SISCER Module 13
Lecture 3: Sensitivity analysis & Intro to the DAG model

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 24, 2025

Recap

We discussed the following in day 1:

- Correlation does not imply causation.
- General theory for randomization inference for randomized experiments.
- Examples: Fisher's exact test; stepped-wedge design.
- Principles of observational study, no unmeasured confounding.
- Propensity score matching, optimal matching, covariate balance.

Plan for this lecture

- Motivating example: Smoking and lung cancer.
- Sensitivity analysis for matching.
- Intro to DAG model: linear SEMs.
- Intro to DAG model: general theory.

Outline

Smoking and lung cancer

Sensitivity analysis for matching

Intro to DAG model: Linear SEMs

Intro to DAG model: General theory

Smoking and lung cancer: A brief review of the history

- A seminal case-control study by Doll and Hill (1950) showed strong correlation between cigarette smoking and lung cancer.
- This was followed up by many prospective studies that match on many covariates, which all pointed to the same causal relationship (Doll and Hill, 1954; Hammond and Horn, 1954).
- 1957 statement by the UK Medical Research Council and 1964 report by the U.S. Surgeon General concluded that smoking is the principal cause of lung cancer.

Smoking and lung cancer: A brief review of the history

- But this was challenged by several statisticians and epidemiologists. For example, Berkson (1958) questioned the usage of risk ratio (instead of risk difference) in the studies and the lack of "specificity".

Table: Standardized death rates (per 1,000 men) in relation to smoking status, reproduced from Table V in Doll and Hill (1956) and Table 29 in Berkson (1958). The last two columns compare the death rates of heavy smokers (>25 g.) versus non-smokers in two different measures.

Cause of death	Smoking a daily average of				Heavy vs. Non- smokers	
	0 g.	1-14 g.	15-24 g.	>25 g.	Ratio	Difference
Lung cancer	0.07	0.47	0.86	1.66	23.71	1.59
Other cancer	2.04	2.01	1.56	2.63	1.29	0.59
Other respiratory diseases	0.81	1.00	1.11	1.41	1.74	0.60
Coronary thrombosis	4.22	4.64	4.60	5.99	1.42	1.77
Other causes	6.11	6.82	6.38	7.19	1.18	1.08

Smoking and lung cancer: A brief review of the history

- More relevant to us is the criticism by Fisher (1958) and response by Cornfield et al. (1959).
- Fisher was also a geneticist and questioned whether the association between smoking and lung cancer can be explained by confounding genotypes. He offered some preliminary twin data suggesting smoking is genetically heritable.
- This prompted the first sensitivity analysis that established a mathematical inequality which amounts to the following in this example.

  If cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers. If the relative prevalence of hormone X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect.

Outline

Smoking and lung cancer

Sensitivity analysis for matching

Intro to DAG model: Linear SEMs

Intro to DAG model: General theory

Matched observational studies

- By matching units in an observational studies with very similar covariates, the hope is that we reconstruct a block randomized experiment.
- Consider the Neyman-Rubin causal model. Suppose treated observation $i = 1, \ldots, n$ is matched to control observation $i + n$. Define

$$\mathcal{M} = \{\boldsymbol{a} \in \{0,1\}^{2n} \mid a_i + a_{i+n} = 1, i = 1, \ldots, n\}$$

- Randomization analysis of matched observational studies assumes

$$\mathbb{P}\Big(\boldsymbol{A} = \boldsymbol{a} \,\Big|\, \boldsymbol{X}, \boldsymbol{A} \in \mathcal{M}\Big) = \begin{cases} 2^{-n_1}, & \text{if } \boldsymbol{a} \in M, \\ 0, & \text{otherwise.} \end{cases}$$

- This would be satisfied if $(\boldsymbol{X}_i, A_i)$ are drawn i.i.d. (independent and identically distributed) from a population and the matching is exact.
- By further assuming no unmeasured confounders $A_i \perp\!\!\!\perp Y_i(a) \mid \boldsymbol{X}_i$ for all a, randomization tests can be constructed as in randomized experiments.

