}^T \boldsymbol{X}_i - A_i(\boldsymbol{\delta}^T \boldsymbol{X}_i))^2. \tag{2.13}$$

It is easy to show that $\hat{\beta}_1$ is the difference-in-means estimator (2.5). The rationale for $\hat{\beta}_2$ is that the adjustment tends to improve precision if $\boldsymbol{X}$ is correlated with Y. This is known as the analysis of covariance (ANCOVA)[10]. The third estimator further models treatment effect heterogeneity through the interaction term $A\boldsymbol{X}$.

The classical linear regression theory for these estimators assumes the regression models are correct. Below we provide a modern analysis that allows model misspecification by using the *M-estimation* theory. Our analysis is a compact version of previous results[11].

Let's first write down the population version of the least squares problems:

$$(\alpha_1, \beta_1) = \arg\min_{\alpha,\beta} \mathbb{E}[(Y - \alpha - \beta A)^2], \tag{2.14}$$

$$(\alpha_2, \beta_2, \boldsymbol{\gamma}_2) = \arg\min_{(\alpha,\beta,\boldsymbol{\gamma})} \mathbb{E}[(Y - \alpha - \beta A - \boldsymbol{\gamma}^T \boldsymbol{X})^2], \tag{2.15}$$

$$(\alpha_3, \beta_3, \boldsymbol{\gamma}_3, \boldsymbol{\delta}_3) = \arg\min_{(\alpha\beta,\boldsymbol{\gamma},\boldsymbol{\delta})} \mathbb{E}[(Y - \alpha - \beta A - \boldsymbol{\gamma}^T \boldsymbol{X} - A \cdot (\boldsymbol{\delta}^T \boldsymbol{X}))^2]. \tag{2.16}$$

By the law of large numbers, we expect $\hat{\beta}_m$ converges to β_m, $m = 1, 2, 3$, as $n \to \infty$. To focus on the essential ideas, below we will omit the regularity conditions (for example, to ensure these parameters exist and the central limit theorems hold).

2.23 Lemma. *Suppose $(\boldsymbol{X}_i, A_i, Y_i)$ are iid, $A \perp\!\!\!\perp X$, $\mathbb{E}[\boldsymbol{X}] = 0$. Then $\alpha_1 = \alpha_2 = \alpha_3$ and $\beta_1 = \beta_2 = \beta_3 = \beta$.*

Proof. By taking partial derivatives of $\mathbb{E}[(Y - \alpha - \beta A - \boldsymbol{\gamma}^T \boldsymbol{X} - A \cdot (\boldsymbol{\delta}^T \boldsymbol{X}))^2]$ with respect to α and β, we obtain

$$\begin{aligned}\mathbb{E}[Y - \alpha_3 - \beta_3 A - \boldsymbol{\gamma}_3^T \boldsymbol{X} - A(\boldsymbol{\delta}_3^T \boldsymbol{X})] = 0,\\ \mathbb{E}[A(Y - \alpha_3 - \beta_3 A - \boldsymbol{\gamma}_3^T \boldsymbol{X} - A(\boldsymbol{\delta}^T \boldsymbol{X}))] = 0.\end{aligned}$$

Using $\mathbb{E}[\boldsymbol{X}] = 0$ and $A \perp\!\!\!\perp \boldsymbol{X}$, they can be simplified to

$$\begin{aligned}\mathbb{E}[Y - \alpha_3 - \beta_3 A] = 0,\\ \mathbb{E}[A(Y - \alpha_3 - \beta_3 A)] = 0.\end{aligned}$$

Following the same derivation, these two equations also hold for the other estimators. By cancelling α_3 in the equations, we get $\beta_3 = \beta$. □

2.24 Exercise. Derive $\alpha_1, \alpha_2, \alpha_3$, $\boldsymbol{\gamma}_2$, $\boldsymbol{\gamma}_3$, and $\boldsymbol{\delta}_3$ in terms of the distribution of $(\boldsymbol{X}, A, Y)$ and in terms of the distribution of $(\boldsymbol{X}, A, Y(0), Y(1))$.

2.25 Remark. Notice that Lemma 2.23 does not rely on the correctness of the linear model. Modern causal inference often tries to make minimal assumptions about the data and avoid relying on specific statistical models ("all models are wrong, but some are useful").

We will use a general result for least squares estimators to study the asymptotic behaviour of $\hat{\beta}_1$,$\hat{\beta}_2$, and $\hat{\beta}_3$.[12]

2.26 Lemma. *Suppose $(\boldsymbol{Z}_i, Y_i), i = 1, \dots, n$ are iid and $\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]$ is positive definite. Let $\boldsymbol{\theta} = (\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T])^{-1}(\mathbb{E}[\boldsymbol{Z}Y])$ be the population least squares parameter and*

$$\hat{\boldsymbol{\theta}} = \Big(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}\boldsymbol{Z}^T\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Z}Y\Big)$$

be its sample estimator. Let $\epsilon_i = Y_i - \boldsymbol{\theta}^T \boldsymbol{Z}_i$ be the regression error. Suppose Y and $\boldsymbol{Z}$ have bounded fourth moments, then $\hat{\boldsymbol{\theta}} \overset{p}{\to} \boldsymbol{\theta}$ and

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \overset{d}{\to} \mathrm{N}\Big(0, \{\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]\}^{-1}\, \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T \epsilon^2]\{\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T]\}^{-1}\Big), \tag{2.17}$$

as $n \to \infty$.

2.27 Remark. The variance in (2.17) is said to be robust to *heteroscedasticity*, meaning that $\mathrm{Var}(\epsilon \mid X)$ depends on $\boldsymbol{X}$ in a regression model. It can be obtained using the general M-estimation (M for maximum/minimum) or Z-estimation (Z for zero) theory.

Informal proof of (2.17). Notice that $\hat{\boldsymbol{\theta}}$ is an empirical solution to the equation

$$\mathbb{E}[\boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{Z}, Y)] = \mathbf{0},$$

where

$$\boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{Z}, Y) = \boldsymbol{Z} \cdot (Y - \boldsymbol{Z}^T \boldsymbol{\theta}) = \boldsymbol{Z}\epsilon. \tag{2.18}$$

For a general function $\boldsymbol{\psi}$, the Z-estimation theory shows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathrm{N}\Big(\mathbf{0}, \Big\{\mathbb{E}\Big[\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big]\Big\}^{-1} \mathbb{E}\left[\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{\theta})^T\right] \Big\{\mathbb{E}\Big[\frac{\partial \boldsymbol{\psi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big]\Big\}^{-T}\Big). \tag{2.19}$$

By plugging in (2.18), we obtain (2.17). The asymptotic normality (2.19) follows from the argument below. Using Taylor's expansion,

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\hat{\boldsymbol{\theta}}; \boldsymbol{Z}_i, Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{Z}_i, Y_i) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{Z}_i, Y_i)\Big] + R_n. \end{aligned}$$

By using $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, it can be shown that the residual term R_n is asymptotically smaller than the other two terms and can be ignored. Thus

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx \mathbb{E}\Big\{\Big[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{Z}, Y)\Big]^{-1}\Big\}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{\theta}; \boldsymbol{Z}_i, Y_i)\Big]. \tag{2.20}$$

The first term on the right hand side converges in probability to $\mathbb{E}[\partial \boldsymbol{\psi}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}]$. The second term converges in distribution to a normal random variable with variance $\mathbb{E}[\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{\theta})^T]$. Using Slutsky's theorem, we arrive at (2.19). □

2.28 Remark. The Z-estimation theory generalises the asymptotic theory for maximum likelihood estimator (MLE), where $\boldsymbol{\psi}$ is the score function (gradient of the log-likelihood). In that case, it can be shown that $\mathbb{E}[\partial \boldsymbol{\psi}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}] = \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{\theta})^T]$ is the Fisher information matrix, which you may recognise from your undergraduate lectures (Part II Principles of Statistics).

2.29 Exercise. Let $\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}$ be the error terms in the three regression estimators:

$$\epsilon_{im} = Y_i - \alpha_m - \beta_m A_i - \boldsymbol{\gamma}_m^T \boldsymbol{X}_i - A_i(\boldsymbol{\delta}_m^T \boldsymbol{X}_i),\ m = 1, 2, 3.$$

Here we are using the convention $\boldsymbol{\gamma}_1 = 0$ and $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \mathbf{0}$. By using Lemma 2.26 with different $\boldsymbol{Z}$ and $\boldsymbol{\theta}$, show that, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_m - \beta) \xrightarrow{d} \mathrm{N}(0, V_m),\ \text{where } V_m = \frac{\mathbb{E}[(A - \pi)^2 \epsilon_m^2]}{\pi^2(1 - \pi)^2},\ m = 1, 2, 3.$$

2.30 Theorem. *Suppose $(\boldsymbol{X}_i, A_i, Y_i)$ are iid, $A \perp\!\!\!\perp X$, $\mathbb{E}[\boldsymbol{X}] = 0$. Then as $n \to \infty$,*

$$\sqrt{n}(\hat{\beta}_m - \beta) \xrightarrow{d} N(0, V_m), \; m = 1, 2, 3,$$

where the asymptotic variances satisfy $V_3 \le \min\{V_1, V_2\}$.

Proof. By differentiating (2.16),

$$\mathbb{E}[\epsilon_3] = \mathbb{E}[A\epsilon_3] = 0, \; \mathbb{E}[\boldsymbol{X}\epsilon_3] = \mathbb{E}[A\boldsymbol{X}\epsilon_3] = \mathbf{0}.$$

By Lemma 2.23,

$$\begin{aligned} \epsilon_1 &= \epsilon_3 + \boldsymbol{\gamma}_3^T \boldsymbol{X} + A(\boldsymbol{\delta}_3^T \boldsymbol{X}), \\ \epsilon_2 &= \epsilon_3 + (\boldsymbol{\gamma}_3 - \boldsymbol{\gamma}_2)^T \boldsymbol{X} + A(\boldsymbol{\delta}_3^T \boldsymbol{X}). \end{aligned}$$

Thus, for $m = 1, 2$,

$$\begin{aligned} & \mathbb{E}[(A - \pi)^2 \epsilon_m^2] - \mathbb{E}[(A - \pi)^2 \epsilon_3^2] \\ = & \mathbb{E}\left[(A - \pi)^2 \big((\boldsymbol{\gamma}_3 - \boldsymbol{\gamma}_m)^T \boldsymbol{X} + A(\boldsymbol{\delta}_3^T \boldsymbol{X})\big)^2\right] \\ \ge & 0. \end{aligned}$$

□

2.31 Exercise. Use your results in Exercise 2.24 to derive the conditions under which $V_1 = V_2 = V_3$. Show that $V_2 \le V_1$ is not always true, therefore, regression adjustment does not always reduce the asymptotic variance (if not done properly)!

2.32 Remark. The assumption $\mathbb{E}[\boldsymbol{X}] = 0$ is useful to simplify the calculations above. In practice, we obviously don't know if this assumption is true, so it is common to centre $\boldsymbol{X}$ before solving the least squares problem. Let $\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3$ be the least squares estimators in eqs. (2.11) to (2.13) with $\boldsymbol{X}_i$ replaced by $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ where $\bar{\boldsymbol{X}} = \sum_{i=1}^n \boldsymbol{X}_i / n$. It is easy to show that $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_2 = \hat{\beta}_2$ (because of the intercept term) and $\tilde{\beta}_3 = \hat{\beta}_3 + \hat{\boldsymbol{\delta}}_3^T \bar{\boldsymbol{X}}$. Therefore, $\tilde{\beta}_1$ and $\tilde{\beta}_2$ have the same asymptotic distributions as $\hat{\beta}_1$ and $\hat{\beta}_2$, even when $\mathbb{E}[\boldsymbol{X}] \ne 0$. However, the asymptotic variance of $\tilde{\beta}_3$ (denoted as $\tilde{V}_3$) is larger than that of $\hat{\beta}_3$. It can be shown that $\tilde{V}_3 = V_3 + \boldsymbol{\delta}_3^T \boldsymbol{\Sigma} \boldsymbol{\delta}_3$ and $\tilde{V}_3 \le \min\{V_1, V_2\}$ still holds.

2.6 Comparison of different modes of inference

See Table 2.3.

Notes

[1] Ye, T., Shao, J., & Zhao, Q. (2020). Principles for covariate adjustment in analyzing randomized clinical trials. arXiv: 2009.11828 `[stat.ME]`.

[2] Li, X., & Ding, P. (2019). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 241–268. doi:10.1111/rssb.12353.

	Neyman's inference	Randomisation test	Regression
Population	Finite	Finite	Super-population
Randomness	Only A	Only A	$A, \boldsymbol{X}, Y$
Point estimator	Difference-in-means	Hodges-Lehmann (example sheet)	Least squares
Inference	Exact variance, CI asymptotic	Exact (approximate if using Monte-Carlo)	Asymptotic
Covariate adjustment	No	Yes	Yes
Effect heterogeneity	Yes	No	Yes

Table 2.3: Side-by-side comparison of the modes of inference

[3] Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, *5*(4), 465–472. doi:10.1214/ss/1177012031.

[4] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350.

[5] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. doi:10.1080/01621459.1986.10478354.

[6] For a recent review, see Li, X., & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, *112*(520), 1759–1769. doi:10.1080/01621459.2017.1295865.

[7] Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). Oliver and Boyd, Edinburgh and London.

[8] See Chapter 2 of Imbens and Rubin, 2015, for a historical account.

[9] Rosenbaum, P. R. (1993). Hodges-Lehmann point estimates of treatment effect in observational studies. *Journal of the American Statistical Association*, *88*(424), 1250–1253. doi:10.1080/01621459.1993.10476405.

[10] Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Oliver and Boyd, Edinburgh and London.

[11] Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, *27*(23), 4658–4677. doi:10.1002/sim.3113. Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences* (1st ed.). Cambridge University Press, Chapter 7. For finite-population randomisation inference, see Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, *7*(1), 295–318. doi:10.1214/12-aoas583

[12] See, for example, White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. doi:10.2307/1912934.

Chapter 3

Path analysis

Last Chapter introduced potential outcomes to describe a causal effect. This approach is most convenient when we focus on a single cause-effect pair.

An alternative and more holistic approach to is to use a graph, in which directed edges represent causal effects. This framework traces back to a series of works on *path analysis* by Sewall Wright a century ago.[1]

3.1 Graph terminology

We start with introducing some terminology in graph theory.

3.1 Definition (Graph and subgraph)**.** A *graph* $\mathcal{G} = (V, E)$ is defined by its finite vertex set V and its edge set $E \subseteq V \times V$ containing ordered pairs of vertices. The *subgraph* of $\mathcal{G}$ restricted to $A \subset V$ is $\mathcal{G}_A = (A, E_A)$ where $E_A = \{(i,j) \in E \mid i,j \in A\}$.

3.2 Definition (Directed edge and graph)**.** An edge (i,j) is called *directed* and written as $i \to j$ if $(i,j) \in E$ but $(j,i) \notin E$. Vertex i is called a *parent* of j and j a *child* of i if $i \to j$. The set of parents of a vertex i is denoted as $pa_{\mathcal{G}}(i)$ or simply $pa(i)$. A graph $\mathcal{G}$ is called a *directed graph* if all its edges are directed.

3.3 Definition (Path and cycle)**.** A *path* between i and j is a sequence of distinct vertices $k_0 = i, k_1, k_2, \ldots, k_m = j$ such that the consecutive vertices are adjacent, that is, $(k_{l-1}, k_l) \in E$ or $(k_l, k_{l-1}) \in E$ for all $l = 1, 2, \ldots, m$.

A *directed path* from i to j is a path in which all the arrows are going "forward", that is, $(k_{l-1}, k_l) \in E$ for all $l = 1, 2, \ldots, m$.

A *cycle* is a directed path with the only modification that the first and last vertices are the same $k_m = k_0$.

A *directed acyclic graph* (DAG) is a directed graph with no cycles.

3.4 Definition (Ancestor and descendant)**.** In a DAG $\mathcal{G}$, a vertex i is an *ancestor* of j if there exists a directed path from i to j; conversely, j is called a *descendant* of i. Let $an_{\mathcal{G}}(I)$ denote the union of ancestors and $de_{\mathcal{G}}(I)$ the union of descendants of all vertices in I.

3.5 Exercise. Show that a directed graph is acyclic if and only if the vertices can be relabeled in a way that the edges are monotone in the label (this is called a *topological ordering*). In other words, there exists a permutation $(k_1, \dots, k_p)$ of $(1, \dots, p)$ such that $(i, j) \in E$ implies $k_i < k_j$.

3.6 Exercise. Show that for any $J \subset [p]$, there exists $i \notin J$ such that all the descendants of i in a DAG $\mathcal{G}$ are in J. *Hint: Use the topological ordering.*

3.7 Definition (Graphical model)**.** A *graphical model* is a graph $\mathcal{G} = (V, E)$ and a bijection from the vertex set V of the graph to a set of random variables $\boldsymbol{X}$.

A *causal graphical model* is a collection of graphical models, each corresponding to an intervention on $\boldsymbol{X}$.

For the rest of this course we will focus on DAG models, which provide a natural setup for causal inference.

Some conventions: We will often consider the random variables $\boldsymbol{X} = \boldsymbol{X}_{[p]} = (X_1, \dots, X_p)$ and the graphical model $\mathcal{G} = (V = [p], E)$ with the map $i \to X_i$. To simplify notation, we won't distinguish $\boldsymbol{X}_{[p]}$ with the set $\{X_1, \dots, X_p\}$. We will often not distinguish between $\mathcal{G}$ and the graph induced by the graphical model, $\mathcal{G}[\boldsymbol{X}] = (\boldsymbol{X}, E[\boldsymbol{X}])$ where $(X_i, X_j) \in E[\boldsymbol{X}]$ if and only if $(i, j) \in E$.

3.2 Linear structural equation models

Wright's path analysis applies to random variables that satisfy the *linear structural equation model (SEM)*, a (causal) graphical model defined below.

3.8 Definition (Linear SEM)**.** The random variables $\boldsymbol{X}_{[p]}$ satisfy a linear SEM with respect to a DAG $\mathcal{G} = (V = [p], E)$ if they satisfy

$$X_i = \beta_{0i} + \sum_{j \in pa(X_i)} \beta_{ji} X_j + \epsilon_i, \tag{3.1}$$

where $\epsilon_1, \dots, \epsilon_p$ are mutually independent with mean 0 and finite variance, and the interventional distributions of $\boldsymbol{X}$ also satisfy (3.1) (see Remark 3.9). The parameter β_{ji} is called a *path coefficient.*

Equation (3.1) looks just like a linear model and can indeed be fitted using linear regression. What makes it *structural* or *causal* is an implicit assumption that (3.1) still holds if we make interventions to one or some of the variables. For example, consider the following linear SEM,

$$\begin{aligned} X_1 &= \epsilon_1, \\ X_2 &= 2X_1 + \epsilon_2, \\ X_3 &= X_1 + 2X_2 + \epsilon_3. \end{aligned}$$

By recursive substitution, we have $X_3 = 5\epsilon_1 + 2\epsilon_2 + \epsilon_3$. If we make an intervention that sets X_1 to some value x_1 (for example, give a treatment to an experimental unit), we can still use the above equations to derive the values of $\boldsymbol{X}$ by simply replacing the equation $X_1 = \epsilon_1$ with $X_1 = x_1$:

$$\begin{aligned} X_1 &= x_1, \\ X_2 &= 2X_1 + \epsilon_2, \\ X_3 &= X_1 + 2X_2 + \epsilon_3. \end{aligned}$$

By recursive substitution, we obtain $X_3 = 5x_1 + 2\epsilon_2 + \epsilon_3$ under the intervention $X_1 = x_1$. Thus a structural equation model not only describes the *distribution* of the random variables but also their *data generating mechanism.*

3.9 Remark. The distinction between a structural equation model and a regression model can be further elucidated using counterfactuals. Later in the course we will define the counterfactuals for the intervention $X_j = x_j$ by replacing X_j with x_j in (3.1) and all other X_i by $X_i(x_j)$, the counterfactual value X_i. In the example above, this amounts to

$$\begin{aligned} X_1(x_1) &= x_1, \\ X_2(x_1) &= 2x_1 + \epsilon_2, \\ X_3(x_1) &= x_1 + 2X_2(x_1) + \epsilon_3. \end{aligned}$$

So SEM is not only a model for the factuals (like regression models) but also a model for the counterfactuals (unlike regression models).

We can use matrix notation to write (3.1) more compactly:

$$\boldsymbol{X}_{[p]} = \boldsymbol{\beta}_0 + \boldsymbol{B}^T \boldsymbol{X}_{[p]} + \boldsymbol{\epsilon}_{[p]},$$

where the (i,j)-entry of $\boldsymbol{B}$ is β_{ij}.

3.10 Exercise. Suppose the vertices in a DAG is labelled according to a topological ordering. What property does the matrix $\boldsymbol{B}$ have? Use this property to show that $\mathrm{Cov}(\boldsymbol{X}_{[p]})$ is positive definite.

Given a linear SEM, we may define causal effect of X_i on X_j as the product of path coefficients along all directed paths from i to j.

3.11 Definition. Let $\mathcal{C}(i,j)$ be the collection of all directed paths ("causal paths") from i to j. The *causal effect* of X_i on X_j in a linear SEM is defined as

$$\beta(X_i \to X_j) = \sum_{(k_0,\dots,k_m)\in\mathcal{C}(i,j)} \prod_{l=1}^{m} \beta_{k_{l-1}k_l}. \tag{3.2}$$

Immediately we have, if j precedes i in a topological ordering of $\mathcal{G}$, then $\beta(X_i \to X_j) = 0$, i.e., X_i has no causal effect on X_j.

3.12 Remark. Notice that $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$ is symmetric, but $\beta(X_i \to X_j)$ is clearly not symmetric.

3.3 Path analysis

Wright's path analysis uses the path coefficients to express the covariance matrix of $\boldsymbol{X}$ and clearly describes why "correlation does not imply causation".

3.13 Definition. A path $k_0 = i, k_1, \dots, k_m = j$ between i and j is called *d-connected* or *open* if it does not contain a *collider* $k_{l-1} \to k_l \leftarrow k_{l+1}$.

The letter d in d-connected stands for dependence. Intuitively, a d-connected path introduces dependence between X_i and X_j.

Let $\mathcal{D}(i,j)$ be the collection of all d-connected paths between i and j.

3.14 Theorem (Wright's path analysis)**.** *Suppose the random variables $\boldsymbol{X}_{[p]}$ satisfy the linear SEM* (3.1) *with respect to a DAG $\mathcal{G}$ and are standardised so that Var*$(X_i) = 1$ *for all i. Then*

$$\mathrm{Cov}(X_i, X_j) = \sum_{(k_0,\dots,k_m)\in\mathcal{D}(i,j)} \prod_{l=1}^{m} \beta_{k_{l-1}k_l}. \tag{3.3}$$

Proof. Without loss of generality, let's assume $(1,\dots,p)$ is a topological order of $\mathcal{G}$ and $i < j$. We prove Theorem 3.14 by induction. Equation (3.3) is obviously true if $i = 1$ and $j = 2$. Now suppose (3.3) is true for any $i < j \le k$, where $2 \le k \le p-1$. It suffices to show that (3.3) also holds for $i < j = k+1$. The key is to realise that $\mathcal{D}(i,j)$ can be obtained by taking a union of all paths in $\mathcal{D}(i,l)$ for $l \in pa(j)$ appended with the edge $l \to j$. See Figure 3.1 for an illustration.

By (3.1),

$$X_j = \sum_{l\in pa(j)} \beta_{lj} X_l + \epsilon_j.$$

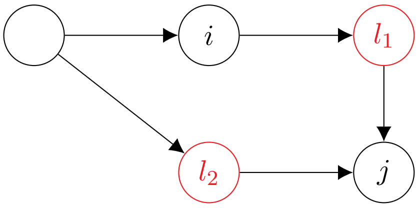


Figure 3.1: Illustration for the proof of Theorem 3.14.

Therefore, using the induction hypothesis and the trivial fact that $X_i \perp\!\!\!\perp \epsilon_j$ (beause i precedes j),

$$\begin{aligned}\operatorname{Cov}(X_i, X_j) &= \sum_{l \in pa(j)} \beta_{lj} \operatorname{Cov}(X_i, X_l) \\ &= \sum_{l \in pa(j)} \beta_{lj} \sum_{(k_0,\ldots,k_m) \in \mathcal{D}(i,l)} \prod_{l=1}^{m} \beta_{k_{l-1}k_l} \\ &= \sum_{l \in pa(j)} \sum_{(k_0,\ldots,k_m) \in \mathcal{D}(i,l)} \Bigl(\prod_{l=1}^{m} \beta_{k_{l-1}k_l}\Bigr) \cdot \beta_{lj} \\ &= \sum_{(k_0,\ldots,k_{m+1}) \in \mathcal{D}(i,j)} \prod_{l=1}^{m+1} \beta_{k_{l-1}k_l}.\end{aligned}$$

□

3.15 Exercise. Modify equation (3.3) so that it is still true when the random variables are not standardised. *Hint: How many "forks" $k_{l-1} \leftarrow k_l \rightarrow k_{l+1}$ can a d-connected path have?*

3.4 Correlation and causation

When is correlation the same as causation? Comparing (3.2) with (3.3), we see that is only the case if all the d-connected paths are directed.

The causal effect of X_i on X_j is said to be *confounded* if i and j shares a common ancestor in the graph. In this case, non-zero correlation between X_i and X_j does not imply a causal relationship.

3.16 Example. Consider the graphical model in Figure 3.2.

Applying path analysis and assuming A, X, Y have unit variance, we obtain

$$\begin{aligned}\operatorname{Cov}(A, X) &= \beta_{XA}, \\ \operatorname{Cov}(X, Y) &= \beta_{XY} + \beta_{XA}\beta_{AY}, \\ \operatorname{Cov}(A, Y) &= \beta_{AY} + \beta_{XA}\beta_{XY}.\end{aligned} \tag{3.4}$$

Thus $\operatorname{Cov}(A, Y)$, the coefficient of regressing Y on A, is generally not equal to $\beta(A \rightarrow Y) = \beta_{AY}$.

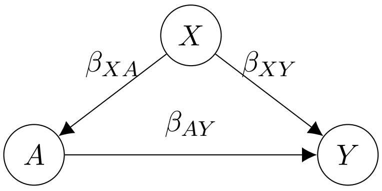


Figure 3.2: X confounds the causal effect of A on Y.

3.17 Example (Continuing Example 3.16)**.** To remove the confounding effect, it is common to add X in the linear regression. Let the coefficient of A in that regression be $\gamma_{AY\cdot X}$. Using least squares theory, this is equal to the partial regression coefficient of $Y - \mathrm{Cov}(X,Y)X$ on $A - \mathrm{Cov}(X,A)X$:

$$\begin{aligned}\gamma_{AY\cdot X} &= \frac{\mathrm{Cov}\left(A - \mathrm{Cov}(X,A)X, Y - \mathrm{Cov}(X,Y)X\right)}{\mathrm{Var}\left(A - \mathrm{Cov}(X,A)X\right)} \\ &= \frac{\mathrm{Cov}(A,Y) - \mathrm{Cov}(X,A)\,\mathrm{Cov}(X,Y)}{1 - \mathrm{Cov}(X,A)^2}.\end{aligned} \tag{3.5}$$

Plug in the expressions in (3.4) into (3.5), we obtain

$$\gamma_{AY\cdot X} = \frac{\beta_{AY} + \beta_{XA}\beta_{XY} - \beta_{XA}(\beta_{XY} + \beta_{XA}\beta_{AY})}{1 - \beta_{XA}^2} = \beta_{AY}.$$

3.18 Remark. The fact that $\gamma_{AY\cdot X} = \beta_{AY}$ in Example 3.16 relies on different assumptions than the regression adjustment discussed in Section 2.5. In regression adjustment, the coefficient of A in the linear regression equals to the average causal effect because A is randomised and $A \perp\!\!\!\perp X$, and the conclusion holds regardless of whether the linear regression correctly specifies $\mathbb{E}[Y \mid A, X]$ (see Remark 2.25). In contrast, two strong assumptions are needed here:

(i) The linear regression correctly specifies $\mathbb{E}[Y \mid A, X]$;

(ii) The linear model (3.8) is structural (see Remark 3.9), so β_{AY} is indeed the causal effect.

Following Example 3.17, a natural question is: in order to identify causal effects by regression coefficients, which variables should be included as regressors ("adjusted for")? We will learn the answer later on in the course but we will examine some negative examples below to gain intuitions.

3.19 Example. Consider the two graphical models in Figure 3.3, in which the random variables are all centred and standardised. In the left diagram, $\beta(A \to Y) = \beta_{AX}\beta_{XY}$ but $\gamma_{AY\cdot X} = 0$. In the right diagram, $\beta(A \to Y) = 0$ but using (3.5),

$$\gamma_{AY\cdot X} = -\frac{\beta_{AX}\beta_{YX}}{1 - \beta_{AX}^2}.$$

This is commonly referred to as *collider bias* because X is a collider in $A \to X \leftarrow Y$.

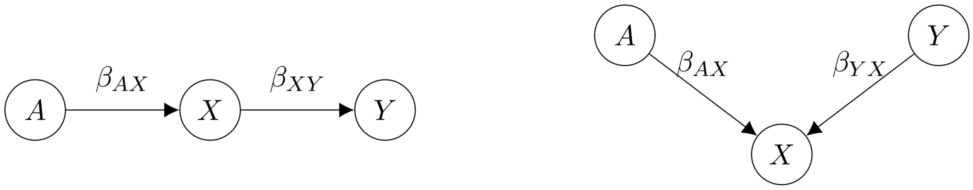


Figure 3.3: Two examples in which adjusting for X in a linear regression introduces bias to estimating the causal effect of A on Y.

An immediate lesson learned from Example 3.19 is that we should not include descendants of the cause in the regression. However, the next Example shows that this is not enough.

3.20 Exercise. In each of the two cases below, give a linear SEMs such that X is not a descendant of A or Y, $\beta(A \to Y) = 0$ but $\gamma_{AY\cdot X} \neq 0$.

(i) There is no d-connected path between A and Y;

(ii) X is on every d-connected path between A and Y.

3.5 Latent variables and identifiability

So far we have assumed that all the variables in the linear SEM are *observed.* A direct consequence is that all the path coefficients are *identifiable* from the distribution of $\boldsymbol{X}$.

3.21 Proposition. *Suppose the random variables $\boldsymbol{X}_{[p]}$ satisfy the linear SEM with respect to a DAG $\mathcal{G}$. Then the path coefficients in $\boldsymbol{B}$ can be written as functions of $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}_{[p]})$.*

Proof. First of all, $\boldsymbol{\Sigma}$ is positive definite (Exercise 3.10), so any principal submatrix of $\boldsymbol{\Sigma}$ is also positive definite. For each variable X_i, the path coefficients from its parents to X_i can be identified by the corresponding linear regression, i.e.,

$$\boldsymbol{\beta}_{pa(X_i),i} = \boldsymbol{\Sigma}^{-1}_{pa(X_i),pa(X_i)} \boldsymbol{\Sigma}_{pa(X_i),i}. \tag{3.6}$$

□

There are at least two reasons to consider SEMs with latent variables (also called *factors*):

(i) *Confounding bias:* Simply ignoring the latent variables (e.g. using the subgraph of $\mathcal{G}$ restricted to the observed variables) lead to biased estimate of the path coefficients. It is thus important to know if we can still identify a causal effect when some variables are unobserved.

(ii) *Proxy measurement:* In many real applications, the variables of interest are not directly measured. This is particularly common in the social sciences where the variable of interest may be socioeconomic status, personality, or political ideology. These variables may only be approximately measured by observable variables (proxies) like human behaviours and questionnaires.

3.22 Example. Excerpt of an educational psychology study (click here).[2]

With latent variables, identifiability of path coefficients no longer follows from Proposition 3.21 because $\boldsymbol{\Sigma}$ is only partially estimable. Path analysis (3.3) allows us to construct a mapping (Exercise 3.10)

$$\boldsymbol{B} \mapsto \boldsymbol{\Sigma}(\boldsymbol{B})$$

between the paths coefficients and the covariance matrix of the observed and unobserved variables.

An entry (or a function) of $\boldsymbol{B}$ is said to be *identifiable* if it can be expressed in terms of the distribution of the observed variables. In linear SEMs with normal errors, this is equivalent to expressing $\boldsymbol{B}$ in terms of the submatrix of $\boldsymbol{\Sigma}$ corresponding to the observed variables (because the multivariate normal distribution is uniquely determined by its mean and covariance matrix).

When the errors are non-normal, we may further use higher moments or the entire distribution of the observed variables to identify $\boldsymbol{\beta}$. However, it is also more sensitive to the distributional assumptions. Below we will restrict our discussion to the case of normal errors.

3.23 Remark. The notion of *identifiability* can depend on the context of the problem. With latent variables, it is often the case that we can only identify some path coefficients up to a sign change. In other problems (such as problems with instrumental variables), the set of nonidentifiable path coefficients has measure zero (this is called *generic identifiability*). We will not differentiate between these concepts in the discussion below.

To identify the entire matrix $\boldsymbol{B}$, a necessary condition is that $\boldsymbol{\Sigma}$ has at least as many entries as $\boldsymbol{B}$. Unfortunately, there is no known necessary and sufficient condition for identifiability in linear SEMs.[3]

3.6 Factor models and measurement models

Below we give some examples in which the path coefficients are indeed identifiable.[4] The basic idea is to use proxies for the latent variables.

Without loss of generality, we assume all the unmeasured variables are standardised so they all have unit variance. In the diagrams below, we will use dashed circles to indicate latent variables.

3.24 Lemma (Three-indicator rule)**.** *Consider any linear SEM for $(U, \boldsymbol{X}_{[p]})$ corresponding to Figure 3.4. Suppose $\boldsymbol{X}_{[p]}$ is observed but U is not. Suppose $\mathrm{Var}(U) = 1$. Then the path coefficients are identifiable (up to a sign change) if $p \geq 3$ and at least 3 coefficients are nonzero.*

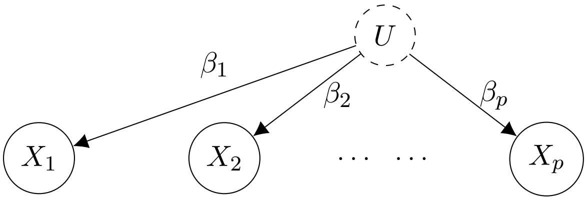


Figure 3.4: Illustration for three-indicator rule ($p \geq 3$).

Proof. Denote the path coefficient for $U_1 \to X_i$ as β_i and the variance of the noise variable for X_i as σ_i^2. It is straightforward to show that

$$\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\beta}\boldsymbol{\beta}^T + \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2).$$

When $p = 3$, this means that

$$\begin{pmatrix} \mathrm{Var}(X_1) & & \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \\ \mathrm{Cov}(X_1, X_3) & \mathrm{Cov}(X_2, X_3) & \mathrm{Var}(X_3) \end{pmatrix} = \begin{pmatrix} \beta_1^2 + \sigma_1^2 & & \\ \beta_1\beta_2 & \beta_2^2 + \sigma_2^2 & \\ \beta_1\beta_3 & \beta_2\beta_3 & \beta_3^2 + \sigma_3^2 \end{pmatrix}.$$

Therefore, we have

$$\beta_1^2 = \frac{\mathrm{Cov}(X_1, X_2) \cdot \mathrm{Cov}(X_1, X_3)}{\mathrm{Cov}(X_2, X_3)},$$

and similarly for β_2^2 and β_3^2. Although the sign of β_1 is not identifiable, it is easy to see that once it is fixed, the signs of β_2 and β_3 are also determined. Thus the vector $\boldsymbol{\beta}$ is identifiable up to the transformation $\boldsymbol{\beta} \mapsto -\boldsymbol{\beta}$.

For $p > 3$, we can apply apply the above result for the 3-subset of $\boldsymbol{X}_{[p]}$ whose corresponding path coefficients are nonzero. □

3.25 Remark. Statistical inference for the graphical model in Figure 3.4 is often called a *confirmatory factor analysis* because the structure is already given. This is different from the *exploratory factor analysis* (e.g., via principal component analysis), which tries to use ovserved data to discover the factor structure.

3.26 Example. For the linear SEM corresponding to the graphical model in Figure 3.5, β_{AY} is identifiable. To see this, we can first use Lemma 3.24 on $\{A, Y, X\}$ and $\{A, Y, Z\}$ to identify (β_{UA}, β_{UY}) (up to a sign change). Without loss of generality we assume A and Y have unit variance, then $\beta_{AY} = \mathrm{Cov}(A, Y) - \beta_{UA}\beta_{UY}$ is also identified.

3.27 Exercise. Show that β_{AY} is non-identifiable if Z is unobserved in Figure 3.5

3.28 Exercise. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ be three random variables/vectors. The *partial covariance* between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ given $\boldsymbol{X}_3$ is defined as

$$\begin{aligned} &\mathrm{PCov}(\boldsymbol{X}_1, \boldsymbol{X}_2 \mid \boldsymbol{X}_3) \\ =&\, \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) - \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_3)\,\mathrm{Var}(\boldsymbol{X}_3)^{-1}\,\mathrm{Cov}(\boldsymbol{X}_3, \boldsymbol{X}_2). \end{aligned}$$

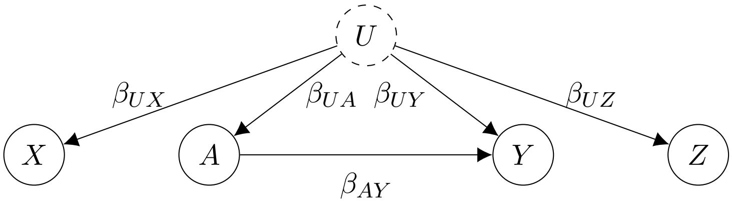


Figure 3.5: Illustration of using proxies of unmeasured confounders to remove unmeasured confounding bias.

Show that if we add a directed edge from X to A in Figure 3.5, β_{AY} is still identifiable by[5]

$$\beta_{AY} = \mathrm{Cov}(A, Y) - \frac{\mathrm{PCov}(X, Y \mid A)}{\mathrm{PCov}(X, Z \mid A)} \mathrm{Cov}(A, Z).$$

The three-indicator rule is also quite useful in the so-called *measurement models.* In this type of problems (an instance is Example 3.22), we are indeed interested in the causal effects between the latent variables (these are often abstract constructs like personalities and academic achievements).

Suppose the latent variables $\boldsymbol{U} \in \mathbb{R}^q$ have unit variances and satisfy a linear SEM with respect to a prespecified DAG. The observed variables (or *measurements*) $\boldsymbol{X} \in \mathbb{R}^p$ satisfy the following model (the intercept term is omitted for simplicity)

$$\boldsymbol{X} = \boldsymbol{\Gamma} \boldsymbol{U} + \boldsymbol{\epsilon}_{\boldsymbol{X}},$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times q}$ is the factor loading matrix, $\boldsymbol{\epsilon}_{\boldsymbol{X}}$ is a vector of mutually independent mean-zero noise variables and $\boldsymbol{\epsilon}_X \perp\!\!\!\perp \boldsymbol{U}$.

3.29 Proposition. *Suppose $(\boldsymbol{U}, \boldsymbol{X})$ satisfy the measurement model described above. The path coefficients between the latent variables $\boldsymbol{U}$ are identifiable (up to sign change of $\boldsymbol{U}$) if the following conditions are satisfied:*

(i) Each row of $\boldsymbol{\Gamma}$ has only 1 nonzero entry (i.e., every measurement loads on only one factor).

(ii) Each column of $\boldsymbol{\Gamma}$ has at least 3 nonzero entries (i.e., each factor has at least three measurements).

Proof. By Lemma 3.24, $\boldsymbol{\Gamma}$ can be identified. The assumptions in the proposition statement also ensures that $\boldsymbol{\Gamma}$ has full column rank, so $\mathrm{Cov}(\boldsymbol{U})$ can be identified from $\mathrm{Cov}(\boldsymbol{X})$. The conclusion then follows from Proposition 3.21. □

3.30 Example. The graphical model in Figure 3.6 satisfies the criterion in Proposition 3.29, thus β_U is identifiable (up to its sign). To see this, $\beta_{11}, \beta_{12}, \ldots, \beta_{26}$ can be identified by confirmatory factor analysis, and by using path analysis, we have

$$\mathrm{Cov}(X_1, X_4) = \beta_{11} \beta_U \beta_{24}.$$

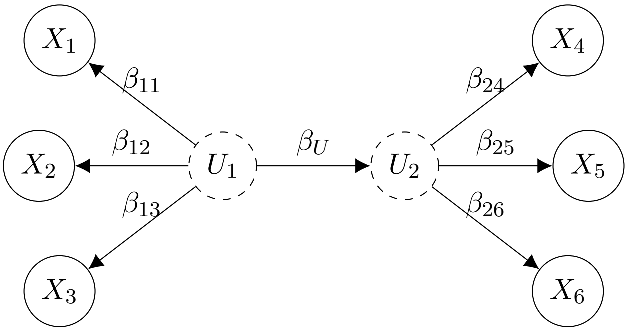


Figure 3.6: Example of a measurement model.

3.31 Exercise. Show that β_U in the last example is still identifiable if each latent variable only has two measurements (i.e. if X_3 and X_6 are deleted from the graph).

3.32 Remark. Although the path coefficients between $\boldsymbol{U}$ can only be identified up to sign changes, this is usually not a problem in practice. Usually we can confidently make assumptions about the signs of certain factor loadings (for example, the loading of a student's maths score on academic achievements is positive).

3.7 Estimation in linear SEMs

Let $\boldsymbol{X} \in \mathbb{R}^p$ be the observed variables in a linear SEM. Let $\boldsymbol{B}$ denote the matrix of path coefficients between all the variables, observed or latent. Suppose $\boldsymbol{B}$ is indeed identifiable.

There are two main approaches to fit a linear SEM and estimate $\boldsymbol{B}$: maximum likelihood and generalised method of moments.

By assuming the noise variable $\boldsymbol{\epsilon}_{[p]}$ in (3.1) follows a multivariate normal distribution, the maximum likelihood estimator of $\boldsymbol{B}$ minimises

$$l(\boldsymbol{B}) = \frac{1}{2} \log\det\Big(\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{B})\Big) + \frac{1}{2}\mathrm{tr}\Big(\boldsymbol{S}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}(\boldsymbol{B})\Big), \tag{3.7}$$

where $\boldsymbol{S}$ is the sample covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{\Sigma}_{\boldsymbol{X}}(\boldsymbol{B})$ is the covariance matrix of $\boldsymbol{X}$ and depends on the path coefficients $\boldsymbol{B}$ through (3.3).

3.33 Exercise. Derive (3.7).

Generalised method of moments (an extension of Z-estimation) tries to directly match the theoretical covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{B})$ with the sample covariance matrix $\boldsymbol{S}$ by minimising

$$l_{\boldsymbol{W}}(\boldsymbol{B}) = \frac{1}{2}\mathrm{tr}\Big(\big\{\big[\boldsymbol{S} - \boldsymbol{\Sigma}(\boldsymbol{B})\big]\boldsymbol{W}^{-1}\big\}^2\Big), \tag{3.8}$$

where $\boldsymbol{W}$ is a $p \times p$ positive definite weighting matrix. This is also called the *generalised least squares* estimator in the SEM literature.

Different choices of $\boldsymbol{W}$ lead to estimators with different asymptotic efficiency. The "optimal" choice is $\boldsymbol{W} = \boldsymbol{\Sigma}(\boldsymbol{B})$ (or any other matrix that converges in probability to $\boldsymbol{\Sigma}(\boldsymbol{B})$). This motivates the practical choice $\boldsymbol{W} = \boldsymbol{S}$, so we estimate $\boldsymbol{B}$ by minimising

$$l_{\boldsymbol{S}}(\boldsymbol{B}) = \frac{1}{2}\mathrm{tr}\Big(\big[\boldsymbol{I} - \boldsymbol{S}^{-1}\boldsymbol{\Sigma}(\boldsymbol{B})\big]^2\Big). \tag{3.9}$$

The generalised method of moments estimator is consistent if the linear SEM is correctly specified (so $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}(\boldsymbol{B})$). Furthermore, if $l_{\boldsymbol{S}}(\boldsymbol{B})$ is used and $\boldsymbol{\epsilon}$ is normally distributed, the estimator is asymptotically equivalent to the maximum likelihood estimator and is thus asymptotically efficient.[6].

3.8 Strengths and weaknesses of linear SEMs

Despite being a century old, linear SEMs are still widely used in many applications for many good reasons:

(i) Graphs and linear SEMs provide an intuitive way to rigorously describe causality that can also be easily understood by applied researchers.

(ii) Path analysis provides a powerful tool to distinguish correlation from causation. Even though we will move away from linearity soon, path analysis provides a straightforward way to disprove some statements and gain intuitions for others[7].

(iii) Linear SEMs allow us to directly put models on unobserved variables. This is especially useful when the causes and effects of interest are abstract constructs.

(iv) Fitting a linear SEM only requires the sample covariance matrix, which can be handy in modern applications with privacy constraints.

Linear SEMs also have important limitations:

(i) The causal structure needs to be known a priori.

(ii) The linear model can be misspecified and does not handle binary variables or discrete variables very well. This is problematic because the causal effect is not well defined if the model is nonlinear. As a consequence, the meaning of structural equation models became obscure and lead many to believe they are just the same as linear regression. This misconception led many researchers to rejected linear SEMs as a tool for causal inference.[8]

(iii) Any model put on the unobserved variables is dangerous, because there is no realistic way to verify those assumptions.

Notes

[1] Wright, S. (1918). On the nature of size factors. *Genetics*, *3*(4), 367–374; Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161–215.

[2] Marsh, H. W. (1990). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology*, *82*(4), 646–656. doi:10.1037/0022-0663.82.4.646.

[3] For a review on recent advances using algebraic geometry, see Drton, M. et al. (2018). Algebraic problems in structural equation modeling. In *The 50th Anniversary of Gröbner Bases* (pp. 35–86). Mathematical Society of Japan.

[4] More discussion and examples can be found at Bollen, K. A. (1989). *Structural equations with latent variables.* doi:10.1002/9781118619179, page 326.

[5] Kuroki, M., & Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, *101*(2), 423–437. doi:10.1093/biomet/ast066.

[6] For more detail, see Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. doi:10.1111/j.2044-8317.1984.tb00789.x.

[7] Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, *1*(1), 155–170. doi:10.1515/jci-2013-0003.

[8] For a historical account, see Pearl, 2009, Section 5.1.

Chapter 4

Graphical models

The linear SEMs are intuitive and easy to interpret, but become inadequate when the structural relations are non-linear. Intuitively, causality should be already entailed in the graphical diagram, and linearity should be unnecessary for causal inference.

To move away from the linearity assumption, we will introduce graphical models for the observed variables in this Chapter and for the unobserved counterfactuals in the next Chapter. The main idea in this Chapter is to describe conditional independences with graphs.

4.1 Markov properties for undirected graphs

Briefly speaking, a graphical model provides a concise representation of all the conditional independence relations (aka Markov properties) between random variables. We will start from undirected graphs.

Let $\mathcal{G} = (V = [p], E)$ be an undirected graphical model for the random variables $\boldsymbol{X}_{[p]} = (X_1, X_2, \ldots, X_p)$. Edges in an undirected graph have no direction. In other words, if $(i, j) \in E$, so does (j, i).

4.1 Definition (Separation in undirected graph)**.** For any disjoint $I, J, K \subset V$, K is said to *separate* I and J in an undirected graph $\mathcal{G}$, denoted as $I \perp\!\!\!\perp J \mid K \; [\mathcal{G}]$, if every path from a node in I to a node in J contains a node in K.

4.2 Definition (Global Markov property)**.** A probability distribution $\mathbb{P}$ is said to satisfy the *global Markov property* with respect to the graph $\mathcal{G}$ if $I \perp\!\!\!\perp J \mid K \; [\mathcal{G}] \Longrightarrow \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K$ for any disjoint $I, J, K \subset V$.

The global Markov property with respect to a graph is closely related to the factorisation of a probability distribution.

4.3 Definition (Factorisation according to an undirected graph)**.** A *clique* in an undirected graph $\mathcal{G}$ is a subset of vertices such that every two distinct vertices in the clique are adjacent.

A probability distribution $\mathbb{P}$ is said to *factorise* according to $\mathcal{G}$ (or a Gibbs random field with respect to $\mathcal{G}$) if $\mathbb{P}$ has a density f that can be written as

$$f(\boldsymbol{x}) = \prod_{\text{clique } C \subseteq V} \psi_C(\boldsymbol{x}_C),$$

for some functions ψ_C, $C \subset V$.

4.4 Theorem (Hammersley-Clifford)**.** *Suppose the probability distribution $\mathbb{P}$ has a positive density function. Then $\mathbb{P}$ satisfies the global Markov property with respect to $\mathcal{G}$ if and only if it factorises according to $\mathcal{G}$.*

Proof. We will only prove the $\Longleftarrow$ direction here.[1]. Let I, J, K be disjoint subsets of V such that $I \perp\!\!\!\perp J \mid K$. A consequence of I and J being separated by K is that I and J must be in different connected components of the subgraph $\mathcal{G}_{V \setminus K}$. (A set of vertices is called connected if there exists a path from any vertex to any other vertex in this set. A connected component is a maximal connected set, meaning no superset of the connected component is connected.)

Let $\tilde{I}$ denote the connected component that I is in and let $\tilde{J} = V \setminus (\tilde{I} \cup K)$. Any clique of $\mathcal{G}$ must either be a subset of $\tilde{I} \cup K$ or of $\tilde{J} \cup K$, otherwise at least one vertex in $\tilde{I}$ is adjacent to a vertex in $\tilde{J}$, violating the maximality of $\tilde{I}$. This implies that

$$\begin{aligned} f(\boldsymbol{x}) &= \prod_{\text{clique } C \subseteq V} \psi_C(\boldsymbol{x}_C) \\ &= \prod_{\text{clique } C \subseteq \tilde{I} \cup K} \psi_C(\boldsymbol{x}_C) \cdot \prod_{\text{clique } C \subseteq \tilde{J} \cup K} \psi_C(\boldsymbol{x}_C) / \prod_{\text{clique} K} \psi_C(\boldsymbol{x}_C) \\ &= h(\boldsymbol{x}_{\tilde{I} \cup K}) \cdot g(\boldsymbol{x}_{\tilde{J} \cup K}). \end{aligned}$$

We have shown that $f(\boldsymbol{x})$ can be written as a function of $\boldsymbol{x}_{\tilde{I} \cup K}$ multiplied by another function $\boldsymbol{x}_{\tilde{J} \cup K}$. By normalising the functions properly, this shows that $\boldsymbol{X}_{\tilde{I}} \perp\!\!\!\perp \boldsymbol{X}_{\tilde{J}} \mid \boldsymbol{X}_K$ and hence $\boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K$. □

Notice that in the proof above we did not use the positive density assumption. It is only needed for the $\Longrightarrow$ direction.

4.2 Markov properties for directed graphs

We now move to DAG models.

4.5 Definition. We say a probability distribution $\mathbb{P}$ *factorises according to a DAG* $\mathcal{G}$ if its density function f satisfies

$$f(\boldsymbol{x}) = \prod_{i \in V} f_{i|pa(i)}(x_i \mid \boldsymbol{x}_{pa(i)}),$$

where $f_{i|pa(i)}(x_i \mid \boldsymbol{x}_{pa(i)})$ is the conditional density of X_i given $\boldsymbol{X}_{pa(i)}$, that is,

$$f_{i|pa(i)}(x_i \mid \boldsymbol{x}_{pa(i)}) = \frac{f_{\{i\} \cup pa(i)}(x_i, \boldsymbol{x}_{pa(i)})}{f_{pa(i)}(\boldsymbol{x}_{pa(i)})},$$

where $f_I(\boldsymbol{x}_I)$ is the marginal density function for $\boldsymbol{X}_I$.

Below we will often suppress the subscript and use f as a generic symbol to indicate a density or conditional density function.

4.6 Example. A probability distribution factorises according to Figure 4.1 if its density can be written as

$$f(\boldsymbol{x}) = f(x_1)f(x_2 \mid x_1)f(x_3 \mid x_1)f(x_4 \mid x_2, x_3)f(x_5 \mid x_2)f(x_6 \mid x_3, x_4)f(x_7 \mid x_4, x_5, x_6).$$

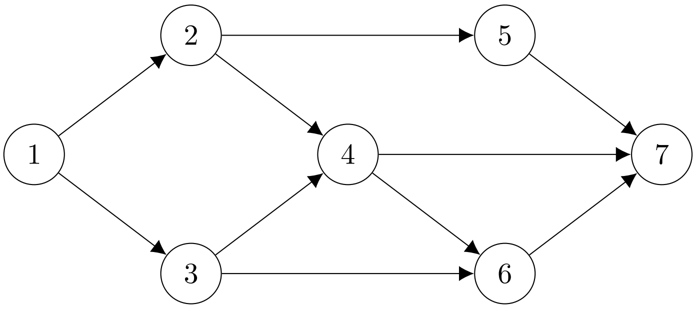


Figure 4.1: Example of a DAG.

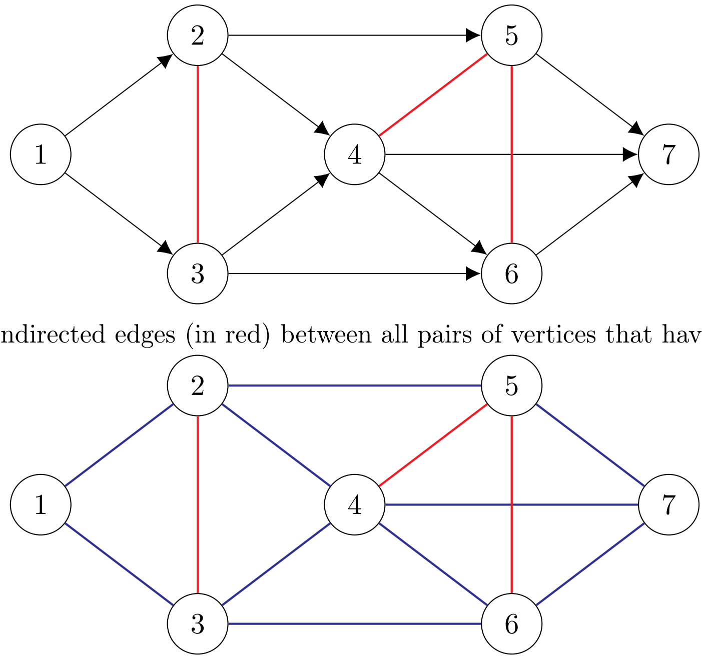
4.7 Example. Probabilitistic graphical models are widely used in Bayesian statistics, machine learning, and engineering.[2] Besides its intuitive representation, another motivation for using graphical models is more efficient storage of probability distributions. Consider p binary random variables. A naive way that stores the entire table of their joint distribution needs to record 2^p probability mass values. In contrast, suppose the random variables factorise according to a DAG $\mathcal{G}$ in which each vertex has no more than d parents. Then it is sufficient to store $p \cdot 2^d$ values.

It is obvious that if $\boldsymbol{X}_{[p]}$ satisfies the linear SEM according to $\mathcal{G}$ (Definition 3.8), then the distribution of $\boldsymbol{X}_{[p]}$ also factorises according to $\mathcal{G}$. Therefore, we can often use linear SEMs to understand properties of DAG models (see, e.g., Exercise 4.16 below). However, linear SEM further makes assumptions about the interventional distributions of $\boldsymbol{X}_{[p]}$. On the contrary, a DAG model only restricts the observational distribution of $\boldsymbol{X}_{[p]}$ like a regression model (see Remark 3.9 for a comparison of linear SEM with linear regression). The next Chapter will introduce DAG models for counterfactuals.

4.8 Definition. Given a DAG $\mathcal{G}$, its undirected *moral graph* $\mathcal{G}^m$ is obtained by first adding undirected edges between all pairs of vertices that have a common child and then erasing the direction of all the directed edges.

This graph is called "moral" because we are marrying all the parents that have a common child.

4.9 Example. Figure 4.2 illustrates the moralisation of the DAG in Figure 4.1. First, three undirected edges (2,3), (4,5), (5,6) are added because they share a common child. Second, the directions of all the edges in the original graph are erased.



(a) Step 1: Add undirected edges (in red) between all pairs of vertices that have a common child.

(b) Step 2: Remove edge directions (in blue).

Figure 4.2: Illustration of moralising the DAG in Figure 4.1.

For any $i \in V$, the subgraph of $\mathcal{G}^m$ restricted to $\{i\} \cup pa(i)$ is a clique. Thus by Definitions 4.3 and 4.5 and Theorem 4.4, we immediately have

4.10 Lemma. *If a probability distribution $\mathbb{P}$ factorises according to a DAG $\mathcal{G}$, it also factorises according to its moral graph $\mathcal{G}^m$ and thus satisfies the global Markov property with respect to $\mathcal{G}^m$.*

Lemma 4.10 gives us a way to obtain conditional independence relations for distributions that factorises according to a DAG (using Definition 4.1).

4.11 Corollary. *Suppose $\mathbb{P}$ factorises according to a DAG $\mathcal{G}$, then*

$$I \perp\!\!\!\perp J \mid K\ [\mathcal{G}^m] \Longrightarrow \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K.$$

This criterion can be improved. Let $\overline{an}(I) = an(I) \cup I$. The next Proposition says that we only need to moralise the subgraph of $\mathcal{G}$ restricted to $\overline{an}(I \cup J \cup K)$.

4.12 Proposition. *Suppose $\mathbb{P}$ factorises according to a DAG $\mathcal{G}$, then*

$$I \perp\!\!\!\perp J \mid K\ \left[(\mathcal{G}_{\overline{an}(I\cup J\cup K)})^m\right] \Longrightarrow \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K.$$

Proof. It is easy to verify that for any $I \subseteq V$, $i \in \overline{an}(I)$ implies that $pa(i) \subseteq \overline{an}(I)$. By Definition 4.5, this implies that the marginal distribution of $\boldsymbol{X}_{\overline{an}(I\cup J\cup K)}$ must factorise according to the subgraph $\mathcal{G}_{\overline{an}(I\cup J\cup K)}$. The proposition then immediately follows from Corollary 4.11. □

4.13 Example. Suppose the distribution $\mathbb{P}$ of $\boldsymbol{X}$ factorises according to the the DAG in Figure 4.1. We can use the criterion in Proposition 4.12 but not the one in Corollary 4.11 to derive $X_4 \perp\!\!\!\perp X_5 \mid \{X_2, X_3\}$.