No unmeasured confounders

Let U be the unmeasured confounder (e.g. $U_i = (Y_i(0), Y_i(1))$). The key assumption above is that

$$\mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}, U_i, U_{i+n}\Big) = \frac{1}{2}.$$

- No unmeasured confounders allows us to discard U_i, U_{i+n}.
- Let $\pi(\boldsymbol{x}) = \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x})$ be the **propensity score**. Matching by $\boldsymbol{X}$ (exactly) then establishes the equality, because

$$\begin{aligned}&\mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}\Big)\\ =&\frac{\pi(\boldsymbol{X}_i)(1-\pi(\boldsymbol{X}_{i+n}))}{\pi(\boldsymbol{X}_i)(1-\pi(\boldsymbol{X}_{i+n})) + (1-\pi(\boldsymbol{X}_i))\pi(\boldsymbol{X}_{i+n})}.\end{aligned}$$

Rosenbaum's sensitivity model

- Inspired by Cornfield's sensitivity analysis, we would like to use a model that bounds the magnitude of unmeasured confounding.
- One option is the following model proposed by Rosenbaum (1987):

$$1/\Gamma \leq \mathrm{OR}(\mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u), \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u')) \leq \Gamma,$$

  where $OR(p, q) = \{p/(1-p)\}/\{q/(1-q)\}$ is the odds ratio and $\Gamma \geq 1$.
- This is equivalent to assume the following logistic model

$$\log \frac{\mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u)}{\mathbb{P}(A_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}, U_i = u)} = g(\boldsymbol{x}) + \gamma u, \;\; 0 \leq \gamma \leq \log \Gamma, 0 \leq U \leq 1.$$

Rosenbaum's sensitivity analysis (Rosenbaum, 2002)

Let $\pi_i = \mathbb{P}(A_i = 1 \mid \boldsymbol{X}_i, U_i), i = 1, \ldots, 2n$. A consequence of Rosenbaum's sensitivity model is that

$$\begin{aligned}\frac{1}{1+\Gamma} \le \mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}, U_i, U_{i+n}\Big)& \\ = \frac{\pi_i(1-\pi_{i+n})}{\pi_i(1-\pi_{i+n}) + (1-\pi_i)\pi_{i+n}} \le \frac{\Gamma}{1+\Gamma}&.\end{aligned}$$

- So within each pair, a fair coin toss is replace by a biased coin toss.
- We then seek the *least favorable* randomization distribution that is allowed by Rosenbaum's sensitivity model. This is usually given by the following (if we are trying to explain away a apparently positive treatment effect):

$$\mathbb{P}\Big(A_i = 1, A_{i+n} = 0 \,\Big|\, A_i + A_{i+n} = 1, \boldsymbol{X}_i, \boldsymbol{X}_{i+n}, U_i, U_{i+n}\Big) = \begin{cases} \frac{1}{1+\Gamma}, & \text{if } Y_i \ge Y_{i+n}, \\ \frac{\Gamma}{1+\Gamma}, & \text{if } Y_i < Y_{i+n}. \end{cases}$$

Sensitivity table and value

- A typical table of results of Rosenbaum's sensitivity analysis looks like the following.

Γ	1.0	2	4	8	**9**	10
Worst-case p-value	0.0001	0.0005	0.001	0.005	**0.01**	0.02

- The value of Γ where the worst-case p-value crosses the significance threshold (e.g. 0.01) is called the **sensitivity value** of the study. This is equal to 9 in Cornfield's example.
- The sensitivity value bears some similarity with the p-value. Both are random quantities determined by the data and indicate the strength of evidence.
- One may consider the problem of how to design an observational studies, not to minimize the p-value, but to maximize the sentivity value (Rosenbaum, 2010; Zhao, 2018).

What we learn from a sensitivity anlaysis

- A sensitivity analysis replaces qualitative claims about whether unmeasured biases are present with an **objective quantitative statement** about the magnitude of bias that would need to be present to change the conclusions.
- In this sense, a sensitivity analysis speaks to the assertion "it might be bias" in much the same way that a P-value speaks to the assertion "it might be bad luck".
- Because a genotype that had as large an effect on smoking and lung cancer might be considered unlikely in light of knowledge of the genotype's effect on other common diseases, the sensitivity analysis **strengthens the evidence** that smoking causes lung cancer although it does not prove that smoking causes lung cancer.

Outline

Smoking and lung cancer

Sensitivity analysis for matching

Intro to DAG model: Linear SEMs

Intro to DAG model: General theory

Introduction to linear structural equation models (SEMs)

- Linear SEMs were first developed by Sewall Wright for genetics problems.

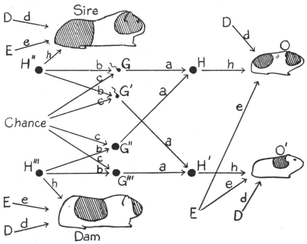


FIG. 5.

Diagram illustrating the casual relations between litter mates (O, O′) and between each of them and their parents. H, H′, H″, H,‴ represent the genetic constitutions of the four individuals, G, G′, G″, and G‴ that of four germ cells. E represents such environmental factors as are common to litter mates. D represents other factors, largely ontogenetic irregularity. The small letters stand for the various path coefficients.

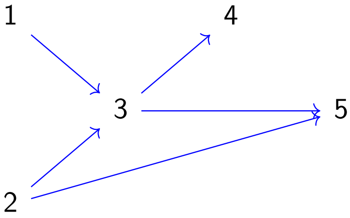
- It is a precursor of the general theory of (causal) graphical models developed since 1980/90s and remains widely used in practice.

From graphs to statistics

- A **directed acyclic graph (DAG)** is a directed graph (containing directed edges $\to$) with no directed cycles like $(j \to \cdots \to j)$.
- A linear SEM assumes the following: ($E_1, \ldots, E_d$ are independent noise terms)

$$V_j = \sum_{k \in \mathrm{pa}(j) = \{k : k \to j\}} \beta_{kj} V_k + E_j, \ j = 1, \ldots, d.$$

Example



$$
\begin{aligned}
V_1 &= E_1,\\
V_2 &= E_2,\\
V_3 &= \beta_{13} V_1 + \beta_{23} V_2 + E_3,\\
V_4 &= \beta_{34} V_3 + E_4,\\
V_5 &= \beta_{25} V_2 + \beta_{35} V_3 + E_5.
\end{aligned}
$$

Causal model

- Importantly, these equations are assumed to be **structural** in the sense that they still hold under interventions. In other words, potential outcomes can be defined from the structural equations in a **recursive** way.

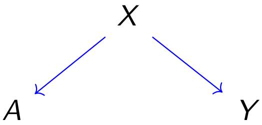
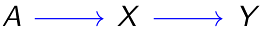
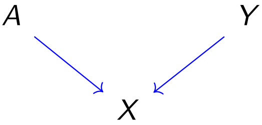
Example

$$
\begin{aligned}
V_1 &= E_1, \\
V_2 &= E_2, \\
V_3 &= \beta_{13} V_1 + \beta_{23} V_2 + E_3, \\
V_4 &= \beta_{34} V_3 + E_4, \\
V_5 &= \beta_{25} V_2 + \beta_{35} V_3 + E_5.
\end{aligned}
\quad \xRightarrow{\text{Intervene } V_2} \quad
\begin{aligned}
V_1 &= E_1, \\
V_2 &= v_2, \\
V_3(v_2) &= \beta_{13} V_1 + \beta_{23} v_2 + E_3, \\
V_4(v_2) &= \beta_{34} V_3(v_2) + E_4, \\
V_5(v_2) &= \beta_{25} v_2 + \beta_{35} V_3(v_2) + E_5.
\end{aligned}
$$

Thus, $V_4(v_1) = \beta_{25} v_2 + \beta_{35}(\beta_{13} V_1 + \beta_{23} v_2 + E_3) + E_5 = (\beta_{25} + \beta_{23}\beta_{35}) v_2 + \cdots$.
So the **total causal effect** is the product of coefficients along all **directed paths**.

Correlation vs. Causation

Consider linear SEMs (with normally distributed noise terms) for the following DAGs

		Causal effect of A on Y	$\mathrm{Cov}(A, Y)$	$\mathrm{Cov}(A, Y \mid X)$
Confounder		$= 0$	$\neq 0$	$= 0$
Mediator		$\neq 0$	$\neq 0$	$= 0$
Collider		$= 0$	$= 0$	$\neq 0$