4.14 Exercise. Suppose the distribution $\mathbb{P}$ of $\boldsymbol{X}$ factorises according to the the DAG in Figure 4.1. Which one(s) of the following conditionally independent relationships can be derived from Proposition 4.12?

(i) $X_2 \perp\!\!\!\perp X_6 \mid X_4$;

(ii) $X_2 \perp\!\!\!\perp X_6 \mid X_3$;

(iii) $X_2 \perp\!\!\!\perp X_7 \mid \{X_4, X_5\}$;

(iv) $X_5 \perp\!\!\!\perp X_6 \mid X_4$;

(v) $X_5 \perp\!\!\!\perp X_6 \mid \{X_3, X_4\}$;

Next we give another criterion called *d-separation* that only uses the original DAG $\mathcal{G}$ and thus is much easier to apply. To gain some intuition, consider the following example.

4.15 Example. There are three possible situations for a DAG with 3 vertices and 2 edges (Figure 4.3). Using Corollary 4.11, it is easy to show that $X_1 \perp\!\!\!\perp X_3 \mid X_2$ is true in the first two cases. However, in the third case, even though X_1 and X_3 are marginally independent, conditioning on the *collider* X_2 (common child of X_1 and X_3) actually introduces dependence, so $X_1 \perp\!\!\!\perp X_3 \mid X_2$ is not true in general. Example 3.19 showed the same phenomenon using the more restrictive linear SEM interpretation of these graphical models.

4.16 Exercise. (i) By directly using the DAG factorisation (without using moralisation), show that $X_1 \perp\!\!\!\perp X_3 \mid X_2$ is true in the first two cases but generally false for the third case in Figure 4.3.

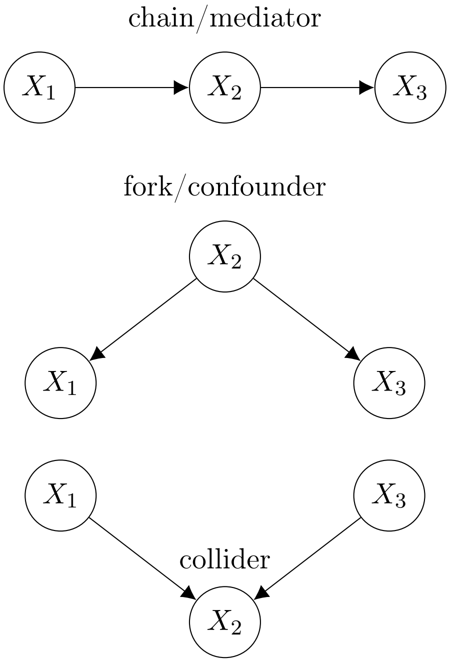


Figure 4.3: Possible DAGs with 3 vertices and 2 edges.

(ii) Alternatively, by assuming (X_1, X_2, X_3) satisfies a linear SEM with respect to the corresponding graph, demonstrate why $X_1 \perp\!\!\!\perp X_3 \mid X_2$ holds or does not hold. For simplicity, you may assume all the path coefficients are equal to 1.

(iii) What happens if there is an additional vertex X_4 that is a child of X_2 and has no other parent, and we condition on X_4 instead of X_2?

4.17 Definition. Given a DAG $\mathcal{G}$, a path is said to be *blocked* by $K \subseteq V$ if there exists a vertex k on the path such that either

(i) k is not a collider on this path and $k \in K$; or

(ii) k is a collider on this path and k and all its descendants are not in K;

For disjoint subsets of nodes $I, J, K \subset V$, we say I and J are *d-separated* by K, written as $I \perp\!\!\!\perp J \mid K$ $[\mathcal{G}]$, if all paths from a vertex in I to a vertex in J are blocked by K.

4.18 Example. For the DAG in Figure 4.1, $K = \{1\}$ blocks the paths $(3, 1, 2, 5)$, $(3, 4, 2, 5)$, $(3, 6, 4, 2, 5)$, $(3, 6, 7, 4, 2, 5)$, and $(3, 6, 7, 5)$. Therefore the nodes 3 and 5 are d-separated by 1.

4.19 Remark. To memorise the definition of d-separation, imagine water flowing along the edges and each vertex acts as a valve. A collider valve is naturally “off”, meaning there is no flow of water from one side of the collider to the other side. A non-collider valve is naturally "on", allowing water to flow freely. Now imagine we can turn on or off the valves (the perhaps non-intuitive part is that turning on any descendant of a collider also turns on that collider). Water can flow from one end of the path to the other end (path is unblocked) if and only if all the valves on the path are "on".

In path analysis (Definition 3.13), we have already seen that a d-connected path can induce dependence between variables. The induced dependence can be removed ("blocked") by conditioning on any non-collider on the path. Conversely, although a closed (not d-connected) path does not induce dependence, it can do so if we condition on all the colliders.

4.20 Lemma. *Consider a DAG $\mathcal{G} = (V, E)$ and disjoint $I, J, K \subset V$. Then*

$$I \perp\!\!\!\perp J \mid K \ \left[\left(\mathcal{G}_{\overline{an}(I \cup J \cup K)}\right)^m\right] \iff I \perp\!\!\!\perp J \mid K \ [\mathcal{G}].$$

The proof of this Lemma is a bit technical and is deferred to Section 4.A.1 (so is non-examinable).

4.21 Theorem. *The distribution $\mathbb{P}$ of $\boldsymbol{X}_{[p]}$ factorises according to a DAG $\mathcal{G}$ if and only if*

$$I \perp\!\!\!\perp J \mid K \ [\mathcal{G}] \Longrightarrow \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K, \ \forall \text{ disjoint } I, J, K \subset V. \tag{4.1}$$

Proof. Proposition 4.12 and Lemma 4.20 immediately imply the $\Longrightarrow$ direction. The $\Longleftarrow$ direction can be shown by induction on $|V|$. Without loss of generality let's assume $V = [p]$ and $(1, 2, \ldots, p)$ is a topological ordering of $\mathcal{G}$, so $(i, j) \in E$ implies that $i < j$. Because the vertex p has no child, it is easy to see that

$$p \perp\!\!\!\perp V \setminus \{p\} \setminus pa(p) \mid pa(p) \ [\mathcal{G}].$$

By (4.1), X_p is independent of the other variables given $\boldsymbol{X}_{pa(p)}$. Thus the joint density of $\boldsymbol{X}_{[p]}$ can be written as

$$f(\boldsymbol{x}_{[p]}) = f(\boldsymbol{x}_{[p-1]}) \cdot f(x_p \mid \boldsymbol{x}_{[p-1]}) = f(\boldsymbol{x}_{[p-1]}) \cdot f(x_p \mid \boldsymbol{x}_{pa(p)}).$$

Using the induction hypothesis for the first term on the right hand side, we thus conclude that $\mathbb{P}$ also factorises according to $\mathcal{G}$ when $|V| = p$. □

Distributions $\mathbb{P}$ satisfying (4.1) are said to satisfy the *global Markov property* with respect to the DAG $\mathcal{G}$. In the last section, Theorem 4.4 establishes the equivalence between global Markov property and factorisation in undirected graphs. Theorem 4.21 extends this equivalence to DAGs, with a small distinction that $\mathbb{P}$ is no longer required to have a positive density function.

4.22 Exercise. Apply the d-separation criterion in Theorem 4.21 to the examples in Exercise 4.14.

4.23 Remark (Completeness of d-separation). The criterion (4.1) cannot be further improved in the following sense. Given any DAG $\mathcal{G}$, it can be shown that there exists a probability distribution $\mathbb{P}^3$ such that

$$I \perp\!\!\!\perp J \mid K \; [\mathcal{G}] \Longleftrightarrow \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K, \; \forall \text{ disjoint } I, J, K \subset V. \tag{4.2}$$

Furthermore, it can be shown that if $\boldsymbol{X}_{[p]}$ is discrete, the set of probability distributions that factorise according to $\mathcal{G}$ but do not satisfy (4.2) has Lebesgue measure zero.[4]

4.24 Example. Consider the setting in Example 3.16 where three random variables (X, A, Y) satisfy a linear SEM (3.4) corresponding to the graph in Figure 3.2. Suppose the structural noise variables are jointly normal and the random variables are standardised so $\mathrm{Var}(X) = \mathrm{Var}(A) = \mathrm{Var}(Y) = 1$. For most values of $\beta_{XA}, \beta_{XY}, \beta_{AY}$, the variables X_1, X_2, X_3 are unconditionally and conditionally dependent. However, very occasionally the distribution of (X, A, Y) may have some "coincidental" independence relations. For example, $A \perp\!\!\!\perp Y$ if the path coefficients happen to satisfy $\beta_{AY} + \beta_{XA}\beta_{XY} = 0$. It is easy to see that this event has Lebesgue measure 0.

4.3 Structure discovery

In structure discovery, the goal is to use conditional independence in the observed data to infer the graphical model, that is, to invert (4.1). Remark 4.23 suggests that this is possible for almost all distributions, which is formalised below.

4.25 Definition. We say a distribution $\mathbb{P}$ of $\boldsymbol{X}$ that factorises according to $\mathcal{G}$ is *faithful* to $\mathcal{G}$ if $I \perp\!\!\!\perp J \mid K \; [\mathcal{G}] \Longleftrightarrow \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K$ for all disjoint $I, J, K \subset V$.

Given that $\mathbb{P}$ is faithful to the unknown DAG $\mathcal{G}$, we can obtain all d-separation relations in $\mathcal{G}$ from the conditional independence relations in $\mathbb{P}$. Without faithfulness, we cannot even exclude the possibility that the underlying DAG is complete.

However, this may not be enough to recover $\mathcal{G}$. A simple counter-example is the two DAGs $X_1 \to X_2$ and $X_2 \to X_1$, both implying $X_1 \not\perp\!\!\!\perp X_2$.

4.26 Definition. Two DAGs are called *Markov equivalent* if they contain the same d-separation relations. A *Markov equivalence class* is the maximal set of Markov equivalent DAGs.

Without additional assumptions, we can only recover the Markov equivalence class that contains in $\mathcal{G}$. The next Theorem gives a complete characterisation of a Markov equivalence class.

4.27 Theorem. *Two DAGs are Markov equivalent if and only if the next two properties are satisfied:*

(i) They have the same "skeleton" (set of edges ignoring the directions);

(ii) They have the same "immoralities" (structures like $i \to k \leftarrow j$ where i and j are not adjacent).

We will only show the $\Longrightarrow$ direction here by proving two Lemmas. The proof for the $\Longleftarrow$ direction can be found in Section 4.A.2.

4.28 Lemma. *Given a DAG $\mathcal{G} = (V, E)$, two vertices $i, j \in V$ are adjacent if and only if they cannot be d-separated by any set $D \subseteq V \setminus \{i, j\}$; otherwise they can be d-separated by $pa(i)$ or $pa(j)$.*

Proof. $\Longrightarrow$ is obvious because no set can block the edge between i and j. For the $\Longleftarrow$ direction, because $i \in an(j)$ and $j \in an(i)$ cannot both be true (otherwise creating a cycle), without loss of generality we assume $j \notin an(i)$. Any path connecting i and j cannot be a directed path from j to i. Consider the directed edge on this path with j as one end. If this edge points into j, this path contains a parent of j that is not a collider; otherwise it must contain a collider that is a descendant of j. In either case the path is blocked by $pa(j)$. □

4.29 Lemma. *For any undirected path $i - k - j$ in $\mathcal{G}$ such that i and j are not adjacent, k is a collider if and only if i and j are not d-separated by any set containing k.*

Proof. This immediately follows from the fact that the path $i - k - j$ is blocked by k if and only if k is not a collider. □

The IC[5] or SGS[6] algorithm uses the conditions in Lemmas 4.28 and 4.29 to recover the Markov equivalence class:

Step 0 Start with an undirected complete graph in which all vertices are adjacent.

Step 1 For every two vertices i, j, remove the edge between i and j if $X_i \perp\!\!\!\perp X_j \mid X_K$ for some $K \subseteq V \setminus \{i, j\}$. This gives us the skeleton of the graph (Lemma 4.28).

Step 2 For every undirected path $i - k - j$ such that i and j are not adjacent in the skeleton obtained in Step 1, orient the edges as $i \to k \leftarrow j$ if $X_i \not\!\perp\!\!\!\perp X_j \mid X_K$ for all $K \subseteq V \setminus \{i, j\}$ containing k (Lemma 4.29).

Step 3 Orient some of the other edges so that the graph contains no cycle or a new immorality would be introduced if the edge is oriented the other way. (In general it is impossible to orient all the edges unless the Markov equivalence class is a singleton.)

The PC algorithm[7] accelerates the Step 1 above using the following trick: to test whether i and j can be d-separated, one only needs to go through subsets of the neighbours of i and subsets of the neighbours of j. The PC algorithm also imposes an order: it starts with $K = \emptyset$ and then gradually increases size of K. For sparse graphs, these two tricks allow us to not only test fewer conditional independences for each pair but also stop the algorithm at a much smaller size for K.

4.30 Exercise. Use the IC/SGS algorithm to derive the Markov equivalence class containing the DAG in Figure 4.1. More specifically, give the conditional independence and dependence relations you used in Steps 1 and 2 of the algorithm. How many DAGs are there in this Markov equivalence class?

4.4 Discussion: Using DAGs to represent causality

It may not be surprising that many people have been fascinated about the prospects of the graphical approach to causality:

(i) Representating causality by directed edges is naturally appealing.

(ii) The theory of probabilistic graphical models is elegant and powerful.

(iii) The possibility of discovering (possibly causal) structures from observational data is exciting.

But structure learning also has important limitations:

(i) Strucutre learning algorithms are computationally intensive;

(ii) Testing conditional independence is known to be a very difficult statistical problem.[8]

(iii) DAGs do not necessarily represent causality: graphical models can just be viewed as a useful tool to describe a probability distribution.

(iv) Additional assumptions like the *causal Markov condition* needed to define causal DAGs are not as transparent as assumptions on counterfactuals or structural noises.[9]

Notes

[1] Proof of the other direction can be found, for example, in Lauritzen, S. L. (1996). *Graphical models.* Clarendon Press, page 36.

[2] See, for example, Wainwright, M. J., & Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1-2), 1–305. doi:10.1561/2200000001.

[3] Geiger, D., & Pearl, J. (1990). On the logic of causal models. R. D. Shachter, T. S. Levitt, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (Vol. 9, pp. 3–14). Machine Intelligence and Pattern Recognition. doi:10.1016/B978-0-444-88650-7.50006-8.

[4] Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 411–418). Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc.

[5] Pearl, J., & Verma, T. S. (1991). A theory of inferred causation. J. Allen, R. Fikes, & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning: Proceedings of the second international conference* (pp. 441–452). Morgan Kaufmann.

[6] Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* doi:10.1007/978-1-4612-2748-9.

[7] Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). doi:10.7551/mitpress/1754.001.0001.

[8] Shah, R. D., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics*, *48*(3), 1514–1538. doi:10.1214/19-aos1857.

[9] Dawid, A. P. (2010). Beware of the DAG!. I. Guyon, D. Janzing, & B. Schölkopf (Eds.), (pp. 59–86). Proceedings of Machine Learning Research. Whistler, Canada: PMLR.

4.A Graphical proofs (non-examinable)

4.A.1 Proof of Lemma 4.20

Suppose K does not d-separate I from J in $\mathcal{G}$, so there exists a path from I to J not blocked by K. The vertices on this path must be contained in $\overline{an}(I \cup J \cup K)$, which can be shown by considering the following two cases:

(i) If this unblocked path contains no collider, then every vertex on this path, if not already in $I \cup J$, must be an ancestor of $I \cup J$.

(ii) If this unblocked path contains at least one collider, then all the colliders must be or have a descendant that is in K. Thus, all the vertices on this path must be in K or ancestors of K.

In case (i), this path cannot contain a vertex in K (because it is unblocked in $\mathcal{G}$), so is unblocked by K in $(\mathcal{G}_{\overline{an}(I\cup J\cup K)})^m$. In case (ii), the path in $(\mathcal{G}_{\overline{an}(I\cup J\cup K)})^m$ that marries the parents of all the colliders is not separated by K (the parents of a collider cannot be colliders and thus do no belong to K). In both cases, K does not separate I from J in the moral graph.

Next we consider the other direction. Suppose I is not separated from J by K in $(\mathcal{G}_{\overline{an}(I\cup J\cup K)})^m$, so there exists a path from a vertex in I to a vertex in J circumventing K in $(\mathcal{G}_{\overline{an}(I\cup J\cup K)})^m$. Edges in the moral graph are either already in $\mathcal{G}$ or added during the "marriage". For each edge added because of a "marriage" by a collider, extend the path to include that collider. This results in a path in $\mathcal{G}$. The set K does not block this path at the non-colliders because the original path in the moral graph is not separated by K.

Consider the subsequence of this path, say from $i \in I$ to $j \in J$, that does not contain any intermediate vertex in $I \cup J$. Consider any collider in this sub-path (let's call it l) that does not belong to $\overline{an}(K)$, so $l \in an(I \cup J)$; without loss of generality assume $l \in an(I)$. By definition, there exists a directed path in $\mathcal{G}$ from l to i' for some $i' \in I$. Consider a new path, tracing back from i' to k and then joining the original path from k to j (see Figure 4.4 for an illustration). Because $l \notin \overline{an}(K)$, the new part of the path from i to l is not blocked by K. Thus we have obtained a path in $\mathcal{G}$ from I to J, unblocked by K at non-colliders, with one fewer collider than the original. Repeating the argument in this paragraph until we end up with a path from I to J whose colliders are in $\overline{an}(K)$ and non-colliders are not in K. By Definition 4.17, this path is not blocked by K.

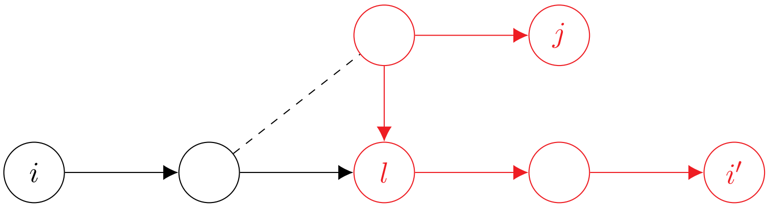


Figure 4.4: Illustration for obtaining a new path with fewer colliders (in red). Dashed line indicates an edge added during the marriage.

4.A.2 Additional proofs for Markov equivalent DAGs

The $\Longleftarrow$ direction of Theorem 4.27 is established through the next two Lemmas.

4.31 Lemma. *Among all paths from i to j that is unblocked by $K \subset V$ in a DAG $\mathcal{G} = (V, E)$, consider the shortest one (if there are several, consider any one of them). Moreover, suppose this path contains a $k - l - m$ such that k, m are adjacent. Then the edges must be oriented as $k \leftarrow l \rightarrow m$, and at least one of k, m is an collider in this path (if $k \rightarrow m$ then k is a collider; if $k \leftarrow m$ then m is a collider). As a corollary, all colliders in this path must be "immoral".*

4.32 Lemma. *Suppose two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeleton and immoralities. If there is a path from i to j unblocked by $K \subset V$ in $\mathcal{G}_1$, then there exists a path from i to j that is unblocked by K in both $\mathcal{G}_1$ and $\mathcal{G}_2$.*

Proof of Lemma 4.31. If $k - l - m$ constitutes a moral collider so the edges are oriented as $k \rightarrow l \leftarrow m$, we show that the shorter path that bypasses l (by directly going through the $k \rightarrow m$ or $k \leftarrow m$ edge) is also unblocked by K, thus contradicting the hypothesis. Because k, m are not colliders in the original path (there cannot be two consecutive colliders) and the original path is unblocked by K, we have $k, m \notin K$. Suppose the path is like $i \cdots k \rightarrow l \leftarrow m \cdots j$ (k could be the same as i and m could be the same as j). The sub-paths $i \cdots k$ and $j \cdots m$ are not blocked by K because the original path is not blocked by K. Although k or m might be a collider on the new path, none of them blocks the path when conditioning on K because k and m are parents of $l \in K$.

The other possibility of $k - l - m$ is that it forms a chain, for example $k \rightarrow l \rightarrow m$. In this case $k - m$ must be oriented as $k \rightarrow m$ in order to not not create a cycle. By observing that any intermediate vertex is a collider in the new path if and only if it is a collider in the original path, it is easy to show that the shorter path bypassing l is also unblocked by K, thus resulting in a contradiction.

The only remaining possibility is $k \leftarrow l \rightarrow m$. Since k and m are adjacent, without loss of generality let's say the orientation is $k \rightarrow m$. It must be the case that k is a collider in the original path, otherwise the shorter path bypassing l would have the same colliders and non-colliders as the original path (except l) and is thus unblocked by K. □

Proof of Lemma 4.32. Consider the shortest unblocked path between i and j in $\mathcal{G}_1$. The goal is to show that the same path (or some path constructed based on this path) is unblocked by K in $\mathcal{G}_2$. By Lemma 4.31, all $\mathcal{G}_1$-colliders in this path are immoral and hence are also $\mathcal{G}_2$-colliders (since $\mathcal{G}_1$ and $\mathcal{G}_2$ share the same immoralities). Consider any intermediate vertex l in this path; if there is none then obviously the path can't be blocked by any K in $\mathcal{G}_2$. Let's say the adjacent vertices are k, m. The vertex l must either be

(i) A non-collider in both $\mathcal{G}_1$ and $\mathcal{G}_2$; or

(ii) A non-collider in $\mathcal{G}_1$ and collider in $\mathcal{G}_2$; or

(iii) A collider in both $\mathcal{G}_1$ and $\mathcal{G}_2$.

Obviously K does not block the path in $\mathcal{G}_2$ at the first kind of vertices, because the path is unblocked in $\mathcal{G}_1$. There can be no l of the second kind for the following reason. Because l is a collider in $\mathcal{G}_2$ but not in $\mathcal{G}_\infty$, the parents k, m of l must be adjacent; otherwise $\mathcal{G}_1$ would not share the immorality $k \to l \leftarrow m$ with $\mathcal{G}_2$. By Lemma 4.31, the orientation in $\mathcal{G}_1$ must be $k \leftarrow l \to m$, and at least one of k, m is an immoral collider in $\mathcal{G}_1$. However, k, m cannot be colliders in $\mathcal{G}_2$, which contradicts the hypothesis that $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same immoralities.

Finally we consider the third case that l is a collider in both $\mathcal{G}_1$ and $\mathcal{G}_2$. Because the shortest path we are considering is unblocked by K in $\mathcal{G}_1$, l or a $\mathcal{G}_1$-descendant of l must be in K. Among $\overline{de}_{\mathcal{G}_1}(l) \cap K$, let o be a vertex that is closest to l (if there are several, let o be any one of them). There are three possibilities:

(i) $o = l$;

(ii) $o \in ch_{\mathcal{G}_1}(l)$;

(iii) The length of the shortest directed path from l to o in $\mathcal{G}_1$ is at least 2.

In the first case, the path not blocked at l in $\mathcal{G}_2$ because $l \in K$. In the second or the third case, the path is blocked at l in $\mathcal{G}_2$ only if edge directions in the shortest direct path from l to o in $\mathcal{G}_1$ are no longer the same in $\mathcal{G}_2$. In the third case, we claim that this shortest directed path from l to o must also a directed path in $\mathcal{G}_2$, hence $o \in \overline{de}_{\mathcal{G}_2}(l)$ and the path is unblocked at l in $\mathcal{G}_2$. If this not the true, this path from l to o must have a $\mathcal{G}_2$-collider. In order to not create an immoral collider that does not exist in $\mathcal{G}_1$, the two $\mathcal{G}_\in -$ parents of this collider must be adjacent. In $\mathcal{G}_1$, this edge either results in a cycle or a shorter directed path from l to o.

We are left with the second case with the edge direction $l \leftarrow o$ in $\mathcal{G}_2$. In order to not create the immorality $k \to l \leftarrow o$ which would be inconsistent with $\mathcal{G}_1$, k, o must be adjacent; furthermore, the direction must be $k \to o$ in $\mathcal{G}_1$ to not create a $k \to l \to o \to k$ cycle. For the same reason, m, o must be adjacent and the direction must be $m \to o$ in $\mathcal{G}_1$. Hence $k \to o \leftarrow m$ is a immorality in $\mathcal{G}_1$ and must also be a immorality in $\mathcal{G}_2$. Consider a new path modified from the original path from i to j with $k \to l \leftarrow m$ replaced by $k \to o \leftarrow m$. It is easy to show that this new path has the same length as the original path and is also unblocked bt K in $\mathcal{G}_1$ and $\mathcal{G}_2$ because $o \in K$. Apart from o, any other vertex in the new path is a $\mathcal{G}_2$-collider if and only if it is a $\mathcal{G}_2$-collider in the original path. We can continue to use the arugment in this paragraph until we no longer have a collider l with a $\mathcal{G}_1$-child but not a $\mathcal{G}_2$-child o in K. □

Chapter 5

A unifying theory of causality

This Chapter introduces a theory that unifies the previous approaches:

(i) The potential outcomes are useful to elucidate the causal effect that is being estimated in a randomised experiment (Chapter 2);

(ii) The linear structural equations can be used to define causal effects and distinguish correlation from causation (Chapter 3);

(iii) The graphical models (particularly DAGs) can encode conditional independence relations and can be used to represent causality (Chapter 4).

5.1 From graphs to structural equations to counterfactuals

The key idea is to define counterfactuals from a DAG by using nonparametric structural equations:[1]

5.1 Definition (NPSEMs)**.** Given a DAG $\mathcal{G} = (V = [p], E)$, the random variables $\boldsymbol{X} = \boldsymbol{X}_{[p]}$ satisfy a *nonparametric structural equation model (NPSEM)* if the observed and interventional distributions of $\boldsymbol{X}_{[p]}$ satisfy

$$X_i = f_i(\boldsymbol{X}_{pa_{\mathcal{G}}(i)}, \epsilon_i),\ i = 1, \ldots, p,$$

for some functions $f_1, \ldots, f_p$ and random variables $\boldsymbol{\epsilon}_{[p]}$.

5.2 Definition (Counterfactuals)**.** Given the above NPSEM, the counterfactual variables $\{X_i(\boldsymbol{X}_J = \boldsymbol{x}_J) \mid i \in [p], J \subseteq [p]\}$ ($X_i(\boldsymbol{X}_J = \boldsymbol{x}_J)$ is often abbreviated as $X_i(\boldsymbol{x}_J)$) are defined as follows:

(i) *Basic counterfactuals:* For each X_i and intervention $\boldsymbol{X}_{pa(i)} = \boldsymbol{x}_{pa(i)}$, define $X_i\big(\boldsymbol{X}_{pa(i)} = \boldsymbol{x}_{pa(i)}\big) = f_i(\boldsymbol{x}_{pa(i)}, \epsilon_i)$.

(ii) *Substantative counterfactuals:* For any $i \in [p]$ and $J \subseteq an(i)$, recursively define

$$X_i(\boldsymbol{X}_J = \boldsymbol{x}_J) = X_i\Bigg(X_k = x_k I(k \in J) + X_k(\boldsymbol{x}_{J \cap an(k)}) I(k \notin J), k \in pa(i)\Bigg).$$

(iii) *Irrelevant counterfactuals:* For any $i \in [p]$ and $J \not\subseteq an(i)$, $X_i(\boldsymbol{x}_J) = X_i(\boldsymbol{x}_{J \cap an(i)})$.

5.3 Remark. The recursive definition of substantative counterfactuals is easy to comprehend but makes the algebra clunky (because we need to keep track of the relevant intervention). An equivalent but more compact definition for the non-basic counterfactual can be obtained via *recursive substitution*: For any $i \in [p], J \subseteq [p]$ and $J \neq pa(i)$, the counterfactual $X_i(\boldsymbol{x}_J)$ is defined recursively by

$$X_i(\boldsymbol{X}_J = \boldsymbol{x}_J) = X_i\big(\boldsymbol{X}_{pa(i) \cap J} = \boldsymbol{x}_{pa(i) \cap J}, \boldsymbol{X}_{pa(i) \setminus J} = \boldsymbol{X}_{pa(i) \setminus J}(\boldsymbol{x}_J)\big).$$

We will often abbreviate $X_i(\boldsymbol{X}_J = \boldsymbol{x}_J)$ as $X_i(\boldsymbol{x}_J)$, where the subscript J indicates the intervention. For disjoint $J, K \subset [p]$, we write $X_i(\boldsymbol{X}_{J \cup K} = \boldsymbol{x}_{J \cup K})$ as $X_i(\boldsymbol{x}_J, \boldsymbol{x}_K)$.

By induction, it is easy to show that $X_i = X_i(\boldsymbol{x}_\emptyset)$. This is consistent with our intution that X_i is the "counterfactual" corresponding to no intervention.

5.4 Example. Consider the graphical model in Figure 5.1. The basic counterfactuals are X_1, $X_2(x_1)$, and $X_3(x_1, x_2)$. The substantative counterfactuals of X_3 are defined using the basic counterfactuals as

$$X_3(x_1) = X_3\big(x_1, X_2(x_1)\big) \text{ and } X_3(x_2) = X_3(X_1, x_2).$$

The other counterfactuals are irrelevant and can be trimmed into the basic or substantative counterfactuals. For example, $X_1(x_1) = X_1(x_2) = X_1$, $X_2(x_2) = X_2$, $X_2(x_1, x_2) = X_2(x_1)$.

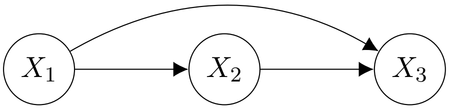


Figure 5.1: A simple graphical model illustrating the concept of counterfactuals.

We can further simplify some of the substantative counterfactuals:

5.5 Proposition. *Consider any disjoint $J, K \subseteq V$ and any $i \in V$. If K blocks all directed paths from J to i, then*

$$\boldsymbol{X}_i(\boldsymbol{x}_J, \boldsymbol{x}_K) = \boldsymbol{X}_i(\boldsymbol{x}_K).$$

Proof. This follows from recursive substitution and the next observation: if K blocks all directed paths from J to i, then K also blocks directed paths from J to $pa(i) \setminus K$. □

A corollary of Proposition 5.5 is that $X_i(\boldsymbol{x}_J) = X_i(\boldsymbol{X}_{J \cap an(i)})$ for all $i \in V$ and $j \subseteq V$. From Definitions 5.1 and 5.2, we immediately obtain

$$X_i = X_i(\boldsymbol{X}_{pa(i)}). \tag{5.1}$$

This generalises the consistency assumption in Chapter 2 (Assumption 2.6). Here, consistency is a property instead of an assumption because the connections between the factuals and counterfactuals are already made explicit in the definition of counterfactuals.

The next result further generalises the consistency property:[2]

5.6 Proposition. *For any disjoint $J, K \subseteq V$ and any $i \in V$,*

$$\boldsymbol{X}_J(\boldsymbol{x}_K) = \boldsymbol{x}_J \Longrightarrow X_i(\boldsymbol{x}_J, \boldsymbol{x}_K) = X_i(\boldsymbol{x}_K), \tag{5.2}$$

in the sense that the event defined on the left hand side is a subset of the event defined on the right hand side.

Proof. By definition,

$$X_i(\boldsymbol{x}_J, \boldsymbol{x}_K) = X_i(\boldsymbol{x}_{pa(i) \cap J}, \boldsymbol{x}_{pa(i) \cap K}, \boldsymbol{X}_{pa(i) \setminus J \setminus K}(\boldsymbol{x}_J, \boldsymbol{x}_K)),$$

$$X_i(\boldsymbol{x}_K) = X_i(\boldsymbol{X}_{pa(i) \cap J}(\boldsymbol{x}_K), \boldsymbol{x}_{pa(i) \cap K}, \boldsymbol{X}_{pa(i) \setminus J \setminus K}(\boldsymbol{x}_K)).$$

By using the assumption $\boldsymbol{X}_J(\boldsymbol{x}_K) = \boldsymbol{x}_J$, we see it suffices to show $\boldsymbol{X}_{pa(i) \setminus J \setminus K}(\boldsymbol{x}_J, \boldsymbol{x}_K) = \boldsymbol{X}_{pa(i) \setminus J \setminus K}(\boldsymbol{x}_K)$. We can then complete the proof by induction. □

Notice that (5.1) is implied by (5.2) by letting $K = \emptyset$.

5.2 Markov properties for counterfactuals

If you come from a statistics background, it may be natural to assume the error variables $\epsilon_1, \dots, \epsilon_p$ are mutually independent. But after translating it into the basic counterfactuals, this assumption may seem rather strong.[3]

5.7 Definition (Basic counterfactual independence)**.** A NPSEM is said to satisfy the *single-world independence assumptions*, if

$$\text{For any } \boldsymbol{x}_{[p]}, \text{ the variables } X_i(\boldsymbol{x}_{pa(i)}),\ i \in [p], \text{ are mutually independent.} \tag{5.3}$$

A NPSEM is said to satisfy the the *multiple-world independence assumptions*, if $\epsilon_1, \dots, \epsilon_p$ are mutually independent. Equivalently,

$$\text{The sets of variables } \{X_i(\boldsymbol{x}_{pa(i)}) \mid \boldsymbol{x}_{pa(i)}\},\ i \in [p] \text{ are mutually independent.}$$

5.8 Example. Consider the graph in Figure 5.1. The single-world independence assumptions assert that X_1, $X_2(x_1)$, $X_3(x_1, x_2)$ are mutually independent for any x_1 and x_2. The multiple-world independence assumptions are

$$X_1 \perp\!\!\!\perp \{X_2(x_1) \mid x_1\} \perp\!\!\!\perp \{X_3(x_1, x_2) \mid x_1, x_2\}.$$

Thus in addition to the single-world independence assumptions, the multiple-world independence assumptions also make the following *cross-world independence assumption*:

$$X_2(x_1) \perp\!\!\!\perp X_3(\tilde{x}_1, x_2) \text{ for any } x_1 \neq \tilde{x}_1, x_2.$$

Cross-world independence is controversial because it is about two counterfactuals that can never be observed together in any experiment. Fortunately, it is not needed in most causal inference problems.

5.9 Remark. Whether the cross-world independence seems "natural" depends on how we approach the problem. If we start from structural equations, the multiple-world independence assumptions may seem natural. If we start from counterfactuals, the same assumptions may seem unnecessarily strong. But the two frameworks are equivalent. To give an example of a NPSEM that satisfies the single-world but not the cross-world independence assumptions, suppose all variables are discrete. We can simply let ϵ_i be the vector of all the basic counterfactuals of X_i and f_i select the basic counterfactual according to $\boldsymbol{x}_{pa(i)}$.

5.10 Definition (Single-world causal model)**.** We say the random variables $\boldsymbol{X}_{[p]}$ satisfy a *single-world causal model* or simply a *causal model* defined by a DAG $\mathcal{G} = (V = [p], E)$, if $\boldsymbol{X}_{[p]}$ satisfies a NPSEM defined by $\mathcal{G}$ and the counterfactuals of $\boldsymbol{X}_{[p]}$ satisfy the single-world independence assumptions.

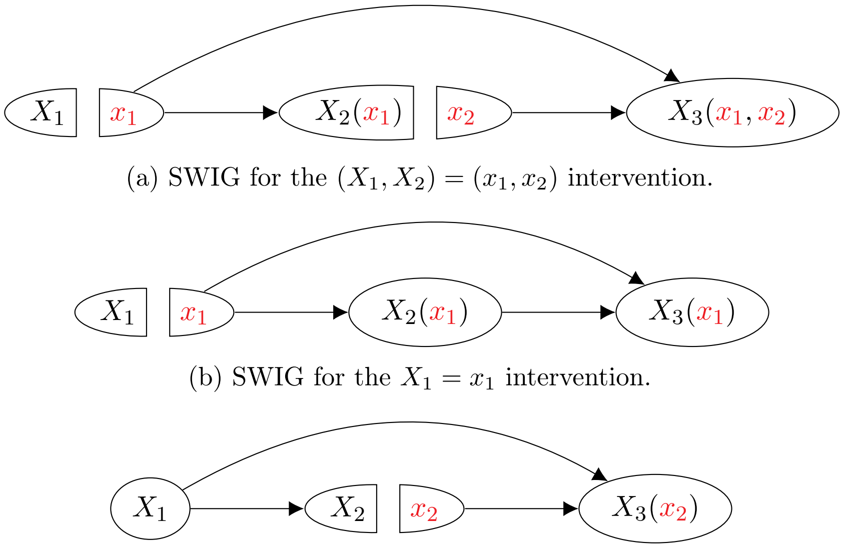
Next we introduce a transformation that maps a graphical model for the factual variables $\boldsymbol{X}$ to a graphical model for the counterfactual variables $\boldsymbol{X}(\boldsymbol{x}_J)$.

5.11 Definition. The *single-world intervention graph* (SWIG) $\mathcal{G}[\boldsymbol{X}(\boldsymbol{x}_J)]$ (sometimes abbreviated as $\mathcal{G}[\boldsymbol{x}_J]$) for the intervention $\boldsymbol{X}_J = \boldsymbol{x}_J$ is constructed from $\mathcal{G}$ via the following two steps:

(i) **Node splitting:** For every $j \in J$, split the vertex X_j into a random and a fixed component, labelled X_j and x_j respectively. The random half inherited all edges into X_j and the fixed half inherited all edges out of X_j.

(ii) **Labelling:** For every random node X_i in the new graph, label it with $X_i(\boldsymbol{x}_J) = X_i(\boldsymbol{x}_{J\cap an(i)})$.

5.12 Example. Figure 5.2 shows the SWIGs for the graphical model in Figure 5.1.



(a) SWIG for the $(X_1, X_2) = (x_1, x_2)$ intervention.

(b) SWIG for the $X_1 = x_1$ intervention.

(c) SWIG for the $X_2 = x_2$ intervention.

Figure 5.2: SWIGs for the graphical model in Figure 5.1.

The next Theorem states that in a single-world causal model, the counterfactuals $\boldsymbol{X}(\boldsymbol{x}_J)$ “factorise” according to the SWIG $\mathcal{G}[\boldsymbol{X}(\boldsymbol{x}_J)]$. “Factorise” is quoted because $\mathcal{G}[\boldsymbol{X}(\boldsymbol{x}_J)]$ has non-random vertices and we have not formally defined a graphical model for a mixture of random and non-random quantities. In this case, we essentially always condition on the fixed quantities, so in the graph they block all the paths they are on.

To simplify this, let $\mathcal{G}^*[\boldsymbol{X}(\boldsymbol{x}_J)]$ be the random part of $\mathcal{G}[\boldsymbol{X}(\boldsymbol{x}_J)]$, i.e., the subgraph of $\mathcal{G}[\boldsymbol{X}(\boldsymbol{x}_J)]$ restricted to $X_i(\boldsymbol{x}_J)$, $i \in [p]$. This is sometimes abbreviated as $\mathcal{G}^*[\boldsymbol{x}_J]$. Thus $\mathcal{G}[\boldsymbol{x}_J]$ has the same number of edges as $\mathcal{G}$ and $\mathcal{G}^*[\boldsymbol{x}_J]$ has the same number of vertices as $\mathcal{G}$.

5.13 Theorem (Factorisation of counterfactual distributions)**.** *Suppose $\boldsymbol{X}$ satisfies the causal model defined by a DAG $\mathcal{G}$, then $\boldsymbol{X}(\boldsymbol{x}_J)$ factorises according to $\mathcal{G}^*[\boldsymbol{X}(\boldsymbol{x}_J)]$ for all $J \subseteq [p]$.*

A key step in the proof of Theorem 5.13 is to establish the following Lemma using induction.

5.14 Lemma. *For any $k \notin J \subseteq [p]$ such that $de(k) \subseteq J$, $i \in [p]$, $\boldsymbol{x}_J$ and $\tilde{\boldsymbol{x}}$,*

$$\begin{aligned}&\mathbb{P}\Big(X_i(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{x}_i \,\Big|\, \boldsymbol{X}_{pa(i)\setminus J\setminus\{k\}}(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J\setminus\{k\}}\Big)\\ =&\mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \,\Big|\, \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big).\end{aligned}$$

Proof of Lemma 5.14 and Theorem 5.13. To simplify the exposition, let $\mathcal{G}^*[J]$ denote the modified graph $\mathcal{G}^*[\boldsymbol{X}(\boldsymbol{x}_J)]$ with the vertex mapping $X_i(\boldsymbol{x}_J) \to i$, so $\mathcal{G}^*[J]$ can be obtained by removing the outgoing arrows from J in $\mathcal{G}$. Notice that for any $i \in [p]$ and $J \subset [p]$, $pa_{\mathcal{G}^*[J]}(i) = pa_{\mathcal{G}}(i) \setminus J$.

The single-world independence assumptions in Equation (5.3) means that this conclusion is true for $J = [p]$. The Theorem can be proven by reverse induction from $J \cup \{k\} \subseteq [p]$ to J where $k \notin J$ and $de(k) \subseteq J$ (Exercise 3.6 shows that such k always exists). By Proposition 5.6 (consistency of counterfactuals),

$$\mathbb{P}\big(\boldsymbol{X}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}\big) = \mathbb{P}\big(\boldsymbol{X}(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{\boldsymbol{x}}\big), \text{ for any } \tilde{\boldsymbol{x}}.$$

Using the induction hypothesis, we have, for any $\tilde{\boldsymbol{x}}$,

$$\mathbb{P}(\boldsymbol{X}(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{\boldsymbol{x}}) = \prod_{i=1}^{p} \mathbb{P}\Big(X_i(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{x}_i \,\Big|\, \boldsymbol{X}_{pa_{\mathcal{G}^*[J\cup\{k\}]}(i)}(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{\boldsymbol{x}}_{pa_{\mathcal{G}^*[J\cup\{k\}]}(i)}\Big), \tag{5.4}$$

By using Lemma 5.14, we have, for any $\tilde{\boldsymbol{x}}$,

$$\mathbb{P}(\boldsymbol{X}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}) = \prod_{i=1}^{p} \mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \,\Big|\, \boldsymbol{X}_{pa_{\mathcal{G}^*[J]}(i)}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa_{\mathcal{G}^*[J]}(i)}\Big). \tag{5.5}$$

This shows that $\boldsymbol{X}(\boldsymbol{x}_J)$ factorises according to $\mathcal{G}^*[J]$.

We next prove Lemma 5.14 using the induction hypothesis. If $i \notin de(k)$, then there is no directed path from k to i or $pa(i)$. If $i \in de(k)$ but $i \notin ch(k)$, all the directed paths from k to i or $pa(i) \setminus J$ (if non-empty) are blocked by J, because $de(k) \subset J$ and there are no directed paths from k to $pa(i) \setminus J$. In both cases above, Proposition 5.5 shows that $X_i(\boldsymbol{x}_J) = X_i(\boldsymbol{x}_J, \tilde{x}_k)$ and $X_{pa(i)\setminus J}(\boldsymbol{x}_J) = X_{pa(i)\setminus J}(\boldsymbol{x}_J, \tilde{x}_k)$, which proves the identity in Lemma 5.14.

If $i \in ch_{\mathcal{G}}(k)$, we have

$$\begin{aligned}
&\mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big) \\
=&\mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J\setminus\{k\}}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J\setminus\{k\}}, X_k(\boldsymbol{x}_J) = \tilde{x}_k\Big) \\
=&\mathbb{P}\Big(X_i(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J\setminus\{k\}}(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J\setminus\{k\}}, X_k(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{x}_k\Big) \\
=&\mathbb{P}\Big(X_i(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J\setminus\{k\}}(\boldsymbol{x}_J, \tilde{x}_k) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J\setminus\{k\}}\Big).
\end{aligned}$$

The second equality follows from the consistency of counterfactuals (Proposition 5.6). The third equality follows from the conditional independence between $X_i(\boldsymbol{x}_{J\cup\{k\}})$ and $X_k(\boldsymbol{x}_{J\cup\{k\}})$ that follows from the induction hypothesis and the observation that $pa_{\mathcal{G}^*[J\cup\{k\}]}(i)$ d-separates i from k in $\mathcal{G}^*[J \cup \{k\}]$. □

5.15 Remark. (Completeness of d-separation in SWIGs) The completeness of d-separation in DAGs (Remark 4.23) also extends to SWIGs, in the sense that if two counterfactuals are d-connected given a third counterfactual, then there exists a distribution of the factuals and counterfactuals obeying Theorem 5.13 and Lemma 5.18 in which the two counterfactuals are dependent given the third.

5.16 Exercise. Consider the causal model defined by the graph in Figure 5.3. Show that $Y(a_1, a_2) \perp\!\!\!\perp A_1$ and $Y(a_2) \perp\!\!\!\perp A_2 \mid A_1, X$.

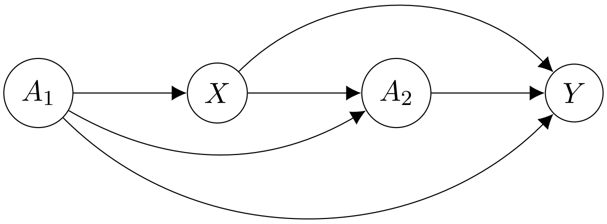


Figure 5.3: A sequentially randomised experiment (A_1 and A_2 are time-varying treatments).

5.3 From counterfactual to factual

Theorem 5.13 allows us to use d-separation to check all conditional independences in a causal DAG model. This can be used, together with the consistency property (Proposition 5.6), to identify causal effects.

5.17 Example (Continuing from Example 5.12)**.** By using d-separation for the SWIG in Figure 5.2c, we have $X_2 \perp\!\!\!\perp X_3(x_2) \mid X_1$ for any x_2. This conditional independence is the same as the randomisation assumption (Assumption 2.10) in Chapter 2. We can then apply Theorem 2.12 with $X = X_1$, $A = X_2$, $Y = X_3$ to obtain

$$\mathbb{P}(X_3(x_2) = x_3 \mid X_1 = x_1) = \mathbb{P}(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2). \tag{5.6}$$

Next, we give some general results that link counterfactual distributions with factual distributions. The first Lemma establishes the *modularity* property of the counterfactual distribution.[4] A proof of this result can be found in the appendix.

5.18 Lemma. *Suppose $\boldsymbol{X}$ satisfies the causal model defined by a DAG $\mathcal{G} = ([p], E)$. Then for any $i \in [p]$, $J \subseteq [p]$ and $\tilde{\boldsymbol{x}}$,*

$$\begin{aligned}&\mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big)\\ =&\mathbb{P}\Big(X_i = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J} = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}, \boldsymbol{X}_{pa(i)\cap J} = \boldsymbol{x}_{pa(i)\cap J}\Big),\end{aligned}$$

5.19 Example (Continuing from Example 5.17)**.** By letting $i = 3$ and $J = \{2\}$ in Lemma 5.18, we obtain

$$\mathbb{P}(X_3(x_2) = \tilde{x}_3 \mid X_1(x_2) = \tilde{x}_1) = \mathbb{P}(X_3 = \tilde{x}_3 \mid X_1 = \tilde{x}_1, X_2 = x_2),$$

which is exactly the same as (5.6).

Because $\boldsymbol{X}(\boldsymbol{x}_J)$ factorises according to $\mathcal{G}^*[\boldsymbol{X}(\boldsymbol{x}_J)]$ (Theorem 5.13), Lemma 5.18 implies that

5.20 Theorem. *Suppose $\boldsymbol{X}$ satisfies the causal model defined by a DAG $\mathcal{G}$, then for any $J \subseteq [p]$,*

$$\mathbb{P}(\boldsymbol{X}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}) = \prod_{i=1}^{p} \mathbb{P}\Big(X_i = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\cap J} = \boldsymbol{x}_{pa(i)\cap J}, \boldsymbol{X}_{pa(i)\setminus J} = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big).$$

In practice, we are often more interested in the marginals of $\boldsymbol{X}(\boldsymbol{x}_J)$. The next result simplifies the marginalisation.

5.21 Corollary. *Suppose $\boldsymbol{X}$ is discrete. Then for any disjoint $I, J \subseteq [p]$, let $K = [p] \setminus (I \cup J)$, then*

$$\mathbb{P}(\boldsymbol{X}_I(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_I) = \sum_{\tilde{\boldsymbol{x}}_K} \prod_{i \in I \cup K} \mathbb{P}\Big(X_i = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\cap J} = \boldsymbol{x}_{pa(i)\cap J}, \boldsymbol{X}_{pa(i)\setminus J} = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big).$$

For continuous $\boldsymbol{X}$, replace the summation by integral.

Proof. By Theorem 5.20,

$$\mathbb{P}(\boldsymbol{X}_I(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_I) = \sum_{\tilde{\boldsymbol{x}}_K, \tilde{\boldsymbol{x}}_J} \prod_{i=1}^{p} \mathbb{P}\Big(X_i = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\cap J} = \boldsymbol{x}_{pa(i)\cap J}, \boldsymbol{X}_{pa(i)\setminus J} = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big).$$

For any $i \in J$, $\tilde{x}_i$ only appears in the ith term in the product. When marginalising over such $\tilde{x}_i$, this term sums up to 1 (because it is a conditional density). □

The identity in Corollary 5.21 is known as the *g-compuation formula* (or simply the *g-formula*, g for generalised)[5] or the *truncated factorisation*[6].

5.22 Example (Continuing from Example 5.19)**.** By applying Theorem 5.20 with $I = \{3\}$ and $J = \{2\}$, we have

$$\mathbb{P}(X_1 = \tilde{x}_1, X_2 = \tilde{x}_2, X_3(x_2) = \tilde{x}_3) = \mathbb{P}(X_1 = \tilde{x}_1)\,\mathbb{P}(X_2 = \tilde{x}_2)\,\mathbb{P}(X_3 = x_3 \mid X_1 = \tilde{x}_1, X_2 = x_2).$$

By summing over $\tilde{x}_1$ and $\tilde{x}_2$, we obtain

$$\begin{aligned}
&\mathbb{P}\left(X_3(x_2) = \tilde{x}_3\right) \\
=& \sum_{\tilde{x}_1, \tilde{x}_2} \mathbb{P}(X_1 = \tilde{x}_1)\,\mathbb{P}(X_2 = \tilde{x}_2)\,\mathbb{P}(X_3 = \tilde{x}_3 \mid X_1 = \tilde{x}_1, X_2 = x_2) \\
=& \sum_{\tilde{x}_1} \mathbb{P}(X_1 = \tilde{x}_1)\,\mathbb{P}(X_3 = \tilde{x}_3 \mid X_1 = \tilde{x}_1, X_2 = x_2) \sum_{\tilde{x}_2} \mathbb{P}(X_2 = \tilde{x}_2) \\
=& \sum_{\tilde{x}_1} \mathbb{P}(X_1 = \tilde{x}_1)\,\mathbb{P}(X_3 = \tilde{x}_3 \mid X_1 = \tilde{x}_1, X_2 = x_2),
\end{aligned}$$

which is what we would obtain if we directly apply the g-formula (Corollary 5.21). A cleaner form is

$$\mathbb{P}\left(X_3(x_2) = x_3\right) = \sum_{x_1} \mathbb{P}(X_1 = x_1)\,\mathbb{P}(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2),$$

which is simply a marginalisation of (5.6).

5.23 Remark. Notice that the formula for $\mathbb{P}(X_3(x_2) = x_3)$ above is generally different from the conditional distribution of X_3 given X_2:

$$\mathbb{P}\left(X_3 = x_3 \mid X_2 = x_2\right) = \sum_{x_1} \mathbb{P}(X_1 = x_1 \mid X_2 = x_2)\,\mathbb{P}(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2),$$

This generalises the discussion in Section 3.4 and demonstrates how "correlation does not imply causation".

5.24 Exercise. Consider the causal model defined by the graph in Figure 5.3. Suppose all the random variables are discrete.

(i) By applying the g-computation formula, show that

$$\mathbb{E}[Y(a_1, a_2)] = \sum_{x} \mathbb{P}(X = x \mid A_1 = a_1) \cdot \mathbb{E}[Y \mid A_1 = a_1, A_2 = a_2, X = x]. \quad (5.7)$$

(ii) Derive (5.7) using the conditional independence in Exercise 5.16 and the consistency of counterfactuals (Proposition 5.6).

5.25 Remark. In the identification formula (5.7), the condition expectation $\mathbb{E}[Y \mid A_1 = a_1, A_2 = a_2, X = x]$ is weighted by $\mathbb{P}(X = x \mid A_1 = a_1)$ instead of the marginal probability $\mathbb{P}(X = x)$ in (5.22). This makes (5.7) a non-trivial extension to the simple case with one treatment variable. Intuitively, the dilemma is that, in order to recover the causal effect of A_2 on Y, we need to condition on their confounder X_2. However, this blocks the directed path $A_1 \to X \to Y$ and makes the estimated causal effect of A_1 on Y biased.

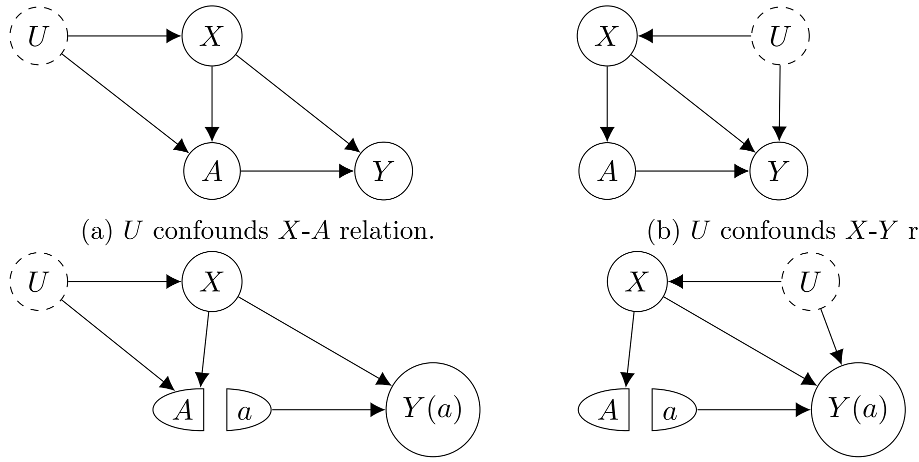
5.4 Causal identification

A major limitation of the results in the last section is that they require all relevant variables in the causal model to be measured. This is rarely the case in practice.

Fortunately, in many problems with unmeasured variables it may still be possible to identify the causal effect. To focus on the main ideas, we will assume all the random variables are discrete in the examples below.

5.4.1 Back-Door formula

5.26 Example. Consider the causal graphs in Figures 5.4a and 5.4b. We cannot directly apply the g-formula to identify the causal effect of A on Y, because there is an unmeasured variable U in both cases. However, the same identification formula is still correct because $A \perp\!\!\!\perp Y(a) \mid X$ still holds (you can check this using the SWIGs in Figures 5.4c and 5.4d). It then follows from Theorem 2.12 that $\mathbb{P}(Y(a) = y \mid X = x) = \mathbb{P}(Y = y \mid A = a, X = x)$ for all a, x, y.



(a) U confounds X-A relation.

(b) U confounds X-Y relation.

(c) SWIG for (a).

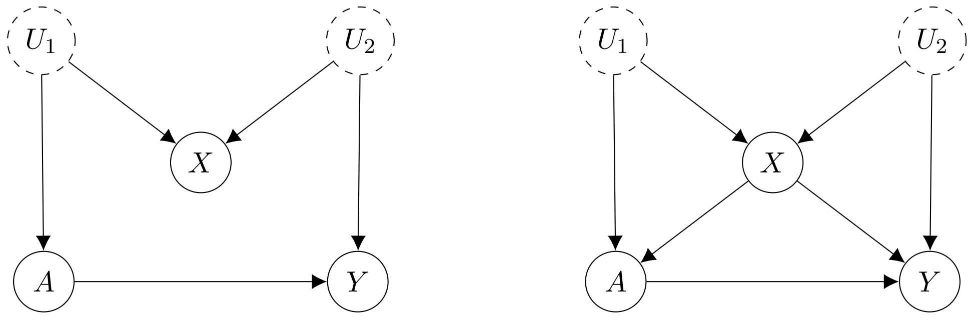
(d) SWIG for (b).

Figure 5.4: It suffices to adjust for X in these scenarios to estimate the average causal effect of A on Y.

The crucial condition here is $A \perp\!\!\!\perp Y(a) \mid X$. This has appeared before in Chapter 2 under the name "randomisation assumption". With observational data, the same assumption is usually called the *no unmeasured confounders* assumption, because we can no longer guarantee it by physically randomising A.

5.27 Remark. A common name for $A \perp\!\!\!\perp Y(a) \mid X$ is *treatment assignment ignorability* or simply *ignorability*[7]. The underlying idea is that A can be treated as a missingness indicator for $Y(0)$ (and $1-A$ for $Y(1)$), and this assumption says that the missingness is "ignorable". Essentially, this assumption allows us to treat the observational study as if the data came from a randomised experiment (but with unknown distribution of A given X). Another name is *exchangeability*[8]. Our nomenclature emphasises on the structural (instead of the statistical) content of the assumption.

5.28 Remark. In observational studies, a widely held belief is that the more pre-treatment covariates being included in $\boldsymbol{X}$, the more "likely" the assumption $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}$ is satisfied. However, this is not necessarily true; see Figure 5.5 for two counterexamples.[9]



(a) M-bias.

(b) Butterfly bias.

Figure 5.5: Counter-examples for the claim that adjusting for all observed variables that temporally precedes the treatment would be sufficient. U_1 and U_2 are unobserved.

In the graphical framework, we can check $Y(a) \perp\!\!\!\perp A \mid X$ by d-separation in the SWIG. Because there is no out-going arrow from A in $\mathcal{G}^*[a]$, this essentially says that every *back-door path* from A to Y (meaning the path has an edge going into A) must be blocked by X.