Outline

Smoking and lung cancer

Sensitivity analysis for matching

Intro to DAG model: Linear SEMs

Intro to DAG model: General theory

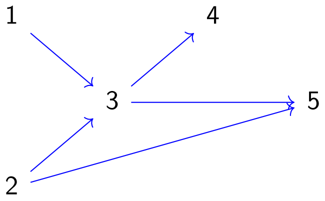
Factorization property

Definition

A probability density function p is said to **factorize** according to a DAG if

$$p(v_1, \ldots, v_d) = \prod_{j=1}^{d} p(v_j \mid v_{\mathrm{pa}(j)}).$$

Example

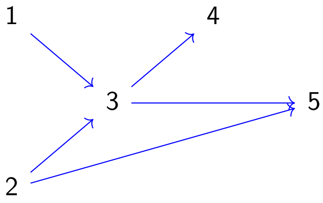


$$p(v_1, \ldots, v_5) = p(v_1)p(v_2)p(v_3 \mid v_1, v_2)p(v_4 \mid v_3)p(v_5 \mid v_2, v_3)$$

d-separation

- A path (no repeated vertices) is said to be **(ancestrally) blocked** by $L \subseteq V$ if
  1. it contains a non-collider V_l such that $V_l \in L$; **OR**
  2. it contains a collider V_m such that $V_m \notin L$ and there is no path like $V_m \to \cdots \to L$.
- Two disjoint sets $J, K \subseteq V$ are said to be **d-separated** by $L \subseteq V$ if all paths from a vertex in J to a vertex in K are blocked by L.
- Imagine the DAG represents how information flows from one variable to another.

Example



- $\{4\}$ and $\{5\}$ are **d-separated** by $\{3\}$.
- $\{4\}$ and $\{5\}$ are **d-separated** by $\{2, 3\}$.
- $\{1\}$ and $\{2\}$ are **d-separated** by $\emptyset$.
- $\{1\}$ and $\{2\}$ are **NOT d-separated** by $\{5\}$.
- $\{1\}$ and $\{5\}$ are **NOT d-separated** by $\{3\}$.

Global Markov model

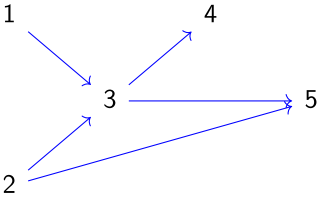
Definition

A probability distribution p is said to satisfy the **global Markov property** with respect to a DAG, if every d-separation in the graph implies a conditional independence.

Fundamental theorem for DAG models

A probability distribution factorizes according to a DAG **if and only if** it satisfies the global Markov property for the same DAG.

Example

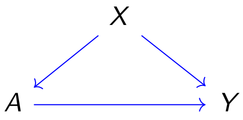


- $V_4 \perp\!\!\!\perp V_5 \mid V_3$.
- $V_4 \perp\!\!\!\perp V_5 \mid V_2, V_3$.
- $V_1 \perp\!\!\!\perp V_2$.
- $V_1 \not\perp\!\!\!\perp V_2 \mid V_5$ (in general).
- $V_1 \not\perp\!\!\!\perp V_5 \mid V_3$ (in general).

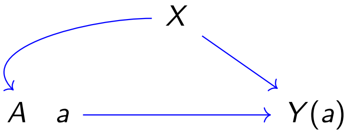
Single-world intervention graphs (SWIG)

- Factorization and global Markov properties say nothing about potential outcomes. They are not causal models.
- To give causal interpretions, we can imagine these graphs entail further graphs for (single-world) interventions creating by **node splitting**.

Example



No conditional independence among X, A, Y.



$A \perp\!\!\!\perp Y(a) \mid X.$

Further reading

Overall best

Hernan, M. A., & Robins, J. M. (2023). *Causal Inference: What If.* CRC Press.

Best introduction/popular science

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

Best mathematical

Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press.

Berkson, J. (1958). Smoking and lung cancer: Some observations on two recent reports. *Journal of the American Statistical Association*, 53(281):28–38.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and 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. **Note:** This classic paper was reprinted in the *International Journal of Epidemiology*, 38(5), 2009 with commentaries.

Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, 2(4682):739–748.

Doll, R. and Hill, A. B. (1954). The mortality of doctors in relation to their smoking habits. *BMJ*, 1(4877):1451–1455.

Doll, R. and Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking. *BMJ*, 2(5001):1071–1081.

Fisher, R. A. (1958). Cancer and smoking. *Nature*, 182(4635):596–596.

Hammond, E. C. and Horn, D. (1954). The relationship between human smoking habits and death rates. *Journal of the American Medical Association*, 155(15):1316–1328.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer Series in Statistics. Springer, New York.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer Series in Statistics. Springer, New York.

Zhao, Q. (2018). On sensitivity value of pair-matched observational studies. *Journal of the American Statistical Association*, 114(526):713–722.