5.29 Proposition (Back-door adjustment)**.** *Suppose* $(\boldsymbol{X}, A, Y)$ *are random variables in a causal model* $\mathcal{G}$ *that may contain other unobserved variables. Suppose* $\boldsymbol{X}$ *contains no descendant of* A *and blocks every back-door path from* A *to* Y *in* $\mathcal{G}$*. Then* $Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}$ *for all* a *and*

$$\mathbb{P}(Y(a) \leq y) = \sum_{\boldsymbol{x}} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \cdot \mathbb{P}(Y \leq y \mid A = a, \boldsymbol{X} = \boldsymbol{x}),\ \forall a, \boldsymbol{x}, y.$$

5.30 Exercise. Why is it necessary to assume $\boldsymbol{X}$ contains no descendant of A?

5.4.2 Front-door formula

The back-door formula cannot be applied when there are unmeasured confounders between A and Y. The front-door formula is designed to overcome this problem by decomposing the causal effect of A on Y into unconfounded mechanisms.[10]

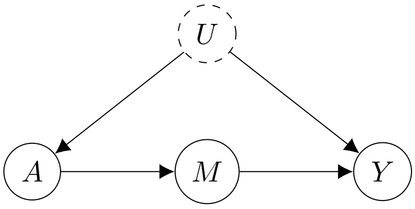


Figure 5.6: A causal model showing the causal effect of A on Y being entirely mediated by M.

5.31 Example. Consider the DAG in Figure 5.6. By recursive substitution,

$$Y(a,m) = Y(m). \tag{5.8}$$

Using d-separation in the corresponding SWIGs, we have

$$Y(m) \perp\!\!\!\perp M(a), M(a) \perp\!\!\!\perp A,\; Y(m) \perp\!\!\!\perp M \mid A. \tag{5.9}$$

In other words, $A \to M$ is unconfounded and $M \to Y$ is unconfounded given A. Using the law of total probability, we obtain

$$\begin{aligned}\mathbb{P}(Y(a) = y) &= \mathbb{P}\big(Y(a, M(a)) = y\big)\\ &= \mathbb{P}\big(Y(M(a)) = y\big)\\ &= \sum_m \mathbb{P}\big(Y(M(a)) = y, M(a) = m\big)\\ &= \sum_m \mathbb{P}\big(Y(m) = y, M(a) = m\big)\\ &= \sum_m \mathbb{P}\big(Y(m) = y\big) \cdot \mathbb{P}\big(M(a) = m\big).\end{aligned}$$

The distribution of $Y(m)$ and $M(a)$ can be identified by the back-door formula. Thus

$$\mathbb{P}(Y(a) = y) = \sum_m \Big\{ \sum_{a'} \mathbb{P}(Y = y \mid M = m, A = a')\, \mathbb{P}(A = a') \Big\}\, \mathbb{P}(M = m \mid A = a). \tag{5.10}$$

There are two key counterfactual relations in Example 5.31. First, the *exclusion restriction* (5.8) allows us to decompose the effect of A on Y to the product of the effect of A on M and the effect of M on Y. This is possible because M blocks all the directed path from A to Y, often referred to as the *front-door condition.* Next, the no unmeasured confounder conditions in (5.9) allow us to use back-door adjustment to identify the effect of A on M and the effect of M on Y.

5.4.3 Counterfactual calculus

Back-door and front-door are two graphical conditions for causal identification. More generally, there are three graphical rules that allow us to simplify counterfactual distributions:[11]

5.32 Proposition (Counterfactual calculus)**.** *Consider a causal model with respect to a DAG $\mathcal{G}$. For any disjoint subsets of variables, $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W} \subset V$, and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}$, we have*

(i) If $\boldsymbol{Z}(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{Y}(\boldsymbol{x}) \mid \boldsymbol{W}(\boldsymbol{x})$ $[\mathcal{G}^*[\boldsymbol{x}]]$, then

$$\mathbb{P}\big(\boldsymbol{Y}(\boldsymbol{x}) = \boldsymbol{y} \mid \boldsymbol{Z}(\boldsymbol{x}), \boldsymbol{W}(\boldsymbol{x})\big) = \mathbb{P}\big(\boldsymbol{Y}(\boldsymbol{x}) = \boldsymbol{y} \mid \boldsymbol{W}(x)\big).$$

(ii) If $\boldsymbol{Y}(\boldsymbol{x}, \boldsymbol{z}) \perp\!\!\!\perp \boldsymbol{Z}(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{z})$ $[\mathcal{G}^*[\boldsymbol{x}, \boldsymbol{z}]]$, then

$$\mathbb{P}\big(\boldsymbol{Y}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{y} \mid \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{w}\big) = \mathbb{P}\big(\boldsymbol{Y}(\boldsymbol{x}) = \boldsymbol{y} \mid \boldsymbol{W}(\boldsymbol{x}) = \boldsymbol{w}, \boldsymbol{Z}(\boldsymbol{x}) = \boldsymbol{z}\big).$$

(iii) If $\boldsymbol{Y}(\boldsymbol{x}, \boldsymbol{z}) \perp\!\!\!\perp \boldsymbol{z}$ $[\mathcal{G}[\boldsymbol{x}, \boldsymbol{z}]]$ ($\boldsymbol{z}$ is the fixed half-vertex), then

$$\mathbb{P}\big(\boldsymbol{Y}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{y}\big) = \mathbb{P}\big(\boldsymbol{Y}(\boldsymbol{x}) = \boldsymbol{y}\big).$$

5.33 Exercise. Prove Proposition 5.32.

The rules in Proposition 5.32 provide a systematic, graphical approach to deduce causal identification. In our framework, they simplify follow from the SWIG Markov properties and basic properties of counterfactuals. These rules are not always user-friendly because all the counterfactual variables must be in the same world. In practice, we do not need to be so dogmatic.

5.34 Remark. Proposition 5.32 is the counterfactual version of the famous *do-calculus*[12]. By using SWIGs and counterfactuals, Proposition 5.32 is, however, much simpler and transparent than the original do-calculus. It has been shown that this calculus is complete for acyclic directed mixed graphs, in the sense that all identifiable counterfactual distributions can be deduced via repeatedly using these rules (and of course the probability calculus).[13]

The completeness of the counterfactual calculus does not preclude the possibility of causal identification with additional assumptions.

5.35 Exercise. Consider the causal diagram in Figure 5.7, where the effect of A on Y is confounded by an unmeasured variable U. The variable Z is called an *instrumental variable* and represents an unconfounded change to A. Show that if the variables satisfy a linear structural equation model according to the diagram, the causal effect of A on Y is identified by the so-called *Wald ratio*[14]

$$\beta_{AY} = \operatorname{Cov}(Z, Y) / \operatorname{Cov}(Z, A). \tag{5.11}$$

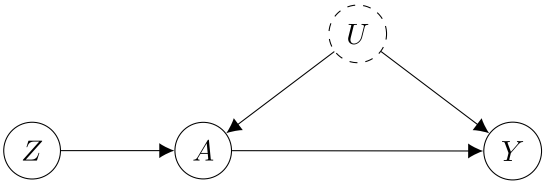


Figure 5.7: Instrumental variables.

In the example sheet you will further explore partial identification of the average treatment effect using instrumental variables without linearity.

5.5 Proofs (non-examinable)

5.5.1 Proof of Lemma 5.18

Consider any $k \in pa(i)$. Then by using counterfactual consistency (Proposition 5.6),

$$\begin{aligned}&\mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big)\\ =&\begin{cases}\mathbb{P}\Big(X_i(\boldsymbol{x}_{J\setminus\{k\}}) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_{J\setminus\{k\}}) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}, X_k(\boldsymbol{x}_{J\setminus\{k\}}) = x_k\Big), & \text{if } k \in J,\\ \mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J\setminus\{k\}}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J\setminus\{k\}}, X_k(\boldsymbol{x}_J) = \tilde{x}_k\Big), & \text{if } k \not\in J.\end{cases}\end{aligned}$$

Repeating this, we obtain

$$\begin{aligned}&\mathbb{P}\Big(X_i(\boldsymbol{x}_J) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_J) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}\Big)\\ =&\mathbb{P}\Big(X_i(\boldsymbol{x}_{J\setminus pa(i)}) = \tilde{x}_i \;\Big|\; \boldsymbol{X}_{pa(i)\setminus J}(\boldsymbol{x}_{J\setminus \mathrm{pa}(i)}) = \tilde{\boldsymbol{x}}_{pa(i)\setminus J}, \boldsymbol{X}_{pa(i)\cap J}(\boldsymbol{x}_{J\setminus pa(i)}) = \boldsymbol{x}_{pa(i)\setminus J}\Big).\end{aligned}$$

From here it is sufficient to remove the intervention $\boldsymbol{x}_{J\setminus pa(i)}$ from the right hand side.

By Proposition 5.5, we can first remove any intervention that is not on an ancestor of i. So without loss of generality we may assume $J \subseteq an(i)$. To achieve our goal, we first add $X_{J\setminus pa(i)}(\boldsymbol{x}_{J\setminus pa(i)}) = \boldsymbol{x}_{J\setminus pa(i)}$ to the conditioning event. This does not change the conditional probability because $X_i(\boldsymbol{x}_{J\setminus pa(i)})$ is d-separated from $X_{J\setminus pa(i)}(\boldsymbol{x}_{J\setminus pa(i)})$ by $X_{pa(i)}(\boldsymbol{x}_{J\setminus pa(i)})$ in the SWIG $\mathcal{G}[\boldsymbol{X}(\boldsymbol{x}_{J\setminus pa(i)})]$. We can then remove the intervention $\boldsymbol{x}_{J\setminus pa(i)}$ from all the counterfactuals by consistency. Finally, we can remove $X_{J\setminus pa(i)} = \boldsymbol{x}_{J\setminus pa(i)}$ from the conditioning event because X_i is d-separated from $X_{J\setminus pa(i)}$ by $X_{pa(i)}$ in $\mathcal{G}$.

Following Richardson and Robins (2013, p. 113), a previous version of the lecture notes suggested that Lemma 5.18 follows from repeatedly applying Lemma 5.14. However, an application of the factorization property seems needed to remove the interventions on $\boldsymbol{X}_{J\setminus pa(i)}$.

Notes

[1] Pearl, J. (2000). *Causality* (1st ed.). Cambridge University Press.

[2] Malinsky, D., Shpitser, I., & Richardson, T. (2019). A potential outcomes calculus for identifying conditional path-specific effects. *Proceedings of Machine Learning Research*, *89*, 3080.

[3] The next two Sections are based on Richardson, T. S., & Robins, J. M. (2013). *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality* (tech. rep. No. 128). Center for the Statistics and the Social Sciences, University of Washington Series.

[4] The term "modularity" is originally due to Pearl (2000). The notion here is due to Richardson and Robins (2013).

[5] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*(9-12), 1393–1512. doi:10.1016/0270-0255(86)90088-6.

[6] Pearl and Verma, 1991.

[7] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

[8] For connections with exchangeability in Bayesian statistics, see Saarela, O., Stephens, D. A., & Moodie, E. E. M. (2020). The role of exchangeability in causal inference. arXiv: 2006.01799 `[stat.ME]`.

[9] For an interesting scientific debate on this from different perspectives, see Sjölander, A. (2009). Propensity scores and m-structures. *Statistics in Medicine*, *28*(9), 1416–1420. doi:10.1002/sim.3532 and the reply in the same issue by Rubin.

[10] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. doi:10.1093/biomet/82.4.669.

[11] Malinsky et al., 2019.

[12] Pearl, 1995.

[13] Huang, Y., & Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the twenty-second conference on uncertainty in artificial intelligence* (pp. 217–224). UAI'06. Cambridge, MA, USA: AUAI Press; Shpitser, I., & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. *Proceedings of the 21st national conference on artificial intelligence - volume 2* (pp. 1219–1226). AAAI'06. Boston, Massachusetts: AAAI Press.

[14] Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, *11*(3), 284–300. doi:10.1214/aoms/1177731868.

Chapter 6

No unmeasured confounders: Randomisation inference

$$\text{Causal inference} \;\approx\; \text{Causal language/model} \;+\; \text{Statistical inference.}$$

We are at the turning point in this course. In the last Chapter, we have unified three seemingly different languages for causality: counterfactuals (Chapter 2), structural equation models (Chapter 3), and graphical models (Chapter 4). If you are a philosopher, the problem of causal inference might seem to have been solved. But for statisticians, the real (and fun) part is just beginning. Starting from this next Chapter, we will put this theory into practice—how can we actually make "good causal inference" in the real world?

6.1 The logic of observational studies

An observational study is an empirical investigation that utilises observation data. By observational (as opposed to experimental) data, we mean the data are collected without manipulation or intervention by the researcher. In fact, in many observational studies the data were recorded before a concrete research question was posed.

6.1 Remark. We will only talk about observational studies for causality below. But many principles below also apply to non-causal inference from observational data (such as estimating the basic reproduction number of COVID-19).

All observational (and experimental) studies have two stages:

(i) *Design:* Empirical data are collected and preprocessed in an organised way.

(ii) *Analysis:* A statistical method is then applied to answer the research question.

In statistics lectures (including this course), you are spending most of time learning useful statistical models for the data already collected and how the analysis can be done correctly and optimally.

In applications, it is the opposite: "design trumps analysis"[1].

Consider the following argument in a randomised experiment:

(i) *Design:* Suppose we let half of the patients to receive the treatment *at random.*

(ii) *Analysis: Significantly more* treated participants have a better outcome,

(iii) *Conclusion:* Therefore, the treatment must be *beneficial.*

The underlying logic is that randomisation allows us to choose between statistical error and causality. This reasoning is *inductive.*

Consider the another argument:

(i) *Design:* Suppose the observed patients are *pair matched*, so that the patients in the same pair have similar demographics and medical history.

(ii) *Analysis:* In *significantly more* pairs, the treated patient has a better outcome,

(iii) *Conclusion:* Therefore, the treatment must be *beneficial.*

In this example, randomisation is replaced by pair matching. As a consequence, apart from statistical error and causality, a third possible explanation is that the treated patients and the control patients are systematically different in some other way (for instance, different lifestyles).

So causal inference in observational studies is always *abductive* (inference to the best explanation). This is summarised in the following equation:[2]

$$\text{Causal estimator} - \text{True causal effect} = \text{Design bias} + \text{Modelling bias} + \text{Statistical noise}. \tag{6.1}$$

This is more than a conceptual statement. To make (6.1) more concrete, let $\boldsymbol{O}$ be all the observed variables (O for observed data), $\mathcal{O}$ is the distribution of $\boldsymbol{O}$. Similarly, let $\boldsymbol{F}$ denote the relevant factuals and counterfactuals in the causal question being asked (F for full data) and $\mathcal{F}$ be its distribution.

Then (6.1) amounts to the decomposition

$$\beta(\boldsymbol{O}_{[n]};\hat{\theta}) - \beta(\mathcal{F}) = \{\beta(\mathcal{O}) - \beta(\mathcal{F})\} + \{\beta(\mathcal{O};\theta) - \beta(\mathcal{O})\} + \{\beta(\boldsymbol{O}_{[n]};\hat{\theta}) - \beta(\mathcal{O};\theta)\}, \tag{6.2}$$

where β is a generic symbol for causal effect functional or estimator, $\boldsymbol{O}_{[n]}$ is the observed data of size n, θ is the parameter in a statistical model and $\hat{\theta} = \hat{\theta}(\boldsymbol{O}_{[n]})$ is an estimator of θ.

6.2 Example. In regression adjustment for randomised experiments, $\boldsymbol{O} = (\boldsymbol{X}, A, Y)$, $\boldsymbol{F} = (Y(0), Y(1))$, $\beta(\mathcal{F}) = \mathbb{E}[Y(1) - Y(0)]$, $\beta(\mathcal{O}) = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$, $\beta(\mathcal{O}, \theta)$ be the any of (2.14), (2.15), or (2.16), $\beta(\boldsymbol{O}_{[n]};\hat{\theta})$ be the corresponding (2.11), (2.12), or (2.13).

6.3 Exercise. In Example 6.2, how much is the design bias? How much is the modelling bias?

Unlike previous Chapters, we now use subscripts to index observations instead of variables. This convention will be used throughout the rest of this course as we are mainly concerned with statistical inference.

6.2 No unmeasured confounders

In this and the next Chapters, we will assume all relevant confounders are measured, so the observational study is mimicing a randomised experiment.

More specifically, let $A_i \in \{0, 1\}$ be a binary treatment for individual i, Y_i be its the outcome of interest with two counterfactuals $Y_i(0)$ and $Y_i(1)$, and $\boldsymbol{X}_i$ be a p-dimensional vector of covariates.

Although unncessary for the randomisation inference, for now we assume $(\boldsymbol{X}_i, A_i, Y_i(0), Y_i(1))$, $i = 1, \ldots, n$, are i.i.d. In this setting, subscripts are often suppressed to indicate a generic random variable.

We restate Assumption 2.10, but now with a different name:

6.4 Assumption (No unmeasured confounders)**.** $A \perp\!\!\!\perp Y(a) \mid \boldsymbol{X}$ for $a = 0, 1$.

In other words, we eliminate one of the possible explanations to observed associations *by assumption*. This is convenient for studying statistical methodologies but obviously optimistic for practical applications.

6.3 Matching algorithms

Matching is a popular observational study design. Matching is essentially a preprocessing algorithm that aims to reconstruct a pairwise randomised experiment (Example 2.4) or a stratified randomised experiment (Exercise 2.5) from observational data.

We will focus on *1-to-1 matching* below, but the algorithms can be easily extended to more general forms of matching.[3]

To simplify the exposition, we will assume $A_i = 1$ for $1 \le i \le n_1$ and $A_i = 0$ for $n_1 + 1 \le i \le n$.

An essential element is a measure of distance $d(\cdot, \cdot)$ between two values of the covariates $\boldsymbol{X}$.

6.5 Example. A commonly used distance measure in matching is the Mahalanobis distance:

$$d_{\mathrm{MA}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \hat{\boldsymbol{\Sigma}} (\boldsymbol{x} - \tilde{\boldsymbol{x}}), \tag{6.3}$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of the covariance matrix of $\boldsymbol{X}_i$ within a treatment group. For example,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \Big[\sum_{i=1}^{n_1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_1)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_1)^T + \sum_{i=n_1+1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_0)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_0)^T \Big],$$

where $\bar{\boldsymbol{X}}_1 = \sum_{i=1}^{n_1} \boldsymbol{X}_i / n_1$ and $\bar{\boldsymbol{X}}_0 = \sum_{i=n_1+1}^{n} \boldsymbol{X}_i / n_0$ are the treated and control sample means.

Nearest-neighbour matching

Given the distance measure d, this naive method matches a treated observation $1 \leq i \leq n_1$ with its nearest control observation,

$$j = \arg\min_{A_j = 0} \{d(\boldsymbol{X}_i, \boldsymbol{X}_j)\}.$$

The problem with this method is that one control individual could be matched to several treated individuals, which never happens in a pairwise randomised experiment.

We can fix this problem by a greedy algorithm that sequentially matches a treated i to its nearest control neighbor that has yet been selected. A drawback is that the result will then depend on the order of the input.

Optimal matching

An improvement is the *optimal matching* that solves the following optimisation problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n_1} d\Big(\boldsymbol{X}_i, \sum_{j=n_1+1}^{n} M_{ij} \boldsymbol{X}_j\Big) \\ \text{subject to} \quad & M_{ij} \in \{0,1\},\ \forall 1 \leq i \leq n_1, n_1+1 \leq j \leq n, \\ & \sum_{j=1}^{n_0} M_{ij} = 1,\ \forall 1 \leq i \leq n_1, \\ & \sum_{i=1}^{n_1} M_{ij} \leq 1,\ \forall n_1+1 \leq j \leq n, \end{aligned} \tag{6.4}$$

where M_{ij} is an indicator for the treated observation i being mached to the control observation j. The last two constraints mean that every treated is matched to exactly one control and every control is matched to at most one treated.

Although combinatorial optimisation is generally NP-complete, the optimal matching problem (6.4) can be recasted as a network flow problem and solved efficiently in polynomial time.[4]

Propensity score matching

When matching was first developed in the 1980s, computing power was limited and it was often desirable to reduce the dimension of $\boldsymbol{X}$ before running a matching algorithm.

We say $b(\boldsymbol{X})$ is a *balancing score* if $A \perp\!\!\!\perp \boldsymbol{X} \mid b(\boldsymbol{X})$, that is, given $b(\boldsymbol{X})$, the covariates $\boldsymbol{X}$ have the same conditional distribution (are "balanced") in different treatment groups. If $b(\boldsymbol{X})$ is a balancing score, then the no unmeasured confounders assumption implies that

$$A \perp\!\!\!\perp Y(a) \mid b(\boldsymbol{X}),\ \text{for } a = 0, 1. \tag{6.5}$$

Among all the balancing scores, of particular interest is the *propensity score*[5], defined as

$$\pi(\boldsymbol{x}) = \mathbb{P}(A = 1 \mid \boldsymbol{X} = \boldsymbol{x}).$$

6.6 Exercise. Prove (6.5), then show that the propensity score is a balancing score. Furthermore, show that $\pi(\boldsymbol{X})$ can be written as a function of any balancing score $b(\boldsymbol{X})$.

The propensity score can be estimated from the observational data, commonly by fitting a logistic regression of A_i on $\boldsymbol{X}_i$. Let the estimated propensity score for individual i be $\hat{\pi}(\boldsymbol{X}_i)$. A popular distance measure is the squared distance between the estimated propensity scores in the logit scale:

$$d_{\text{PS}}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \left[\log\left(\frac{\hat{\pi}(\boldsymbol{X}_i)}{1 - \hat{\pi}(\boldsymbol{X}_i)}\right) - \log\left(\frac{\hat{\pi}(\boldsymbol{X}_j)}{1 - \hat{\pi}(\boldsymbol{X}_j)}\right)\right]^2.$$

Propensity score caliper

The distance measure d can be freely modified according to the problem. As an example, we may use the Mahalanobis distance with a propensity score caliper:

$$d(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} d_{\text{MA}}(\boldsymbol{X}_i, \boldsymbol{X}_j) & \text{if } d_{\text{PS}}(\boldsymbol{X}_i, \boldsymbol{X}_j) \le \tau^2, \\ \infty & \text{otherwise,} \end{cases}$$

where $\tau > 0$ is some tuning parameter. In this case, an treated observation is only allowed to be matched to a control observation whose propensity score is no more different than τ^2 (in the logit scale).

6.4 Covariate balance

Recall the logic of randomised experiments (Section 6.1) is that randomisation allows us to choose between causality and statistical error. This is reasonable because randomisation balances all pre-treatment covariates in simple Bernoulli trials and pairwise/stratified experiments. In other words, all pre-treatment covariates—measured or unmeasured—have the same distribution in the treated and control groups. Therefore, we cannot attribute any difference in the outcome (beyond some statistical error) to systematic differences in the covariates.

Following this logic, we can assess whether the matching is satisfactory by checking covariate balance. A common measure of covariate imbalance is the standardised covariate differences,[6]

$$B_k(\boldsymbol{M}) = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}\left(X_{ik} - \sum_{j=n_1+1}^{n} M_{ij}X_{jk}\right)}{\sqrt{(s_{1k}^2 + s_{0k}^2)/2}}, \ k = 1, \dots, p,$$

where X_{ik} is the kth covariate for the ith observation and s_{1k}^2 and s_{0k}^2 are the sample variances of X_k in the treated and control groups before matching.

A rule of thumb is that the k-th covariate X_k is considered approximately balanced if $|B_k| < 0.1$, but obviously we would like the entire vector $\boldsymbol{B}$ to be as close to $\boldsymbol{0}$ as possible.

If the covariate balance is unsatisfactory, a common practice is to rerun the matching algorithm with a different distance measure or remove treated units that have extreme propensity scores. This is often called the “propensity score tautology”[7]. In modern optimal matching algorithms, it is possible to include $|B_k(\boldsymbol{M})| \le \eta$, $\forall k$ as a constraint in the combinatorial optimisation problem.

6.5 Randomisation inference

Next we consider the statistical inference after matching.

To simplify the exposition, we assume treated observation i is matched to control observation $i + n_1$, $i = 1, \dots, n_1$. Let

$$D_i = (A_i - A_{i+n_1})(Y_i - Y_{i+n_1}),\ i = 1, \dots, n_1,$$

be the treated-minus-control difference in pair i. Let

$$M = \{\boldsymbol{a}_{[2n_1]} \in \{0,1\}^{2n_1} \mid a_i + a_{i+n_1} = 1, \forall i \in [n_1]\}$$

be all the treatment assignments such that within a matched pair, exactly one observation receives the treatment. Let $\boldsymbol{C}_i = (\boldsymbol{X}_i, Y_i(0), Y_i(1))$.

There are two ways to proceed from here. The first approach is to use the sample average of D_i,

$$\bar{D} = \frac{1}{n_1} \sum_{i=1}^{n_1} D_i$$

to estimate $\mathbb{E}[Y(1) - Y(0) \mid A = 1]$, the average treatment effect on the treated (ATT). We will not say more about this estimator other than it is commonly used in practice but its statistical inference is not as straightforward as one might imagine.[8]

The second and perhaps more interesting approach is to use a randomisation test to mimic what is done for randomised experiments in Section 2.4. The next assumption mimics Example 2.4.

6.7 Assumption. We assume matching reconstructs a pairwise randomised experiment, so

$$\mathbb{P}\left(\boldsymbol{A}_{[2n_1]} = \boldsymbol{a} \,\middle|\, \boldsymbol{C}_{[2n_1]}, \boldsymbol{A}_{[2n_1]} \in M\right) = \begin{cases} 2^{-n_1}, & \text{if } \boldsymbol{a} \in M, \\ 0, & \text{otherwise.} \end{cases}$$

This assumption is satisfied if there are no unmeasured confounders and the (true) propensity scores are exactly matched.

6.8 Proposition. *Suppose the data are i.i.d and Assumption 6.4 is satisfied. Then Assumption 6.7 holds if $\pi(\boldsymbol{X}_i) = \pi(\boldsymbol{X}_{i+n_1})$ for all $i \in [n_1]$.*

6.9 Exercise. Prove Proposition 6.8

Assumption 6.7 allows us to apply the randomisation test described in Section 2.4. Because the observations are matched, it is common to construct test statistics based on $\boldsymbol{D}_{[n_1]}$.

Consider the sharp null hypothesis $H_0 : Y_i(1) - Y_i(0) = \beta,\ \forall i$, where β is given. Under H_0 and by using the consistency assumption (Assumption 2.6), the counterfactual values of $\boldsymbol{D}_{[n_1]}$ can be imputed as

$$D_i(\boldsymbol{a}_{[2n_1]}) = (a_i - a_{i+n_1}) \cdot (Y_i(a_i) - Y_{i+n_1}(a_{i+n_1})) = \begin{cases} D_i, & \text{if } a_i = 1, a_{i+n_1} = 0, \\ 2\beta - D_i, & \text{if } a_i = 0, a_{i+n_1} = 1. \end{cases} \tag{6.6}$$

Consider any test statistic $T = T(\boldsymbol{D}_{[n_1]})$. Next we construct a randomisation test based on the randomisation distribution of $T(\boldsymbol{D}_{[n_1]}(\boldsymbol{A}_{[2n_1]}))$.

Let $F(t)$ denote its cumulative distribution function given $\boldsymbol{C}_{[2n_1]}$ and $\boldsymbol{A}_{[2n_1]} \in M$ under H_0,

$$\begin{aligned} F\big(t; \boldsymbol{D}_{[n_1]}, \beta\big) &= \mathbb{P}\left(T \le t \mid \boldsymbol{C}_{[n_1]}, \boldsymbol{A}_{[2n_1]} \in M, H_\beta\right), \\ &= \sum_{\boldsymbol{a}_{[2n_1]} \in M} \left(\frac{1}{2}\right)^{n_1} \cdot I\Big(T\big(\boldsymbol{D}_{[n_1]}(\boldsymbol{a}_{[2n_1]})\big) \le t\Big). \end{aligned} \tag{6.7}$$

We then compute the *p-value* for this randomisation test as $P_2 = F(T)$ and reject the hypothesis H_β if P_2 is less than a significance threshold $0 < \alpha < 1$. Following the same argument as in the proof of Theorem 2.19, this is valid test of H_0.

6.10 Theorem. *Under Assumptions 2.6 and 6.7, $\mathbb{P}(P_2 \le \alpha) \le \alpha$ under H_0 for all $0 < \alpha < 1$.*

6.11 Example (Signed score statistic)**.** Let $\psi : [0,1] \to \mathbb{R}^+$ be a positive function on the unit interval. The *signed score statistic* is defined as

$$T_\psi(\boldsymbol{D}_{[n_1]}) = \sum_{i=1}^{n_1} \operatorname{sgn}(D_i)\psi\Big(\frac{\operatorname{rank}(|D_i|)}{n_1 + 1}\Big), \tag{6.8}$$

where sgn is the sign function and $\operatorname{rank}(|D_i|)$ is the rank of the absolute difference $|D_i|$ among $|D_1|, \dots, |D_{n_1}|$. The widely used Wilcoxon signed rank statistic corresponds to the choice $\psi(t) = (n_1 + 1)t$.

6.12 Remark. We have been following the second approach of randomisation test in Section 2.4. We can also follow the first approach by considering the distribution of $T(\boldsymbol{A}_{[2n_1]}, \boldsymbol{Y}_{[2n_1]}(0))$, although this is less intuitive because $\boldsymbol{A}_{[2n_1]} = \boldsymbol{0}$ is not in the allowed set M.

6.13 Exercise. Consider the signed score statistic in (6.8) (treated as a function of $\boldsymbol{A}_{[2n_1]}$, $\boldsymbol{Y}_{[2n_1]}$). Derive the randomisation test based on the randomisation distribution of

$T(\boldsymbol{A}_{[2n_1]}, \boldsymbol{Y}_{[2n_1]}(0))$ and show that, given H_0 and conditioning on $\boldsymbol{C}_{[2n_1]}$,

$$T\big(\boldsymbol{A}_{[2n_1]}, \boldsymbol{Y}_{[2n_1]}(0)\big) \mid \boldsymbol{A}_{[2n_1]} \in M \overset{d}{=} \sum_{i=1}^{n_1} S_i \psi\Big(\frac{\mathrm{rank}(|Y_i(0) - Y_{n_1+i}(0)|)}{n_1 + 1}\Big), \tag{6.9}$$

where $S_i = (A_i - A_{i+n_1}) \cdot \mathrm{sgn}(Y_i(0) - Y_{i+n_1}(0)) \sim \mathrm{Bernoulli}(1/2)$. Justify this test using the symmetry of $D_i - \beta$ under Assumption 6.7 and H_0. Establish the equivalence between P_1 and P_2 (see also Exercise 2.21).

Equation (6.9) is indeed the more commonly used test because it is more computationally friendly. (Why?)

Notes

[1] Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*(3), 808–840. doi:10.1214/08-aoas187.

[2] Zhao, Q., Keele, L. J., & Small, D. S. (2019). Comment: Will competition-winning methods for causal inference also succeed in practice? *Statistical Science*, *34*(1), 72–76. doi:10.1214/18-sts680.

[3] Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. doi:10.1214/09-sts313.

[4] Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, *7*(1), 143–176. doi:10.1146/annurev-statistics-031219-041058.

[5] This was proposed in one of the most cited statistics paper by Rosenbaum and Rubin, 1983

[6] Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. doi:10.1080/00031305.1985.10479383.

[7] Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502. doi:10.1111/j.1467-985x.2007.00527.x.

[8] Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*(1), 235–267. doi:10.1111/j.1468-0262.2006.00655.x.

Chapter 7

No unmeasured confounders: Semiparametric inference

In this Chapter, we extend the super-population inference in randomised experiments (Section 2.5) to observational studies. The main challenge is that the treatment assignment mechanism is now unknown and must be estimated.

7.1 An introduction to semiparametric inference

As in the Chapter, we will focus on the case of a binary treatment A and assume $(\boldsymbol{X}_i, A_i, Y_i(0), Y_i(1)),\ i \in [n]$ are iid.

We maintain the assumptionss of no unmeasured confounders (Assumption 6.4), consistency (Assumption 2.6), and positivity:

7.1 Assumption (Positivity)**.** $\pi_a(\boldsymbol{x}) = \mathbb{P}(A = a \mid \boldsymbol{X} = \boldsymbol{x}) > 0, \forall a, \boldsymbol{x}$.

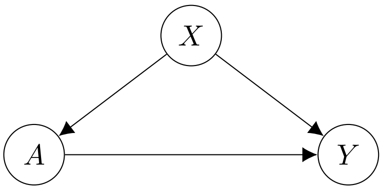


Figure 7.1: An observational study with no unmeasured confounders.

7.2 Remark. We have seen in Chapter 5 that no unmeasured confounders and consistency necessarily follow from assuming the single-world causal model corresponding to Figure 7.1. Assumption 7.1 is also called the *overlap* assumption, because by the Bayes rule, it is equivalent to assuming that the distribution $\boldsymbol{X}$ has the same support given $A = a$ for all a.

By Theorem 2.12, we have

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\big\{\mathbb{E}[Y \mid A = 1, \boldsymbol{X}] - \mathbb{E}[Y \mid A = 0, \boldsymbol{X}]\big\}. \tag{7.1}$$

Our goal is estimate the right hand side of the above quation, a functional of the observed data distribution.

This is where semiparametric inference becomes useful. A *semiparametric model* is a statistical model with parametric and nonparametric components.

7.3 Example. An example of semiparametric model is the partially linear model

$$\mathbb{E}[Y \mid A, \boldsymbol{X}] = \beta \cdot A + g(\boldsymbol{X}),$$

where β and $g(\cdot)$ are unknown. Other well known examples include single index models, varying coefficient models, and Cox's proportional hazards model in survival analysis.

Semiparametric inference is mainly concerned with estimating and making inference for the (finite-dimensional) parametric component. This is called the *parameter of interest*, and the (infinite-dimensional) nonparametric component is called the *nuisance parameter*.

An alternative setup is to consider estimating a functional $\beta(P)$ using an iid sample $\boldsymbol{D}_1, \dots, \boldsymbol{D}_n$ from the distribution P that is known to belong to a set $\mathcal{P}$ of probability distributions. This is very general and is well suited for causal inference problems: the causal identification theory (Section 5.4) often equates a causal effect of interest with a functional of the observed variables; an example is (7.1).

Formally deriving the semiparametric inference theory is beyond the scope of this course.[1] Below we will just informally describe some key results in this theory.

Roughly speaking, semiparametric inference provides a theory for well-behaved (so-called regular) estimators[2] that admit the so-called *asymptotic linear* expansion:

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_\beta(\boldsymbol{D}_i) + o_p(1), \tag{7.2}$$

where the *influence function* $\psi_\beta(\cdot)$ has mean 0 and finite variance, and $o_p(1)$ means that the residual converges to 0 in probability as $n \to \infty$.

The asymptotic linearity (7.2) implies that $\hat{\beta}$ has an asymptotic normal distribution

$$\sqrt{n}(\hat{\beta} - \beta) \overset{d}{\to} \mathrm{N}(0, \mathrm{Var}(\psi_\beta\{\boldsymbol{D})\}).$$

The influence function that has the smallest variance, $\psi_{\beta,\text{eff}}(\cdot)$ is called the *efficient influence function.*

Semiparametric inference theory gives a geometric characterisation of the space of influence functions. A key conclusion is that $\psi_{\beta,\text{eff}}(\cdot)$ can be obtained by projecting any influence function onto the so-called *tangent space* consisting of all score functions of the model. In consequence, $\mathrm{Var}(\psi_{\beta,\text{eff}}(\boldsymbol{D})) \leq \mathrm{Var}(\psi_\beta(\boldsymbol{D}))$. This generalises the Cràmer-Rao lower bound and asymptotic efficiency of the maximum likelihood estimator in parametric models to semiparametric models.

7.4 Example. Following the proof of Lemma 2.26, a Z-estimator is generally asymptotic linear and its influence function is given in (2.20).

7.5 Exercise. Derive the influence function for the regression estimator $\hat{\beta}_1$ in (2.11). Veryify that it has mean 0.

7.2 Discrete covariates

The semiparametric inference theory is very general and abstract. To obtain some intuitions for our problem, we first consider the case of discrete covariates $\boldsymbol{X}$ and the estimation of

$$\beta_a = \mathbb{E}\{\mathbb{E}[Y \mid A = a, \boldsymbol{X}]\} = \sum_{\boldsymbol{x}} \mu_a(\boldsymbol{x})\,\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}),\ a = 0, 1 \tag{7.3}$$

where $\mu_a(\boldsymbol{x}) = \mathbb{E}[Y_i \mid A_i = a, \boldsymbol{X}_i = \boldsymbol{x}]$. The ATE can be written as $\beta = \beta_1 - \beta_0$.

Given an iid sample from this population, we can empirically estimate the quantities in (7.3) by

$$\hat{\mu}_a(\boldsymbol{x}) = \frac{\sum_{i=1}^n I(A_i = a, \boldsymbol{X}_i = \boldsymbol{x}) Y_i}{\sum_{i=1}^n I(A_i = a, \boldsymbol{X}_i = \boldsymbol{x})},$$

$$\hat{\mathbb{P}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{X}_i = \boldsymbol{x}).$$

The first estimator is well defined if the denominator is non-zero, an event with probability tending to 1 as $n \to \infty$. By plugging these into (7.3), we obtain the *outcome regression* (OR) estimator

$$\hat{\beta}_{a,\mathrm{OR}} = \sum_{\boldsymbol{x}} \hat{\mu}_a(\boldsymbol{x})\hat{\mathbb{P}}(\boldsymbol{X}_i = \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \sum_{\boldsymbol{x}} \hat{\mu}_a(\boldsymbol{x}) I(\boldsymbol{X}_i = \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n \hat{\mu}_a(\boldsymbol{X}_i). \tag{7.4}$$

The ATE can be subsequently estimated by

$$\hat{\beta}_{\mathrm{OR}} = \hat{\beta}_{1,\mathrm{OR}} - \hat{\beta}_{0,\mathrm{OR}} = \frac{1}{n}\sum_{i=1}^n \hat{\mu}_1(\boldsymbol{X}_i) - \hat{\mu}_0(\boldsymbol{X}_i). \tag{7.5}$$

7.6 Remark. Since (7.4) does not depend on the form of $\hat{\mu}_a(\boldsymbol{x})$, this outcome estimator can be easily extended to the continuous $\boldsymbol{X}$ case.

Next, we analyse the asymptotic behaviour of $\hat{\beta}_{\mathrm{OR}}$ with discrete $\boldsymbol{X}$. This does not follow trivially from the central limit theorem because the summands in (7.4) are not independent. To solve this problem, we derive an alternative representation of the outcome regression estimator.

Recall $\pi_a(\boldsymbol{x}) = \mathbb{P}(A = a \mid \boldsymbol{X} = \boldsymbol{x})$, which can be estimated by

$$\hat{\pi}_a(\boldsymbol{x}) = \frac{\sum_{i=1}^n I(A_i = a, \boldsymbol{X}_i = \boldsymbol{x})}{\sum_{i=1}^n I(\boldsymbol{X}_i = \boldsymbol{x})}.$$

Because A is binary, $\pi_0(\boldsymbol{x}) = 1 - \pi_1(\boldsymbol{x})$. Note that $\pi_1(\boldsymbol{x}) = \pi(\boldsymbol{x})$, the propensity score defined in Section 6.3.

The *inverse probability weighted* (IPW) estimator[3] is given by

$$\hat{\beta}_{a,\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^n \frac{I(A_i = a)}{\hat{\pi}_a(\boldsymbol{X}_i)} Y_i. \tag{7.6}$$

The name is derived from the form of the estimator. The average treatment effect can be subsequently estimated by

$$\hat{\beta}_{\text{IPW}} = \hat{\beta}_{1,\text{IPW}} - \hat{\beta}_{0,\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{A_i}{\hat{\pi}(\boldsymbol{X}_i)} - \frac{1-A_i}{1-\hat{\pi}(\boldsymbol{X}_i)}\Big]Y_i. \tag{7.7}$$

7.7 Proposition. *Suppose $\boldsymbol{X}$ is discrete and $\hat{\pi}_a(\boldsymbol{x}) > 0$ for all a and $\boldsymbol{x}$. Then $\hat{\beta}_{a,OR} = \hat{\beta}_{a,IPW}$, $a = 0, 1$ and $\hat{\beta}_{OR} = \hat{\beta}_{IPW}$.*

Notice that the summands in (7.7) are still not independent. To break through this impasse, a key observation is that

7.8 Lemma. *Under the sample assumptions in Proposition 7.7, the identity*

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i=a)}{\hat{\pi}_a(\boldsymbol{X}_i)}\mu(\boldsymbol{X}_i) = \frac{1}{n}\sum_{i=1}^{n}\mu(\boldsymbol{X}_i),\ a = 0, 1.$$

holds for any function $\mu(\cdot)$.

Intuitively, Lemma 7.8 says that the distribution of $\boldsymbol{X}$ in the entire sample can be obtained from the $A = a$ subsample by inverse probability weighting.

7.9 Exercise. Prove Proposition 7.7 and Lemma 7.8.

Using Lemma 7.8, we have, by adding and subtracting $\mu_a(\boldsymbol{X})$,

$$\begin{aligned}&\sqrt{n}\big(\hat{\beta}_{a,\text{IPW}} - \beta_a\big)\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{I(A_i=a)}{\hat{\pi}_a(\boldsymbol{X}_i)}[Y_i - \mu_a(\boldsymbol{X}_i)] + \mu_a(\boldsymbol{X}_i) - \beta_a\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\frac{I(A_i=a)}{\pi_a(\boldsymbol{X}_i)}[Y_i - \mu_a(\boldsymbol{X}_i)] + \mu_a(\boldsymbol{X}_i) - \beta_a\Big\} + R_n.\end{aligned}$$

The residual term

$$R_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(A_i = a)\Big[\frac{1}{\hat{\pi}_a(\boldsymbol{X}_i)} - \frac{1}{\pi_a(\boldsymbol{X}_i)}\Big][Y_i - \mu_a(\boldsymbol{X}_i)] \xrightarrow{p} 0$$

as $n \to \infty$. This is because $\hat{\pi}_a(\boldsymbol{x})$ generally converges to $\pi_a(\boldsymbol{x})$ at $1/\sqrt{n}$ rate and the other term $I(A_i = a)[Y_i - \mu_a(\boldsymbol{X}_i)]$ is iid with mean 0. This shows that $\hat{\beta}_{a,\text{IPW}}$ admits the asymptotic linear expansion in (7.2) with the influence function

$$\psi_{\beta_a}(\boldsymbol{D}) = \frac{I(A=a)}{\pi_a(\boldsymbol{X})}[Y - \mu_a(\boldsymbol{X})] + \mu_a(\boldsymbol{X}) - \beta_a.$$

The asymptotic results are summarised in the next Theorem.

7.10 Theorem. *Suppose the positivity assumption (Assumption 7.1) is given. Under iid sampling with discrete $\boldsymbol{X}$ and regularity conditions, we have*

$$\sqrt{n}\big(\hat{\beta}_{a,OR} - \beta_a\big) \xrightarrow{d} \mathrm{N}\Big(0, \mathrm{Var}\left(\psi_{\beta_a}(\boldsymbol{D}_i)\right)\Big),\ a = 0, 1,$$

and

$$\sqrt{n}\big(\hat{\beta}_{OR} - \beta\big) \xrightarrow{d} \mathrm{N}\Big(0, \mathrm{Var}\left(\psi_{\beta_1}(\boldsymbol{D}_i) - \psi_{\beta_0}(\boldsymbol{D}_i)\right)\Big).$$

7.11 Remark. In the discrete $\boldsymbol{X}$ case, $\psi_{\beta_a}(\boldsymbol{D})$ is the only influence function for estimating β_a because we are considering the nonparametric model and the tangent space contains all square-integrable functions with mean 0. As a consequence, $\psi_{\beta_a}(\boldsymbol{D})$ is also the efficient influence function. This last conclusion is still true when $\boldsymbol{X}$ contains continuous covariates, although there are many other possible influence functions.[4]

7.12 Exercise. Derive the same results for estimating the average treatment effect on the treated ATT $= \mathbb{E}[Y(1) - Y(0) \mid A = 1]$

(i) Show that

$$\mathbb{E}[Y(0) \mid A = 1] = \frac{\mathbb{E}[\mu_0(\boldsymbol{X})\pi(\boldsymbol{X})]}{\mathbb{P}(A = 1)} := \beta_{0|1}.$$

(ii) Let $n_1 = \sum_{i=1}^n I(A_i = 1)$. Show that the OR estimator of $\beta_{0|1}$ is given by

$$\hat{\beta}_{0|1,\mathrm{OR}} = \frac{1}{n_1}\sum_{i=1}^n I(A_i = 1)\hat{\mu}_0(\boldsymbol{X}_i),$$

the IPW estimator of $\beta_{0|1}$ is given by

$$\hat{\beta}_{0|1,\mathrm{IPW}} = \frac{1}{n_1}\sum_{i=1}^n I(A_i = 0)\frac{\hat{\pi}(\boldsymbol{X}_i)}{1 - \hat{\pi}(\boldsymbol{X}_i)}Y_i,$$

and $\hat{\beta}_{0|1,\mathrm{OR}} = \hat{\beta}_{0|1,\mathrm{IPW}}$.

(iii) Show that

$$\begin{aligned}&\sqrt{n_1}\big(\hat{\beta}_{0|1,\mathrm{OR}} - \beta_{0|1}\big)\\ =&\frac{1}{\sqrt{n_1}}\sum_{i=1}^n(1 - A_i)\Big[\frac{\pi(\boldsymbol{X}_i)}{1 - \pi(\boldsymbol{X}_i)}\big(Y_i - \mu_0(\boldsymbol{X}_i)\big)\Big] + A_i\big[\mu_0(\boldsymbol{X}_i) - \beta_{0|1}\big] + o_p(1).\end{aligned}$$

(iv) Complete the asymptotic theory for estimating ATT with discrete $\boldsymbol{X}$.

7.3 Outcome regression and inverse probability weighting

When $\boldsymbol{X}$ contains continuous covariates, Remark 7.6 suggests that the OR and IPW estimators can still be applied by plugging in empirical estimates of the *nuisance parameters*

$\mu_a(\boldsymbol{x}) = \mathbb{E}[Y \mid A = a, \boldsymbol{X} = \boldsymbol{x}]$ and $\pi_a(\boldsymbol{x}) = \mathbb{P}(A = a \mid \boldsymbol{X} = \boldsymbol{x})$. However, unlike the discrete $\boldsymbol{X}$ case, the OR and IPW estimators are generally different.

Intuitively, these estimators are reasonable because of the following dual representation of β_a:

7.13 Proposition. *Under the positivity assumption (Assumption 7.1), we have*

$$\beta_a = \mathbb{E}[\mu_a(\boldsymbol{X})] = \mathbb{E}\left[\frac{I(A=a)}{\pi_a(\boldsymbol{X})}Y\right], \ a = 0, 1,$$

thus

$$\beta = \mathbb{E}[\mu_1(\boldsymbol{X}) - \mu_0(\boldsymbol{X})] = \mathbb{E}\left[\frac{A}{\pi(\boldsymbol{X})}Y - \frac{1-A}{1-\pi(\boldsymbol{X})}Y\right].$$

7.14 Exercise. Prove Proposition 7.13.

Thus, the asymptotic behaviour of the OR and IPW estimators crucially depend on how well $\mu_a(\boldsymbol{x})$ and $\pi(\boldsymbol{x})$ are estimated. When $\mu_a(\boldsymbol{x})$ is modelled parametrically, $\hat{\beta}_{a,\mathrm{OR}}$ is generally consistent if the model is correctly specified. The same conclusion holds for $\hat{\beta}_{a,\mathrm{IPW}}$ and modelling $\pi_a(\boldsymbol{x})$.

7.15 Example. In practice, a common model for $\mu_a(\boldsymbol{x})$ is the linear regression

$$\mu_a(\boldsymbol{x}; \boldsymbol{\eta}_\mu) = \gamma_{\mu_a} + \boldsymbol{\delta}_{\mu_a}^T \boldsymbol{x}, \ \text{for } a = 0, 1,$$

where $\boldsymbol{\eta}_\mu = (\gamma_{\mu_0}, \boldsymbol{\delta}_{\mu_0}, \gamma_{\mu_1}, \boldsymbol{\delta}_{\mu_1})$. A common model for $\pi(\boldsymbol{x}) = \pi_1(\boldsymbol{x})$ is the logistic regression

$$\pi(\boldsymbol{x}; \boldsymbol{\eta}_\pi) = \frac{\exp(\gamma_\pi + \boldsymbol{\delta}_\pi^T \boldsymbol{x})}{1 + \exp(\gamma_\pi + \boldsymbol{\delta}_\pi^T \boldsymbol{x})},$$

where $\boldsymbol{\eta}_\pi = (\gamma_\pi, \boldsymbol{\delta}_\pi)$. The parameters $\boldsymbol{\eta}_\mu$ and $\boldsymbol{\eta}_\pi$ are usually estimated by maximum likelihood in the corresponding generalised linear models (Y is normally distributed and A is Bernoulli). When the regression models are correctly specified, $\mu_a(\boldsymbol{x}; \hat{\boldsymbol{\eta}}_\mu)$ and $\pi(\boldsymbol{x}; \hat{\boldsymbol{\eta}}_\pi)$ generally converge to $\mu_a(\boldsymbol{x})$ and $\pi(\boldsymbol{x})$. However, when the regression models are incorrectly specified, they only converge to the best parametric approximations to the true $\mu_a(\boldsymbol{x})$ and $\pi(\boldsymbol{x})$.

7.16 Remark. To reduce the modelling bias (recall (6.1)), we can estimate the nuisance parameters $\mu_a(\cdot), \pi_a(\cdot), a = 0, 1$ using more flexible models. Many researchers have suggested to use machine learning methods to estimate the nuisance parameters in hope that they can adapt to unspecified patterns in the data. These methods indeed perform much better than using traditional statistical models in simulated datasets[5], but the discrepancy tends to be much smaller in real datasets[6].

7.4 Doubly robust estimator

Next we will combine the OR estimator and the IPW estimator to obtain a more efficient and robust estimator. The idea is to use the efficient influence function $\psi_{\beta_a}(\boldsymbol{D})$ derived in Section 7.2, which can be written as

$$\psi_{\beta_a}(\boldsymbol{D}; \mu_a, \pi_a) = m_a(\boldsymbol{D}; \mu_a, \pi_a) - \beta_a,$$

where

$$m_a(\boldsymbol{D}; \eta) = \frac{I(A = a)}{\pi_a(\boldsymbol{X})}\big(Y - \mu_a(\boldsymbol{X})\big) + \mu_a(\boldsymbol{X}). \tag{7.8}$$

Let $\hat{\mu}_a(\cdot)$ be an estimator of $\mu_a(\cdot)$ and $\hat{\pi}_a(\cdot)$ be an estimator of $\pi_a(\cdot)$. Because influence functions have mean 0, the above representation motivates the estimator

$$\hat{\beta}_{a,\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n} m_a(\boldsymbol{D}_i, \hat{\mu}_a, \hat{\pi}_a), \tag{7.9}$$

and $\hat{\beta}_{\mathrm{DR}} = \hat{\beta}_{1,\mathrm{DR}} - \hat{\beta}_{0,\mathrm{DR}}$.

The efficient influence function has the following appealing property.

7.17 Proposition (Double robustness). *Under positivity (Assumption 7.1), we have, for any functions $\tilde{\mu}_a(\cdot)$ and $\tilde{\pi}(\cdot)$,*

$$0 = \mathbb{E}[\psi_{\beta_a}(\boldsymbol{D}; \mu_a, \pi_a)] = \mathbb{E}[\psi_{\beta_a}(\boldsymbol{D}; \tilde{\mu}, \pi_a)] = \mathbb{E}[\psi_{\beta_a}(\boldsymbol{D}; \mu_a, \tilde{\pi})].$$

7.18 Exercise. Prove Proposition 7.17.

In other words, we have

$$\beta_a = \mathbb{E}[m_a(\boldsymbol{V}_i; \mu_a, \pi)] = \mathbb{E}[m_a(\boldsymbol{V}_i; \mu_a, \tilde{\pi})] = \mathbb{E}[m_a(\boldsymbol{V}_i; \tilde{\mu}_a, \pi)].$$

This shows that if either $\mu_a(\cdot)$ or $\pi_a(\cdot)$ is consistently estimated, the estimator $\hat{\beta}_{a,\mathrm{DR}}$ is generally consistent for estimating β_a. This is why we call $\hat{\beta}_{a,\mathrm{DR}}$ the *doubly robust* estimator.

The doubly robust estimator is also useful when the nuisance parameters are estimated using flexible machine learning methods. To see this, we examine the residual term in the asymptotic linear expansion of $\hat{\beta}_{a,\mathrm{DR}}$:

$$\begin{aligned} R_n &= \sqrt{n}\big(\hat{\beta}_{a,\mathrm{DR}} - \beta_a\big) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\beta_a}(\boldsymbol{D}_i) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_a(\boldsymbol{D}_i; \hat{\mu}_a, \hat{\pi}_a) - m_a(\boldsymbol{D}_i; \mu_a, \pi_a). \end{aligned}$$

We cannot immediately conclude that $R_n \overset{p}{\to} 0$ using the law of large numbers, because the summands on the right hand side are not iid ($\hat{\mu}_a$ and $\hat{\pi}_a$ are obtained using $\boldsymbol{D}_{[n]}$).

There are two ways to resolve this issue. First, if the models we used for μ_a and π_a are not too complex (they are in the so-called Donsker function class)[7], one can deduce from the empirical process theory that the dependence of $\hat{\mu}_a$ and $\hat{\pi}_a$ on the data can be ignored:

$$R_n \approx \sqrt{n}\,\mathbb{E}_{\boldsymbol{D}_{n+1}}[m_a(\boldsymbol{D}_{n+1};\hat{\boldsymbol{\eta}}) - m_a(\boldsymbol{D}_{n+1};\boldsymbol{\eta})],$$

where $\mathbb{E}_{\boldsymbol{D}_{n+1}}$ indicates the expectation is taken over a new and independence observation $\boldsymbol{D}_{n+1}$.

Another approach is to use sample splitting, so that the nuisance parameters are estimated using an independent subsample.[8] This techniques allows us to avoid restricting the complexity of the nuisance models and becomes popular with machine learning methods (sometimes under the name "cross-fitting").

Using (7.8) and by taking expectation over A_{n+1}, Y_{n+1} given $\boldsymbol{X}_{n+1}$, it can be shown that

$$R_n \approx \sqrt{n}\,\mathbb{E}_{\boldsymbol{X}_{n+1}}\left[\frac{\{\hat{\pi}_a(\boldsymbol{X}_{n+1}) - \pi_a(\boldsymbol{X}_{n+1})\}\{\hat{\mu}_a(\boldsymbol{X}_{n+1}) - \mu_a(\boldsymbol{X}_{n+1})\}}{\hat{\pi}_a(\boldsymbol{X}_{n+1})}\right]. \tag{7.10}$$

In order words, the residual term R_n depends on a product of the estimation error of $\hat{\pi}$ and $\hat{\mu}_a$.

7.19 Exercise. Derive (7.10) and use it to (informally) prove the double robustness of $\hat{\beta}$.

Let's define the mean squared error (MSE) of $\hat{\mu}_a(\cdot)$ as

$$\text{MSE}\left(\hat{\mu}_a(\cdot)\right) = \mathbb{E}_{\boldsymbol{X}_{n+1}}\left[\left\{\hat{\mu}_a(\boldsymbol{X}_{n+1}) - \mu_a(\boldsymbol{X}_{n+1})\right\}^2\right].$$

Similarly, we may define the MSE of $\hat{\pi}_a(\boldsymbol{x})$. By applying the Cauchy-Scharwz inequality to (7.10), we obtain

$$\text{RHS of (7.10)} \leq \sqrt{n}\big[\sup_{\boldsymbol{x}} \hat{\pi}_a^{-1}(\boldsymbol{x})\big]\,\text{MSE}\left(\hat{\mu}_a(\cdot)\right) \cdot \text{MSE}\left(\hat{\pi}_a(\cdot)\right).$$

This result is summarised in the next Lemma.

7.20 Lemma. *Under i.i.d. sampling and mild regularity conditions, the above residual term* $R_n \overset{p}{\to} 0$ *if*

(i) There exists $C > 0$ *such that* $\mathbb{P}(\hat{\pi}_a(\boldsymbol{x}) \geq C, \forall \boldsymbol{x}) \to 1$ *as* $n \to \infty$*; and*

(ii) $\sqrt{n}\,\text{MSE}\left(\hat{\mu}_a(\cdot)\right) \cdot \text{MSE}\left(\hat{\pi}_a(\cdot)\right) \overset{p}{\to} 0$ *as* $n \to \infty$.

Suppose $\hat{\pi}_0(\boldsymbol{x}) = 1 - \hat{\pi}(\boldsymbol{x})$. By combining the previous results, we obtain the next Theorem.

7.21 Theorem (Semiparametric efficiency of the DR estimator)**.** *Under i.i.d. sampling and mild regularity conditions, suppose there exists $C > 0$ such that*

$$\mathbb{P}(C \leq \hat{\pi}(\boldsymbol{x}) \leq 1 - C, \forall \boldsymbol{x}) \to 1 \text{ as } n \to \infty. \tag{7.11}$$

Furthermore, suppose

$$\sqrt{n} \max \Big\{ \text{MSE}\left(\hat{\mu}_0(\boldsymbol{x})\right), \text{MSE}\left(\hat{\mu}_1(\boldsymbol{x})\right) \Big\} \cdot \text{MSE}\left(\hat{\pi}(\boldsymbol{x})\right) \xrightarrow{p} 0 \text{ as } n \to \infty. \tag{7.12}$$

Then the estimator $\hat{\beta}_{DR}$ satisfies

$$\sqrt{n}(\hat{\beta}_{DR} - \beta) \xrightarrow{d} \text{N}\Big(0, \text{Var}\left(\psi_{\beta_1}(\boldsymbol{D}_i) - \psi_{\beta_0}(\boldsymbol{D}_i)\right)\Big).$$

7.22 Remark. The condition (7.11) highlights the role of the positivity assumption and is needed because of the weighting by the inverse of $\hat{\pi}_a(\boldsymbol{X})$. It is satisfied, for example, if $\pi(\boldsymbol{x})$ is bounded away from 0 and 1 and $\hat{\pi}(\cdot)$ is consistent. The condition (7.12) is satisfied if both $\text{MSE}\left(\hat{\mu}_a(\cdot)\right)$ and $\text{MSE}\left(\hat{\pi}_a(\cdot)\right)$ are $o_p(n^{-1/4})$, for $a = 0, 1$.

7.5 A comparison of the statistical methods

We have covered several statistical methods for observational studies with no unmeasured confounders. Each method has its own strengths and weaknesses and may be preferrable in different practical problems.

Matching and randomisation inference

- Advantages: transparent; easy implementation; can incorporate prior knowledge; ensures well overlapping covariates.
- Disadvantages: Less efficient (though can be improved).

Inverse probability weighting

- Advantages: extends matching; generalisable to more complex problems[9]; can reach the semiparametric efficiency bound[10]; can be doubly robust[11].
- Disadvantages: can be unstable if the estimated probabilities are close to zero[12]; not robust to model misspecification.

Outcome regression

- Advantages: can reach the parametric Cràmer-Rao bound (smaller than the semiparametric bound); can easily incorporate machine learning methods.
- Disadvantages: not robust to model misspecification.

Doubly robust estimator

- Advantages: doubly robust; modelling bias is reduced; can reach the semiparametric efficiency bound.
- Disadvantages: can be unstable if the estimated probabilities are close to zero[13].

Notes

[1] For the general theory, see Van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge University Press, Chapter 25. For a less formal treatment with examples in causal inference, see Tsiatis, A. (2007). *Semiparametric theory and missing data.* Springer.

[2] The regularity condition is needed to rule out estimators (for example, Hodges' estimator) that are "super-efficient" at some parameter values but have erratic behaviours nearby. See Van der Vaart, 2000, Example 8.1.

[3] This is also called the Horvitz-Thompson estimator, which is first proposed in survey sampling; see Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. doi:10.1080/01621459.1952.10483446.

[4] Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*(429), 106–121. doi:10.1080/01621459.1995.10476493. See also Tsiatis, 2007, Chapters 7 and 13.

[5] Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43–68. doi:10.1214/18-sts667.

[6] Keele, L., & Small, D. (2018). Comparing covariate prioritization via matching to machine learning methods for causal inference using five empirical applications. arXiv: 1805.03743 `[stat.AP]`.

[7] Van der Vaart, 2000, Chapter 19.

[8] This idea is originally due to Hájek, J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, *33*(3), 1124–1147. doi:10.1214/aoms/1177704476. See also Van der Vaart, 2000, Section 25.8

[9] Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. doi:10.1097/00001648-200009000-00011, For example, see.

[10] Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. doi:10.1111/1468-0262.00442.

[11] Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, *47*(2), 965–993. doi:10.1214/18-aos1698.

[12] Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*(4), 523–539. doi:10.1214/07-sts227.

[13] Kang and Schafer, 2007.

Chapter 8

Sensitivity analysis

No unmeasured confounders, as introduced in Section 6.2, is a rather optimistic assumption in practice. Does a limited violation of this assumption render our statistical analysis useless? This is answered by a sensitivity analysis.

8.1 A roadmap

Broadly speaking, a sensitivity analysis consists of three steps:

(i) Model augmentation: Specify a family of distributions $\mathcal{F}_{\theta,\eta}$ for the full data $\boldsymbol{F}$ of all the relevant factuals and counterfactuals, where η is a sensitivity parameter. It is customary to let $\eta = 0$ corresponds to the case of no unmeasured confounders.

(ii) Statistical inference: Test a causal hypothesis regarding $\mathcal{F}_{\theta,\eta}$ or estimate a causal effect $\beta(\mathcal{F}_{\theta,\eta})$. This is usually done in one of the two following senses:

   (i) *Point identification:* The inference is valid at a given η;

   (ii) *Partially identifiation:* The inference is valid for all η in a given set H.

(iii) Interpretatation: Assess the strength of evidence by examining how sensitive the conclusions are to unmeasured confounders. This typically involves finding the "tipping point" η and make a heuristic interpretation.

Below we will give two sensitivity analysis methods that rely on different model augmentations and provide different statistical guarantees.

8.2 Rosenbaum's sensitivity analysis

We first describe a sensitivity analysis that applies to randomisation inference for matched observational studies. Recall the setting in Section 6.5 that observations $1, \dots, n_1$ are treated and matched to control observations $n_1 + 1, \dots, 2n_1$, respectively.

Suppose the data are iid and denote

$$\pi_i = \mathbb{P}(A_i = 1 \mid \boldsymbol{C}_i),\ i \in [2n_1],$$

where $\boldsymbol{C}_i = (\boldsymbol{X}_i, Y_i(0), Y_i(1))$.

The following sensitivity model is widely used in observational studies.[1]

8.1 Assumption (Rosenbaum's sensitivity model)**.** For a given value $\Gamma \geq 1$, we have

$$\frac{1}{\Gamma} \leq \frac{\pi_i/(1-\pi_i)}{\pi_{n_1+i}/(1-\pi_{n_1+i})} \leq \Gamma, \ \forall i \in [n_1]. \tag{8.1}$$

In fact, $\Gamma = 1$ recovers no unmeasured confounders (Assumption 6.7). This is because

$$\mathbb{P}\left(A_i = 1, A_{n_1+i} = 0 \,\middle|\, \boldsymbol{C}_{[2n_1]}, A_i + A_{n_1+i} = 1\right) = \frac{\pi_i(1-\pi_{n_1+i})}{\pi_i(1-\pi_{n_1+i}) + \pi_{n_1+i}(1-\pi_i)}.$$

(corrections from the lectures.) The odds ratio bound (8.1) further implies that

$$\frac{1}{1+\Gamma} \leq \mathbb{P}\left(A_i = 1, A_{n_1+i} = 0 \,\middle|\, \boldsymbol{C}_{[2n_1]}, A_i + A_{n_1+i} = 1\right) \leq \frac{\Gamma}{1+\Gamma}. \tag{8.2}$$

8.2 Remark. An alternative and prehaps more intuitive formulation of Rosenbaum's sensitivity model is the following. Suppose there exists an unmeasured confounder $U \in [0,1]$ so that $A \perp\!\!\!\perp \{Y(0), Y(1)\} \mid \boldsymbol{X}, U$. Then if we let $\pi_i = \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i, U_i)$, the sensitivity model (8.1) is equivalent to assuming the logistic regression model

$$\mathbb{P}(A = 1 \mid \boldsymbol{X}, U) = \operatorname{expit}(g(\boldsymbol{X}) + \gamma U), \ 0 \leq \gamma \leq \log \Gamma, \tag{8.3}$$

where $g(\cdot)$ is an arbitrary function and $\operatorname{expit}(\eta) = e^{\eta}/(1+e^{\eta})$.

8.3 Exercise. Show that (8.3) implies (8.1) if $\boldsymbol{X}_i = \boldsymbol{X}_{n_1+i}$.

Next we consider the randomisation distribution of the signed score statistic (6.8) under Rosenbaum's sensitivity model. Following Exercise 6.13, given H_0 and conditioning on $\boldsymbol{C}_{[2n_1]}$

$$T\big(\boldsymbol{A}_{[2n_1]}, \boldsymbol{Y}_{[2n_1]}(0)\big) \mid \boldsymbol{A}_{[2n_1]} \in M \ \overset{d}{=} \ \sum_{i=1}^{n_1} S_i \psi\Big(\frac{\operatorname{rank}(|Y_i(0) - Y_{n_1+i}(0)|)}{n_1+1}\Big),$$

where $S_i = (A_i - A_{n_1+i}) \cdot \operatorname{sgn}(Y_i(0) - Y_{n_1+i}(0))$.

A random variable X is said to *stochastically dominate* another random variable Y, written as $X \succeq Y$, if $\mathbb{P}(X > t) \geq \mathbb{P}(Y > t)$ for all t. The distribution of S_i given $Y_i(0), Y_{n_1+i}(0)$ is unkonwn, but by using (8.2), S_i stochastically dominates the following random variable

$$S_i^- = \begin{cases} -1, & \text{with probability } \Gamma/(1+\Gamma), \\ 1, & \text{with probability } 1/(1+\Gamma). \end{cases}$$

This can be used to obtain a (sharp) bound on the p-value:

8.4 Theorem. *Suppose Assumption 8.1 holds and we are using the signed score statistic* (6.8)*. Given* H_0*,*

$$T\big(\boldsymbol{A}_{[2n_1]}, \boldsymbol{Y}_{[2n_1]}(0)\big) \succeq \sum_{i=1}^{n_1} S_i^- \psi\Big(\frac{rank(|D_i - \beta|)}{n_1 + 1}\Big). \tag{8.4}$$

Proof. This Theorem follows from noticing $|D_i - \beta| = |Y_i(0) - Y_{n_1+i}(0)|$ and the following property of stochastic ordering: If $X_i \succeq Y_i$ for $i \in [n]$ and $X_i \perp\!\!\!\perp X_j, Y_i \perp\!\!\!\perp Y_j$ for all $i \neq j$, then $\sum_{i=1}^n X_i \succeq \sum_{i=1}^n Y_i$. □

Theorem 8.4 allows us to upper bound the randomisation p-value in Rosenbaum's sensitivity model. Let $F_1^-(\cdot)$ denote the cumulative distribution function of the right hand side of (8.4) given $\boldsymbol{C}_{[2n_1]}$. Let $P_1^- = F_1^-(T(\boldsymbol{A}_{[2n_1]}, \boldsymbol{Y}_{[2n_1]} - \beta \boldsymbol{A}_{[2n_1]}))$. Then under the assumptions in Theorem 8.4, P_1^- is a valid p-value if the distribution of $(A_i, \boldsymbol{X}_i, Y_i(0), Y_i(1))$ satisfies Rosenbaum's sensitivity model.

We can compute the bounding p-value P_1^- by using the exact distribution of the right hand side of (8.4) or by using Monte Carlo. Alternatively, we can approximate it by the central limit theorem.

8.5 Exercise. For the sign statistic $\psi(t) \equiv 1$, derive an asymptotic p-value based on a central limit theorem for the bounding variable.

8.3 Sensitivity analysis in semiparametric inference

We will give another example of sensitivity analysis in the semiparametric inference framework described in Chapter 7.

The idea is quite simple. Suppose the treatment A is binary. By using the law of total expectation, we have, for any $a = 0, 1$,

$$\begin{aligned}
&\mathbb{E}[Y(a)] \\
=& \mathbb{E}\{\mathbb{E}[Y(a) \mid \boldsymbol{X}]\} \\
=& \mathbb{E}\{\mathbb{E}[Y(a) \mid A = a, \boldsymbol{X}]\, \mathbb{P}(A = a \mid \boldsymbol{X})\} + \mathbb{E}\{\mathbb{E}[Y(a) \mid A = 1 - a, \boldsymbol{X}]\, \mathbb{P}(A = 1 - a \mid \boldsymbol{X})\} \\
=& \mathbb{E}\{\mathbb{E}[Y \mid A = a, \boldsymbol{X}] \cdot \pi_a(\boldsymbol{X})\} + \mathbb{E}\{\mathbb{E}[Y(a) \mid A = 1 - a, \boldsymbol{X}] \cdot \pi_{1-a}(\boldsymbol{X})\}
\end{aligned}$$

The last equality used the consistency of counterfactuals.

The only non-identifiable term here is $\mathbb{E}[Y(a) \mid \boldsymbol{X}, A = 1 - a]$. The no unmeasured confounders assumption, $A \perp\!\!\!\perp Y(a) \mid \boldsymbol{X}$, renders this term identifiable:

$$\mathbb{E}[Y(a) \mid A = 1 - a, \boldsymbol{X}] = \mathbb{E}[Y(a) \mid A = a, \boldsymbol{X}] = \mathbb{E}[Y \mid A = a, \boldsymbol{X}].$$

This motivates us to specify the contrast between the identifiable and non-identifiable counterfactual quantities as a sensitivity parameter:[2]

$$\delta_a(\boldsymbol{x}) = \mathbb{E}[Y(a) \mid A = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y(a) \mid A = 0, \boldsymbol{X} = \boldsymbol{x}]. \tag{8.5}$$

8.6 Exercise. Show that the design bias for estimating the average treatment effect is given by

$$\begin{aligned}&\mathbb{E}\{\mathbb{E}[Y \mid A=1, \boldsymbol{X}]\} - \mathbb{E}\{\mathbb{E}[Y \mid A=0, \boldsymbol{X}]\} - \mathbb{E}[Y(1) - Y(0)] \\ =&\,\mathbb{E}\left[(1-\pi(\boldsymbol{X}))\delta_1(\boldsymbol{X}) + \pi(\boldsymbol{X})\delta_0(\boldsymbol{X})\right].\end{aligned}$$

So when $\delta_0(\boldsymbol{x}) = \delta_1(\boldsymbol{x}) = \delta$ for all $\boldsymbol{x}$, the design bias is simply δ.

Given the functions $\delta_0(\boldsymbol{x})$ and $\delta_1(\boldsymbol{x})$, estimating $\mathbb{E}[Y(1) - Y(0)]$ becomes another semiparametric inference problem. For example, we can estimate the design bias by plugging in the estimated propensity score:

$$\widehat{\text{bias}} = \frac{1}{n}\sum_{i=1}^{n}(1-\hat{\pi}(\boldsymbol{X}_i))\delta_1(\boldsymbol{X}_i) + \hat{\pi}(\boldsymbol{X}_i)\delta_0(\boldsymbol{X}_i).$$

We can then estimate the ATE by $\hat{\beta}_{\text{IPW}} - \widehat{\text{bias}}$.

8.7 Exercise. Suggest an outcome regression estimator and a doubly robust estimator in this setting.

Notes

[1] Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*(1), 13–26. doi:10.1093/biomet/74.1.13.

[2] Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, *121*, 151–179. doi:10.1023/a:1005285815569.

Chapter 9

Unmeasured confounders: Leveraging specificity

Recall the decomposition (6.1) of the error of a causal estimator:

Causal estimator − True causal effect = Design bias + Modelling bias + Statistical noise.

In the last three Chapters about observational studies, we assumed that the unmeasured confounders are non-existent or have a limited strength. In other words, this amounts to assuming that the design bias is zero or determined by the sensitivity model. Besides trying to measure all the confounders, is it possible to design the observational study cleverly to reduce the design bias?

9.1 Structural specificity

To overcome unmeasured confounders, the key idea is leveraging specificity of the causal structure.

Specificity is one of Bradford Hill's nine principles[1] for causality in epidemiological studies:

> One reason, needless to say, is the specificity of the association, the third characteristic which invariably we must consider. If, as here, the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favour of causation.

Hill is the coauthor of a landmark observational study on smoking and lung cancer[2]. Somewhat ironically, smoking has many detrimental health effects and is often used as a counterexample to the specificity principle. Nonetheless, specificity unifies several causal inference approaches that do not require no unmeasured confounders.

In graphical terminology, specificity refers to the lack of certain causal pathways. One classical example is the use of instrumental variables, which will be a central topic in this Chapter. In Exercise 5.35, it is shown that in a linear structural equation model corresponding to Figure 9.1, the causal effect of A on Y can be identified by the Wald ratio $\mathrm{Cov}(Z,Y)/\,\mathrm{Cov}(Z,A)$. This relies on two structural specificities in Figure 9.1: Z is independent of U, and Z has no direct effect on Y.

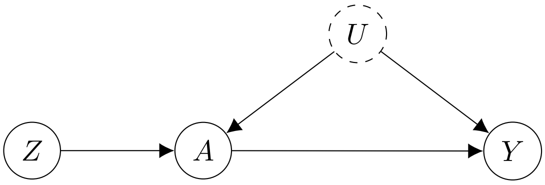


Figure 9.1: Instrumental variables.

By itself, structural specificity is not enough for causal identification. Additional assumptions are needed. One typical assumption is linearity, and the role of structural specificity can indeed be intuitively understood in the linear SEM framework (Chapter 3). Assuming the absence of certain edges reduces the dimension of the unknown parameters. If sufficiently many edges are non-existent, the path coefficients become identifiable; see Section 3.6 for some examples.

Besides linearity, other assumptions can be used with specificity to establish causal identification. One example is monotonicity of instrumental variables; see Section 9.4 below. Another example is the difference-in-difference estimator, a popular method to evaluate the effect of a policy. The corresponding causal diagram is Figure 9.2. The structural specificity here is the lack of causal pathway from A to W; due to this reason, W is often called a *negative control outcome.*

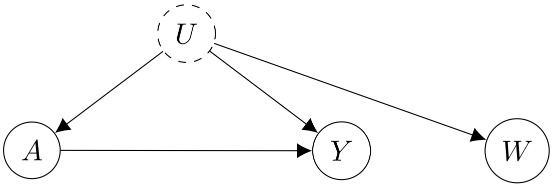


Figure 9.2: The use of negative control in the difference-in-difference estimator.

9.1 Exercise. Consider the causal diagram in Figure 9.2. Suppose the negative control outcome W has the same confounding bias as Y in the following sense:

$$\mathbb{E}[Y(0) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0] = \mathbb{E}[W(0) \mid A = 1] - \mathbb{E}[W(0) \mid A = 0].$$

Show that the so-called *parallel trend* assumption

$$\mathbb{E}[Y(0) - W \mid A = 1] = \mathbb{E}[Y(0) - W \mid A = 0]$$

is satisfied, and use it to show that the average treatment effect on the treated is identified by the so-called difference-in-differences estimator:

$$\mathbb{E}[Y(1) - Y(0) \mid A = 1] = \mathbb{E}[Y - W \mid A = 1] - \mathbb{E}[Y - W \mid A = 0].$$

9.2 Instrumental variables and two-stage least squares

Among all the study designs that leverage specificity, the method of instrumental variables (IV) has the longest history and is the most well established.

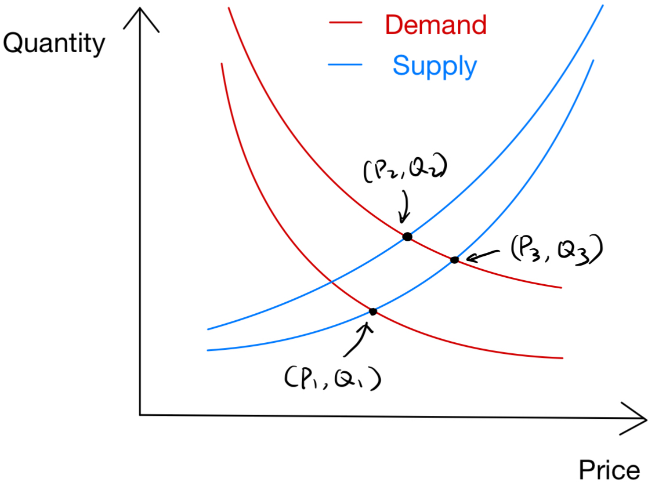


Figure 9.3: Estimating price elasticity using instrumental variables. The price elasticitiy of supply (demand) at (P_1, Q_1) is defined as as slope of the price-supply (price-demand) curve.

IV was invented by economist Philip Wright (father of Sewall Wright) in 1928 to estimate the price elasticities of (the causal effects of price on) demand and supply.[3] This is a challenging problem because price is determined simultaneously by demand and supply (Figure 9.3). To estimate the causal effect of price on supply, we cannot use observational data that correspond to different demand and supply curves (e.g., (P_1, Q_1) and (P_2, Q_2) in Figure 9.3). Instead, we need to use "exogenous" events that change the demand but not the supply (e.g., (P_1, Q_1) and (P_3, Q_3) in Figure 9.3). For example, we can use the COVID-19 outbreak as an instrumental variable for the demand of masks.

The simulatenous determination of price and quantity does not immediately fit in our causal inference framework,[4] but the same idea applies to observational studies with unmeasured confounders. In this case, the instrumental variable needs to be independent of the unmeasured confounders and changes the outcome only through changing the treatment (Figure 9.1). The "exogenous" variability in the IV can then be used to make unbiased causal inference.

In the rest of this section, we extend Exercise 5.35 to the setting with multiple IVs and observed confounders. Given iid observations of treatment A_i, outcome Y_i, instrumental variables $\boldsymbol{Z}_i$, observed confounders $\boldsymbol{X}_i$, and unobserved confounders $\boldsymbol{U}_i$, $i = 1, \dots, n$, we assume the structural equations for A and outcome Y are given by

$$A = \beta_{0A} + \boldsymbol{\beta}_{ZA}^T \boldsymbol{Z} + \boldsymbol{\beta}_{XA}^T \boldsymbol{X} + \boldsymbol{\beta}_{UA}^T \boldsymbol{U} + \epsilon_A, \tag{9.1}$$

$$Y = \beta_{0Y} + \beta_{AY} A + \boldsymbol{\beta}_{XY}^T \boldsymbol{X} + \boldsymbol{\beta}_{UY}^T \boldsymbol{U} + \epsilon_Y. \tag{9.2}$$

See Figure 9.4 for an example with one treatment and two IVs. The other linear structural equations can be derived from Definition 3.8 and are omitted. In fact, we do not need the other equations to be structural (causal) in the derivations below.

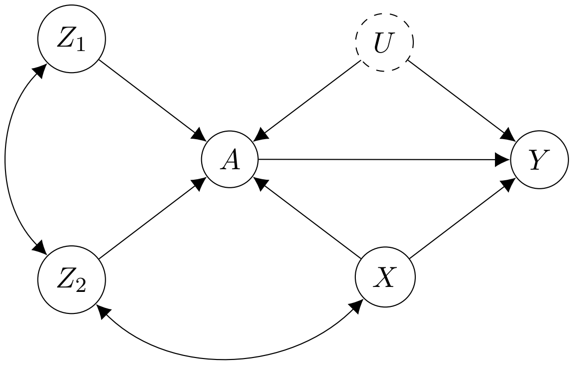


Figure 9.4: An example showing two IVs, Z_1 and Z_2, and one treatment A. There is one measured confounder X and one unmeasured confounder U. A bi-directional arrow indicates that the variables can be dependent.

By using $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{U}, \epsilon_A, \epsilon_Y$, we obtain from (9.1) and (9.2) that

$$\mathbb{E}[A \mid \boldsymbol{Z}, \boldsymbol{X}] = \tilde{\beta}_{0A} + \boldsymbol{\beta}_{ZA}^T \boldsymbol{Z} + \tilde{\boldsymbol{\beta}}_{XA}^T \boldsymbol{X}, \tag{9.3}$$

$$\mathbb{E}[Y \mid \boldsymbol{Z}, \boldsymbol{X}] = \tilde{\beta}_{0Y} + \beta_{AY}\, \mathbb{E}[A \mid \boldsymbol{Z}, \boldsymbol{X}] + \tilde{\boldsymbol{\beta}}_{XY}^T \boldsymbol{X}, \tag{9.4}$$

for some $\tilde{\beta}_{0A}, \tilde{\boldsymbol{\beta}}_{XA}, \tilde{\beta}_{0Y}, \tilde{\boldsymbol{\beta}}_{XY}$.

The key observation is that the confounded treatment effect β_{AY} is the coefficient of $\mathbb{E}[A \mid \boldsymbol{Z}, \boldsymbol{X}]$ in (9.4).

This motivates the *two-stage least squares* estimator of β_{AY}:

(i) Estimate $\mathbb{E}[A \mid \boldsymbol{Z}, \boldsymbol{X}]$ by a least squares regression of A on $\boldsymbol{Z}$ and $\boldsymbol{X}$. Let the fitted model be $\hat{\mathbb{E}}[A \mid \boldsymbol{Z}, \boldsymbol{X}]$.

(ii) Fit another regression of Y on $\hat{\mathbb{E}}[A \mid \boldsymbol{Z}, \boldsymbol{X}]$ and $\boldsymbol{X}$ by least squares, and let $\hat{\beta}_{AY}$ be the coefficient of $\hat{\mathbb{E}}[A \mid \boldsymbol{Z}, \boldsymbol{X}]$.

9.3 Instrumental variables: Method of moments

To study the asymptotic properties of the two-stage least squares estimator, we consider a more general counterfactual setup. We will skip the observed confounders $\boldsymbol{X}$ and consider the diagram in Figure 9.1 with a possibly multi-dimensional instrumental variables $\boldsymbol{Z}$.

9.2 Assumption (Core IV assumptions)**.** We make the following assumptions:

(i) *Relevance:* $\boldsymbol{Z} \not\perp\!\!\!\perp A$;

(ii) *Exogeneity:* $\boldsymbol{Z} \perp\!\!\!\perp \{A(\boldsymbol{z}), Y(\boldsymbol{z}, a)\}$ for all $\boldsymbol{z}, a$.

(iii) *Exclusion restriction:* $Y(\boldsymbol{z}, a) = Y(a)$ for all $\boldsymbol{z}, a$.

The core IV assumptions in Assumption 9.2 are structural and nonparametric. The three assumption would follow from (i) assuming the distribution is faithful to Figure 9.1; (ii) assuming the variables satisfy a single world causal model according to Figure 9.1; (iii) the recursive substitution of counterfactuals.

9.3 Remark. Different authors state these assumptions slightly differently, but they all reflect the structural assumptions: Z and A are dependent; there are no unmeasured Z-Y confounders; there is no direct effect from Z on Y.

As mentioned in Section 9.1, structural assumptions alone are generally not enough to overcome unmeasured confounders. For the rest of this Section, we further assume the causal effect of A on Y is a constant β:

$$Y(a) - Y(\tilde{a}) = (a - \tilde{a})\beta. \tag{9.5}$$

This would be satisfied if we assume the linear structural equation model (9.2).

9.4 Remark. When A is binary, this reduces to the constant treatment effect assumption (2.8) in randomisation inference. The main distinction is that randomisation inference is concerned with testing a given β, while the derivations below focus on the estimation of β. But we can also use randomisation inference for instrumental variables[5] and method of moments (Z-estimation) for models like (9.5)[6].

The estimation of β in (9.5) is a semiparametric inference problem. We first express it in terms of the observed data. Let $\tilde{a}$ be a reference level of the treatment; for simplicity, let $\tilde{a} = 0$. Like randomisation inference, (9.5) gives the imputation $Y(0) = Y - \beta A$. Let $\alpha = \mathbb{E}[Y(0)]$. The exogeneity assumption $\boldsymbol{Z} \perp\!\!\!\perp Y(\boldsymbol{z}, a)$ and exclusion restriction $Y(\boldsymbol{z}, a) = Y(a)$ imply that, for any function $g(\boldsymbol{z})$,

$$\mathbb{E}\left[(Y - \alpha - \beta A)\, g(\boldsymbol{Z})\right] = \mathbb{E}[(Y(0) - \alpha)\, g(\boldsymbol{Z})] = 0, \tag{9.6}$$

Let $\hat{\alpha} = \bar{Y} - \beta\bar{A}$, where $\bar{A} = \sum_{i=1}^{n} A_i/n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$. The method of moments estimator of β is given by solving the empirical version of (9.6) (and with α replaced by $\hat{\alpha}$). After some algebra, we obtain

$$\hat{\beta}_g = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})g(\boldsymbol{Z}_i)}{\frac{1}{n}\sum_{i=1}^{n}(A_i - \bar{A})g(\boldsymbol{Z}_i)}. \tag{9.7}$$

This as an empirical estimator of $\operatorname{Cov}(Y, g(\boldsymbol{Z}))/\operatorname{Cov}(A, g(\boldsymbol{Z}))$, which is the Wald ratio (5.11) with Z replaced by $g(\boldsymbol{Z})$ in. In view of this, $g(\boldsymbol{Z})$ is a one-dimensional summary statistic of all the instrumental variables.

9.5 Theorem. *Under Assumption 9.2, model* (9.5)*, iid sampling, and suitable regularity conditions including* $\mathrm{Cov}(A, g(\boldsymbol{Z})) > 0$*, we have*

$$\sqrt{n}(\hat{\beta}_g - \beta) \xrightarrow{d} \mathrm{N}(0, \sigma_g^2) \text{ as } n \to \infty,$$

where

$$\sigma_g^2 = \frac{\mathrm{Var}(Y - \beta A)\ \mathrm{Var}(g(\boldsymbol{Z}))}{\mathrm{Cov}^2(A, g(\boldsymbol{Z}))} \tag{9.8}$$

is minimised at $g^*(\boldsymbol{z}) = \mathbb{E}[A \mid \boldsymbol{Z} = \boldsymbol{z}]$ *by the Cauchy-Schwarz inequality.*

9.6 Exercise. Prove Theorem 9.5 by showing the influence function of $\hat{\beta}_g$ is given by

$$\psi_g(\boldsymbol{Z}, A, Y) = \frac{[\{Y - \mathbb{E}(Y)\} - \beta\{A - \mathbb{E}(A)\}][g(\boldsymbol{Z}) - \mathbb{E}\{g(\boldsymbol{Z})\}]}{\mathrm{Cov}(A, g(\boldsymbol{Z}))}.$$

The choice $g(\boldsymbol{Z}) = g^*(\boldsymbol{Z})$ that minimises the variance of $\hat{\beta}_g$ is often called the *optimal instrument.* To reduce the variance of $\hat{\beta}$, we can first estimate $g^*(\boldsymbol{Z})$ and then plug it in $\hat{\beta}$. Let the resulting estimator be $\hat{\beta}_{\hat{g}}$.

It is common to estimate $g^*(\boldsymbol{Z})$ by a linear regression, which returns the two-stage least squares estimator in Section 9.2. Other regression models including machine learning methods can also be used.

9.7 Exercise. Show that $\hat{\beta}_{\hat{g}}$ reduces to the two-stage least squares estimator (with no observed covariates $\boldsymbol{X}$), if we use a linear regression model for $g^*(\boldsymbol{Z})$ and obtain $\hat{g}(\boldsymbol{Z})$ by least squares.

Suppose $g(\boldsymbol{z})$ is the probability limit of $\hat{g}(\boldsymbol{z})$ as $n \to \infty$ (assuming it exists). If the first stage regression model is not too complex or the sample splitting technique is used (see Section 7.4), the difference between $\hat{\beta}_{\hat{g}}$ and $\hat{\beta}_g$ is negligible and $\hat{\beta}_{\hat{g}}$ has the same asymptotic distribution as $\hat{\beta}_g$.

9.8 Remark. There is a remarkable robustness property here: Theorem 9.5 shows that regardless of the choice of $g(\boldsymbol{z})$, $\hat{\beta}_g$ is $\sqrt{n}$-consistent as long as $\mathrm{Cov}(A, g(\boldsymbol{Z})) > 0$. A better model for $g^*(\boldsymbol{Z})$ provides a more efficient estimator of β. This bears some similarities to the regression adjustment methods in Section 2.5. However, there is no free lunch (as A is not randomised here): this robustness relies on the rather optimistic constant treatment effect assumption (9.5).

9.4 Complier average treatment effect

Can we relax the constant treatment effect assumption (9.5)? The answer is yes, but alternative assumptions need to be made. In this Section we will show how a monotonicity assumption allows us to identify the so-called complier average treatment effect.

Before we go into any detail, let us first introduce an example to motivate the definition of compliance classes.

9.9 Example. A common problem in randomised experiments is that not all experimental subjects would comply with the assigned treatment. This problem may be described by the IV diagram Figure 9.1, where

- Z is the initial treatment assignment (randomised);
- A is the actual treatment taken by the patient;
- Y is some clinical outcome

Because A is not randomised, the effect of A on Y can be confounded.

There are two solutions to this noncompliance problem:

(i) The *intention-to-treat* analysis that ignores A and estimates the causal effect of Z on Y (which is not confounded). This has been discussed in Chapter 2.

(ii) Use Z as an instrumental variable to estimate the causal effect of A on Y. This will be discussed next.

We will focus on the case of binary Z and A. As before, $A = 1$ refers to receiving the treatment and $A = 0$ refers to the control. The same terminology is used for the levels of Z.

9.10 Exercise. Verify that, if Z is binary, the Wald ratio can be written as

$$\frac{\mathrm{Cov}(Z,Y)}{\mathrm{Cov}(Z,A)} = \frac{\mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0]}{\mathbb{E}[A \mid Z=1] - \mathbb{E}[A \mid Z=0]}.$$

Motivated by the noncompliance problem, we may define four intuitive *compliance classes* based on the counterfactual treatments $A(0)$ and $A(1)$:

$$C = \begin{cases} \text{always taker (at)}, & \text{if } A(0) = A(1) = 1, \\ \text{never taker (nt)}, & \text{if } A(0) = A(1) = 0, \\ \text{complier (co)}, & \text{if } A(0) = 0, A(1) = 1, \\ \text{defier (de)}, & \text{if } A(0) = 1, A(1) = 0. \end{cases} \tag{9.9}$$

Notice that the definition of C involves cross-world counterfactuals. So the identification result below requires multiple-world counterfactual independence (see Section 5.2):

> **9.11 Assumption.** We make all the core IV assumptions in Assumption 9.2 and additionally assume cross-world counterfactual independence according to Figure 9.1. That is, (ii) in Assumption 9.2 is replaced by (ii') $\boldsymbol{Z} \perp\!\!\!\perp \{A(a), Y(\boldsymbol{z}, a) \mid \boldsymbol{z}, a\}$.

Assumption 9.2 is a structural assumption. As discussed in Section 9.1, additional assumptions are needed to make causal identification. The noncompliance problem motivates the following assumption:

9.12 Assumption (Monotonicity). $\mathbb{P}(A(1) \geq A(0)) = 1$, or equivalently, $\mathbb{P}(C = \mathrm{de}) = 0$.

For example, Assumption 9.12 is reasonable if the control patients have no access to the new treatment drug.

The next Theorem shows that under the above assumptions,

9.13 Theorem. *Under Assumptions 9.11 and 9.12, the complier average treatment effect is identified by*

$$\mathbb{E}[Y(1) - Y(0) \mid C = co] = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0]}.$$

Proof. Let us expand the term in the numerator using the compliance classes. Using the law of total expectation,

$$\mathbb{E}[Y \mid Z = 1] = \sum_{c \in \{\mathrm{at,nt,co,de}\}} \mathbb{E}[Y \mid Z = 1, C = c]\, \mathbb{P}(C = c \mid Z = 1).$$

By exclusion restriction (Assumption 9.2(iii)), $Y = Y(A(1))$ given $Z = 1$. Thus

$$\begin{aligned}\mathbb{E}[Y \mid Z = 1] =& \,\mathbb{E}[Y(1) \mid Z = 1, C = \mathrm{at}]\, \mathbb{P}(C = \mathrm{at} \mid Z = 1)\\ &+ \mathbb{E}[Y(0) \mid Z = 1, C = \mathrm{nt}]\, \mathbb{P}(C = \mathrm{nt} \mid Z = 1)\\ &+ \mathbb{E}[Y(1) \mid Z = 1, C = \mathrm{co}]\, \mathbb{P}(C = \mathrm{co} \mid Z = 1)\\ &+ \mathbb{E}[Y(0) \mid Z = 1, C = \mathrm{de}]\, \mathbb{P}(C = \mathrm{de} \mid Z = 1)\end{aligned}$$

By using the exogeneity of Z (Assumption 9.11(ii')) and the fact that C is a deterministic function of $A(0)$ and $A(1)$, we can drop the conditioning on $Z = 1$

$$\begin{aligned}\mathbb{E}[Y \mid Z = 1] =& \,\mathbb{E}[Y(1) \mid C = \mathrm{at}]\, \mathbb{P}(C = \mathrm{at}) + \mathbb{E}[Y(0) \mid C = \mathrm{nt}]\, \mathbb{P}(C = \mathrm{nt})\\ &+ \mathbb{E}[Y(1) \mid C = \mathrm{co}]\, \mathbb{P}(C = \mathrm{co}) + \mathbb{E}[Y(0) \mid C = \mathrm{de}]\, \mathbb{P}(C = \mathrm{de}).\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}[Y \mid Z = 0] =& \,\mathbb{E}[Y(1) \mid C = \mathrm{at}]\, \mathbb{P}(C = \mathrm{at}) + \mathbb{E}[Y(0) \mid C = \mathrm{nt}]\, \mathbb{P}(C = \mathrm{nt})\\ &+ \mathbb{E}[Y(0) \mid C = \mathrm{co}]\, \mathbb{P}(C = \mathrm{co}) + \mathbb{E}[Y(1) \mid C = \mathrm{de}]\, \mathbb{P}(C = \mathrm{de}).\end{aligned}$$

Therefore

$$\begin{aligned}&\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]\\ =& \,\mathbb{E}[Y(1) - Y(0) \mid C = \mathrm{co}]\, \mathbb{P}(C = \mathrm{co}) - \mathbb{E}[Y(1) - Y(0) \mid C = \mathrm{de}]\, \mathbb{P}(C = \mathrm{de}).\end{aligned}$$

Finally, by using the monotonicity assumption (Assumption 9.12), we obtain

$$\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] = \mathbb{E}[Y(1) - Y(0) \mid C = \mathrm{co}]\, \mathbb{P}(C = \mathrm{co}).$$

Using a similar argument, it can be shown that the denominator of β is

$$\mathbb{E}[A \mid Z = 1] - \mathbb{E}[A \mid Z = 0] = \mathbb{P}(C = \text{co}).$$

By dividing the two equations above, we obtain the identification formula of $\mathbb{E}[Y(1) - Y(0) \mid C = \text{co}]$. □

9.14 Remark. The complier average treatment effect $\mathbb{E}[Y(1) - Y(0) \mid C = \text{co}]$ is an instance of the *local average treatment effect.*[7] Here, local means the treatment effect is averaged over a specific subpopulation. What is unusual about the complier average treatment effect is that the subpopulation is defined in terms of cross-world counterfactuals (so it can never be observed). This demonstrates the utility of the counterfactual language as compliance class as a concept does not exist in a purely graphical setup. Pearl, 2009, page 29 discussed three layers of data queries: predictions, interventions, and counterfactuals. The meaning of "counterfactual" in Pearl's classification is not immediately clear. It is helpful to base the classification on whether the query only contains factuals (can be answered without causal inference), only contains counterfactuals in the same world (can be answered with randomised intervention), or contains cross-world counterfactuals (cannot be answered unless some non-verifiable cross-world independence is assumed).[8]

Notes

[1] Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. doi:10.1177/003591576505800503.

[2] Doll and Hill, 1950.

[3] Stock, J. H., & Trebbi, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, *17*(3), 177–194. doi:10.1257/089533003769204416.

[4] This is studied in simultaneous equations models (the variables are determined simultaneously). See, for example, Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics.* Oxford University Press, Chapter 18.

[5] Imbens, G. W., & Rosenbaum, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(1), 109–126. doi:10.1111/j.1467-985x.2004.00339.x.

[6] Equation (9.5) belongs to a more general class of models called structural nested models proposed by Robins; for a review, see Vansteelandt, S., & Joffe, M. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, *29*(4), 707–731. doi:10.1214/14-sts493

[7] Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467. doi:10.2307/2951620.

[8] See also Robins, J. M., & Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. In P. Shrout, K. Keyes, & K. Ornstein (Eds.), *Causality and psychopathology: Finding the determinants of disorders and their cures* (pp. 103–158). Oxford University Press.

Chapter 10

Mediation analysis

In the last Chapter, we saw how specificity (absence of causal mechanisms) could be useful to overcome unmeasured confounding.

In many other problems, the research question is about the causal mechanism itself. For example, instead of simply concluding that smoking causes lung cancer, it is more informative to determine which chemicals in the cigarettes are carcinogenic.

The problem of inferring causal mechanisms is called *mediation analysis*. It is a challenging problem and this Chapter will introduce you to some basic ideas.[1]

10.1 Linear SEM

We start with the simplest setting with three variables A (treatment), Y (outcome), and M (mediator) and no measured or unmeasured confounder (Figure 10.1).

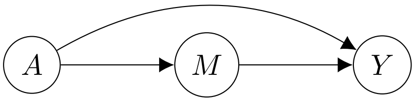


Figure 10.1: The basic mediation analysis problem with three variables and no confounder.

A linear SEM with respect to this causal diagram (Definition 3.8) assumes that the variables are generated by

$$M = \beta_{AM} A + \epsilon_M, \tag{10.1}$$
$$Y = \beta_{AY} A + \beta_{MY} M + \epsilon_Y, \tag{10.2}$$

where ϵ_M, ϵ_Y are mutually independent noise variables that are also independent of A.

Wright's path analysis formula (Theorem 3.14) shows that the total effect of A on Y is $\beta_{AY} + \beta_{AM}\beta_{MY}$. This can be seen directly from the reduced-form equation that plugs equation (10.1) into (10.2):

$$Y = (\underbrace{\beta_{AY}}_{\text{direct}} + \underbrace{\beta_{AM}\beta_{MY}}_{\text{indirect}})A + (\beta_{MY}\epsilon_M + \epsilon_Y).$$

The path coefficient β_{AY} is the *direct effect* of A on Y, and the product $\beta_{AM}\beta_{MY}$ is the *indirect effect* of A on Y through the mediator M. They can be estimated by first estimating the path coefficients using linear regression.

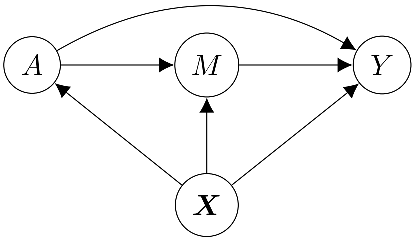


Figure 10.2: The basic mediation analysis problem with covariates.

This approach can be easily extended to allow for covariates (Figure 10.2). In this case, the linear SEM is

$$\begin{aligned} M &= \beta_{AM}A + \boldsymbol{\beta}_{XM}^T\boldsymbol{X} + \epsilon_M, \\ Y &= \beta_{AY}A + \beta_{MY}M + \boldsymbol{\beta}_{XY}^T\boldsymbol{X} + \epsilon_Y. \end{aligned}$$

The direct and indirect effects are still β_{AY} and $\beta_{XM}\beta_{MY}$, though to estimate them one now needs to include $\boldsymbol{X}$ in the regression models for M and Y.

This regression approach to mediation analysis is intuitive and very popular in practice.[2]

The obvious drawback is the strong linearity assumption, which enables us to express direct and indirect effects using regression coefficients. In more sophisticated scenarios (with nonlinearities and interactions), we can no longer rely on this decomposition.

10.2 Identification of controlled direct effect

The rest of this Chapter will develop a counterfactual approach to mediation analysis. For simplicity, we will again focus on the case of binary treatment. The counterfactuals $Y(a,m)$, $Y(a)$, and $Y(m)$ are defined as before via a nonparametric SEM (see Chapter 5).

The *controlled direct effect* (CDE) of A on Y when M is fixed at m is defined as

$$\text{CDE}(m) = \mathbb{E}[Y(1,m) - Y(0,m)]. \tag{10.3}$$

This quantity is of practical interest if we can intervene on both A and M.

CDE is a single-world quantity and can be identified using the g-formula if all the confounders are measured (Figure 10.2).

10.1 Theorem. *In a single-world causal model for the graph in Figure 10.2, we have*

$$\textit{CDE}(m) = \mathbb{E}\left[\mathbb{E}[Y \mid A = 1, M = m, \boldsymbol{X}]\right] - \mathbb{E}\left[\mathbb{E}[Y \mid A = 0, M = m, \boldsymbol{X}]\right]$$

Proof. By the law of total expectation and the single-world independence assumptions $Y(a,m) \perp\!\!\!\perp A \mid \boldsymbol{X}$ and $Y(m) \perp\!\!\!\perp M \mid A, \boldsymbol{X}$, we have, for any a and m,

$$\begin{aligned}
\mathbb{E}[Y(a,m)] &= \mathbb{E}\left\{\mathbb{E}[Y(a,m) \mid \boldsymbol{X}]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y(a,m) \mid A=a, \boldsymbol{X}]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y(m) \mid A=a, \boldsymbol{X}]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y(m) \mid A=a, M=m, \boldsymbol{X}]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}[Y \mid A=a, M=m, \boldsymbol{X}]\right\}.
\end{aligned}$$

The third and the last equalities used consistency of the counterfactual. □

Estimating CDE(m) is similar to estimating ATE in Chapter 7 (with the $M = m$ subsample).

However, the controlled direct effect does not motivate a definition of indirect effects.

10.3 Natural direct and indirect effects

Another approach to mediation analysis is to consider the *natural direct effect* (NDE) and *natural indirect effect* (NIE):[3]

$$\begin{aligned}
\text{NDE} &= \mathbb{E}\left[Y(1, M(0)) - Y(0, M(0))\right], \\
\text{NIE} &= \mathbb{E}\left[Y(1, M(1)) - Y(1, M(0))\right].
\end{aligned}$$

Compared to the CDE, these new quantities allow the mediator M to vary naturally according to some treatment level. By definition, they provide a decomposition of the average treatment effect:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \text{NDE} + \text{NIE}.$$

Both the NDE and NIE depend on a *cross-world* counterfactual, $Y(1, M(0))$. This means that some cross-world independence is needed for causal identification.

To focus on this issue, let's consider the basic mediation analysis problem with no covariates (Figure 10.1). The single-world independence assumptions of the graph in Figure 10.1 are

$$A \perp\!\!\!\perp M(a) \perp\!\!\!\perp Y(a,m),\ \forall\, a, m. \tag{10.4}$$

Consecutive $\perp\!\!\!\perp$ means mutual independence.

To identify $\mathbb{E}[Y(1, M(0))]$, we need an additional cross-world independence:

$$Y(1,m) \perp\!\!\!\perp M(0),\ \forall m. \tag{10.5}$$

10.2 Exercise. Suppose both A and M are binary. Count the number of pairwise independences in the single-world and multiple-world independence assumptions introduced in Definition 5.7.

10.3 Proposition. *Consider a causal model corresponding to the graph in Figure 10.1, in which the counterfactuals satisfy the multiple-world independence assumptions. When M is discrete, we have*

$$\mathbb{E}[Y(1, M(0))] = \sum_m \mathbb{E}[Y \mid A = 1, M = m] \cdot \mathbb{P}(M = m \mid A = 0).$$

In consequence,

$$NDE = \sum_m \big(\mathbb{E}[Y \mid A = 1, M = m] - \mathbb{E}[Y \mid A = 0, M = m]\big) \cdot \mathbb{P}(M = m \mid A = 0), \tag{10.6}$$

$$NIE = \sum_m \mathbb{E}[Y \mid A = 1, M = m] \cdot \big(\mathbb{P}(M = m \mid A = 1) - \mathbb{P}(M = m \mid A = 0)\big). \tag{10.7}$$

Proof. Using the (conditional) independence assumptions and consistency of counterfactuals,

$$\begin{aligned}\mathbb{E}[Y(1, M(0))] &= \sum_m \mathbb{E}[Y(1, m) \mid M(0) = m] \cdot \mathbb{P}(M(0) = m)\\ &= \sum_m \mathbb{E}[Y(1, m)] \cdot \mathbb{P}(M(0) = m \mid A = 0)\\ &= \sum_m \mathbb{E}[Y(1, m) \mid A = 1] \cdot \mathbb{P}(M = m \mid A = 0)\\ &= \sum_m \mathbb{E}[Y(m) \mid A = 1, M = m] \cdot \mathbb{P}(M = m \mid A = 0)\\ &= \sum_m \mathbb{E}[Y \mid A = 1, M = m] \cdot \mathbb{P}(M = m \mid A = 0)\end{aligned}$$

The identification formulas for NDE and NIE can be derived accordingly. □

10.4 Remark. In the proof of Proposition 10.3 only the second equality uses the cross-world independence. Thus without this assumption, we can still interpret the right hand side of (10.6) as the expectation of the controlled direct effect $Y(1, M') - Y(0, M')$, where M' is a randomised interventional analogue in the sense that M' is an independent random variable with the same distribution as $M(0)$.[4]

10.4 Observed confounders

To extend the identification results to more complex situations with covariates $\boldsymbol{X}$, let's examine the proof of Proposition 10.3. The first equality is just the law of total expectation. The second equality uses cross-world independence $Y(1, m) \perp\!\!\!\perp M(0)$ and the independence $M(0) \perp\!\!\!\perp A$. The third equality use $Y(1, m) \perp\!\!\!\perp A$ and consistency. The fourth equality uses consistency and $Y(m) \perp\!\!\!\perp M \mid A$. The last equality uses consistency.

To extend this proof, we can assume all the independences we used still hold when conditioning on $\boldsymbol{X}$:

10.5 Assumption (No unmeasured confounders)**.** We assume there are

(i) No unmeasured treatment-outcome confounders: $Y(a,m) \perp\!\!\!\perp A \mid \boldsymbol{X},\ \forall a,m$;

(ii) No unmeasured mediator-outcome confounders: $Y(m) \perp\!\!\!\perp M \mid A, \boldsymbol{X},\ \forall a,m$;

(iii) No unmeasured treatment-mediator confounders: $M(a) \perp\!\!\!\perp A \mid \boldsymbol{X},\ \forall a$.

10.6 Assumption. Assumption 10.5 is strengthened to include the cross-world independence $Y(a,m) \perp\!\!\!\perp M(a') \mid \boldsymbol{X},\ \forall a \neq a', m$.

The above assumptions are stated in terms of the counterfactuals. Assumption 10.5 can be checked by d-separation in the corresponding SWIGs. Assumption 10.6 cannot be checked in SWIGs because the counterfactuals are in different "worlds".

Importantly, Assumption 10.6 is not implied by Assumption 10.5 and multiple-world independence, as illustrated in the next example.

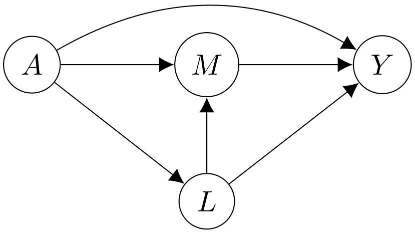


Figure 10.3: An illustration of treatment-induced mediator-outcome confounding.

10.7 Example. Consider the causal diagram in Figure 10.3, where the mediator M and the outcome Y are confounded by an observed variable L that is a descendant of A. In other words, L is another mediator that precedes M. The NPSEM corresponding to Figure 10.3 is

$$\begin{aligned} A &= f_A(\epsilon_A), \\ L &= f_L(A, \epsilon_L), \\ M &= f_M(A, L, \epsilon_M), \\ Y &= f_Y(A, L, M, \epsilon_Y). \end{aligned}$$

The counterfactuals are defined according to Definition 5.2 and the multiple-world independence assumptions (Definition 5.7) assert that the noise variables $\epsilon_A, \epsilon_L, \epsilon_M, \epsilon_Y$

are mutually independent. In this case, Assumption 10.5 is satisfied but Assumption 10.6 is generally not. To see this, the counterfactuals are, by definition,

$$M(a') = f_M(a', L(a'), \epsilon_M),$$
$$Y(a, m) = f_Y(a, L(a), m, \epsilon_Y).$$

So $Y(a,m) \perp\!\!\!\perp M(a') \mid L(a), L(a')$. However, the variable $L = f_L(A, \epsilon_L)$ does not contain as much information as $L(a) = f_L(a, \epsilon_L)$ and $L(a') = f_L(a', \epsilon_L)$ together, unless $f_L(a, \epsilon_L)$ does not depend on a (in which case $L \perp\!\!\!\perp A$ and the graph in Figure 10.3 is not faithful).

This example motivates the following structural assumption for $\boldsymbol{X}$:

10.8 Assumption (No treatment-induced mediator-outcome confounding)**.** $\boldsymbol{X} \cap de(A) = \emptyset$.

10.9 Lemma. *Under the multiple-world independence assumptions and Assumption 10.8, if $Y(m)$ and M are d-separated by $\{A, \boldsymbol{X}\}$ in $\mathcal{G}(m)$, then Assumption 10.6 is satisfied.*

Proof. Assumption 10.8 implies that $\boldsymbol{X}$ does not contain a descendant of M or Y. Given $\boldsymbol{X}$, the randomness of $M(a)$ comes from all (the noise variables of) the ancestors of $M(a)$ which have a directed path to M that is not blocked by $\{A, \boldsymbol{X}\}$; denote those ancestors as $an(M \mid A, \boldsymbol{X})$. Similarly, let $an(Y \mid A, M, \boldsymbol{X})$ be the ancestors of Y that are d-connected with Y given $\{A, M, \boldsymbol{X}\}$. So given $\boldsymbol{X}$, the randomness of $Y(a', m)$ comes from (the noise variables of) the variables in $an(Y \mid A, M, \boldsymbol{X})$.

We claim that $an(M \mid A, \boldsymbol{X})$ and $an(Y \mid A, M, \boldsymbol{X})$ are d-separated by $\boldsymbol{X}$. Otherwise, say $V \in an(M \mid A, \boldsymbol{X})$ and $U \in an(Y \mid A, M, \boldsymbol{X})$ are d-connected given $\boldsymbol{X}$; we can then append that d-connected path with the directed paths from U to M and from V to Y to create a d-connected path from M to Y, which contradicts the assumption that $Y(m) \perp\!\!\!\perp M \mid A, \boldsymbol{X}[\mathcal{G}(m)]$ (see Figure 10.4 for an illustration).

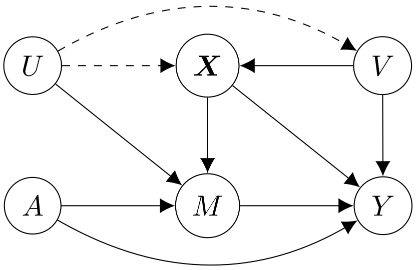


Figure 10.4: Illustration for the proof of Lemma 10.9. Adding the edge from U to $\boldsymbol{X}$ or from U to V creates a d-connected path from M to Y given $\boldsymbol{X}$.

Thus, the noise variables corresponding to $an(M \mid A, \boldsymbol{X})$ and $an(Y \mid A, M, \boldsymbol{X})$ are independent given $\boldsymbol{X}$. In consequence, $Y(a,m) \perp\!\!\!\perp M(a') \mid \boldsymbol{X}$. □

Our proof of Proposition 10.3 and Lemma 10.9 then imply the following identification result.

10.10 Theorem. *Suppose M and $\boldsymbol{X}$ are discrete. Under Assumptions 10.5 and 10.8,*

$$\begin{aligned}&\mathbb{E}[Y(1, M(0))]\\ =&\sum_{m,\boldsymbol{x}} \mathbb{E}[Y \mid A=1, M=m, \boldsymbol{X}=\boldsymbol{x}] \cdot \mathbb{P}(M=m \mid A=0, \boldsymbol{X}=\boldsymbol{x}) \cdot \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}).\end{aligned}$$

10.5 Extended graph interpretation

The cross-world independence in Assumption 10.6 is concerning. It is impossible to verify it empirically, because we can never observe $Y(a, m)$ and $M(a')$ together if $a \neq a'$.

To move beyond the impasse, one proposal is to consider an extension to the causal graph.[5] Consider the following example.

10.11 Example. Suppose a new process can completely remove the nicotine from tobacco, allowing the production of a nicotine-free cigarette to begin next year. The goal is to use the collected data on smoking status A, hypertensive status M and heart disease status Y from a randomised smoking cessation trial to estimate the incidence of heart disease in smokers were all smokers to change to nicotine-free cigarettes. Suppose a scientific theory tells us that the entire effect of nicotine on heart disease is through changing the hypertensive status, while the non-nicotine toxins in cigarettes have no effect on hypertension. Then, under the additional assumption that there are no confounders (besides A) for the effect of hypertension on heart disease, the causal DAG in Figure 10.1 can be used to describe the assumptions. The heart disease incidence rate among smokers of the new nicotine-free cigarettes is equal to $\mathbb{E}[Y(a=1, M(a=0))]$.

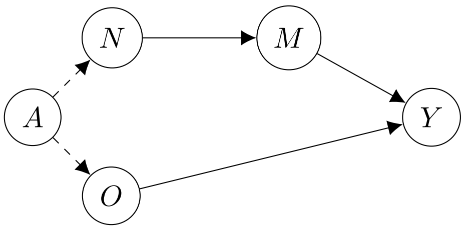


Figure 10.5: Extended causal diagram for mediation analysis.

The scientific story in this example allows as to extend the graph and include two additional variables N and O to represent the nicotine and non-nicotine content in cigarettes (Figure 10.5). Since the nicotine-free cigarette is not available till next year, we have $\mathbb{P}(A=N=O)=1$ in our current data.

Using this new graph, the heart disease incidence rate among the future smokers of nicotine-free cigarettes is given by

$$\mathbb{E}[Y(a=1, M(a=0))] = \mathbb{E}[Y(N=0,\, O=1)],$$

which no longer involves cross-counterfactuals.

Although the event $\{N = 0,\, O = 1\}$ has probability 0 in the current data, once the nicotine-free cigarettes become available, it will become possible to estimate $\mathbb{E}[Y(N = 0,\, O = 1)]$ by randomising this new treatment.

What is really interesting here is that we can still identify the distribution of $Y(N = 0,\, O = 1)$ with the current data, even though $\mathbb{P}(N = 0,\, O = 1) = 0$. Using the g-formula and $\mathbb{P}(A = N = O) = 1$,

$$
\begin{aligned}
&\mathbb{P}\big(Y(N = 0,\, O = 1) = y, M(N = 0) = m\big) \\
=&\,\mathbb{P}\big(Y(N = 0,\, O = 1) = y \mid M(N = 0) = m) \cdot \mathbb{P}(M(N = 0) = m\big) \\
=&\,\mathbb{P}\big(Y(O = 1) = y \mid M = m, N = 0) \cdot \mathbb{P}(M = m \mid N = 0\big) \\
=&\,\mathbb{P}\big(Y = y \mid M = m, N = 0, O = 1) \cdot \mathbb{P}(M = m \mid N = 0\big) \\
=&\,\mathbb{P}\big(Y = y \mid M = m, O = 1\big) \cdot \mathbb{P}\big(M = m \mid N = 0\big) \\
=&\,\mathbb{P}\big(Y = y \mid M = m, A = 1\big) \cdot \mathbb{P}\big(M = m \mid A = 0\big).
\end{aligned}
$$

Summing over m, we arrive at the formula in Proposition 10.3.

Notes

[1] A more comprehensive treatment is given in VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction.* Oxford University Press.

[2] To give you a sense of how popular mediation question is in psychology, the paper that made this regression analysis popular is now one of most cited of all times (close to 100,000 citations on Google Scholar): Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. doi:10.1037/0022-3514.51.6.1173

[3] This terminology is due to Pearl, 2000. These quantities are first proposed under the name “pure direct effect” and “total indirect effect” by Robins, J. M., & Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, *3*(2), 143–155. doi:10.1097/00001648-199203000-00013.

[4] Didelez, V., Dawid, A. P., & Geneletti, S. (2006). Direct and indirect effects of sequential treatments. In *Proceedings of the twenty-second conference on uncertainty in artificial intelligence* (pp. 138–146). UAI'06. Cambridge, MA, USA: AUAI Press.

[5] Robins and Richardson, 2010; see also Didelez, V. (2018). Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime Data Analysis*, *25*(4), 593–610. doi:10.1007/s10985-018-9449-0.