Advanced Methods in Biostatistics

STAT 438

Winter 2022 (1221)[1]

Cameron Roopnarine[2]   Yeying Zhu[3]

29th March 2022

[1] Online Course until February 7$^{\text{th}}$, 2022

[2] LaTeXer

[3] Instructor

Contents

Chapter 1

Introduction

Week 1
5th to 7th January

About this Course

Three topics covered in this course:

- Causal Inference.
- Missing Data.
- Measurement Error.

Basics in Biostatistics

Review:

- Experimental Studies vs. Observational Studies.
- Statistics of Interest.
- Using Regression Models.
- Association vs. Causation.

Research Questions

Questions to ask when studying a disease:

- Which factors are associated with a given disease? These so-called risk factors are sometimes referred to as predictors, explanatory variables, covariates, independent variables, or exposure variables, etc.
- Which factors are associated with the duration of a given disease?
- Correlation (Association) does not imply causation.
- Ultimately, we want to ask: which factors cause the disease, or which factors determine the duration of the disease?

Types of Studies

- Experimental studies.
- Observational studies.

1.1 Experimental Studies

- In an experimental study, the investigator can manipulate the main (risk) factor of interest, while controlling for other factors.
- In a randomized experimental study, such as a clinical trial, eligible people are randomly assigned to one of two or more groups. One group receives the treatment (such as a new drug) while the control group receives nothing or an inactive placebo.
- Due to randomization, the investigator can control for both known and unknown factors, while investigating, typically, a treatment comparison.

Randomization and Causal Inference:

- Randomization is the perfect/golden design for causal inference.
- Random assignment of treatment (exposure) ensures balance across study arms with respect to observed and unobserved risk factors.
- Direct comparisons between treatment groups can be made.
- Any difference can be attributed to the causal effect of treatment.
- Randomization is not always feasible due to ethical/economic reasons.
- Even the treatment is randomized, the participant may not comply with the assigned treatment: compliance issue.

1.2 Observational Studies

- These studies are typically based on sampling populations with subsequent measurement of various factors of interest. In this setting, we cannot even take advantage of a naturally occurring experiment that changed risk factor status conveniently.
- It is sometimes useful to use these studies to look at the natural history of a disease, but any attempt to identify causality between a risk factor and outcome must be done with great caution.
- There is no experimental setting, as study participants typically self-reflect their exposure categories. Nevertheless, in large part due to ethics, such studies are most often to what we have access in Biostatistics.

  Examples of Observational Studies

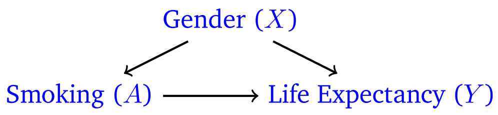
  1.
     - **Risk factor**: cigarette smoking.
     - **Outcome**: bladder cancer.
  2.
     - **Risk factor**: distance of home from hazardous waste site.
     - **Outcome**: respiratory disease.

- Three most popular observational studies:

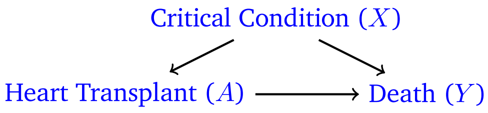
1. Cross-sectional studies.
2. Cohort studies.
3. Case-control studies.

- No control over which subjects have the exposure and which do not.
- Exposed and Unexposed groups may be quite different with respect to other subject characteristics.
- Differences in the outcome are not only due to the (risk) factor of interest, but also because of the masking effect of other covariates (confounders).

Confounding Issue



Another Example of Confounding



1.2.1 Cross-sectional Studies

- Individuals are selected from the target population and their status with respect to the risk factor and the disease status is ascertained at the same time.
- The data represents a snapshot view of the relation between the risk factor and the event occurrence.
- Surveys are often cross-section in nature where associations are of interest and less priority is given to establishing causation.
- Advantage: cross-sectional studies are typically short.
- Disadvantage: a serious problem with such cross-sectional studies is the inability to determine whether the disease outcome or the risk factor occurred first, again this makes causal inferences more problematic or almost impossible.

1.2.2 Cohort Studies

- Cohort studies typically include obtaining two groups from a pre-determined # of individuals, one possessing and the other not possessing a risk factor of interest. Subsequent counts of cases (and non-cases) of a disease of interest are then recorded.
- Much more often than not, cohort studies are prospective, but there are retrospective (or historical) cohort studies as well.

Table representing simple cohort study with sampling based on risk-factor status:

	Disease		
Risk Factor	Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

- $a \sim \text{BIN}\big(n_1, \mathbb{P}(D \mid E)\big)$.
- $c \sim \text{BIN}\big(n_2, \mathbb{P}(D \mid E^c)\big)$.

1.2.3 Case-control Studies

- In case-control studies, the direction of sampling differs from that of cohort studies. Specifically, the investigator selects a pre-determined # of disease cases and non-cases (i.e., controls), then looks retrospectively to see the # of individuals with and without the risk factor in each group.
- Case-control studies are retrospective studies.

Table representing simple case-control study with sampling based on disease status:

	Disease	
Risk Factor	Present	Absent
Present	a	b
Absent	c	d
Total	n_1	n_2

- $a \sim \text{BIN}\big(n_1, \mathbb{P}(E \mid D)\big)$.
- $b \sim \text{BIN}\big(n_2, \mathbb{P}(E \mid D^c)\big)$.

Week 2
10th to 14th January

Statistics of Interest

- Relative Risk.
- Excess Risk.
- Odds Ratio.
- Others: such as attributable risk, hazard ratio.

1.3 Relative Risk

The **relative risk** (RR) of an outcome (e.g., disease) D associated with a binary risk factor E is:

$$\text{RR} = \frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)},$$

where $0 \le \text{RR} < \infty$.

Remarks:

(1) The upper limit in practice typically will have a finite constraint. Noting that $\mathbb{P}(D \mid E) \le 1$, we have

$$\text{RR} \le \frac{1}{\mathbb{P}(D \mid E^c)} < \infty,$$

assuming $\mathbb{P}(D \mid E^c) \neq 0$.

(2) If there exists absolutely no association between D and E, this results in $\mathrm{RR} = 1$, that is, this will happen when $\mathbb{P}(D \mid E) = \mathbb{P}(D \mid E^c)$.

(3) If $\mathrm{RR} > 1$, there is greater risk or probability of D when E is present versus absent.

(4) If $\mathrm{RR} < 1$, there is lower risk or probability of D when E is present versus absent.

RR Calculation

- Recall the table for a cohort study.

	Disease		
Risk Factor	Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

Then,

$$\widehat{\mathrm{RR}} = \frac{a/(a+b)}{c/(c+d)} = \frac{a/n_1}{c/n_2}.$$

- To make inference, we have, approximately,

$$\log(\widehat{\mathrm{RR}}) \sim \mathcal{N}\Big(\log(\mathrm{RR}), \mathrm{Var}\big(\log(\mathrm{RR})\big)\Big),$$

where

$$\mathrm{Var}\big(\log(\mathrm{RR})\big) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.$$

- The (approximate) 95% confidence interval for $\log(\mathrm{RR})$ is

$$\log(\widehat{\mathrm{RR}}) \pm 1.96\sqrt{\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RR}})\big)}.$$

- The (approximate) 95% confidence interval for RR is:

$$\exp\Big\{\log(\widehat{\mathrm{RR}}) \pm 1.96\sqrt{\widehat{\mathrm{Var}}\big(\log(\widehat{\mathrm{RR}})\big)}\Big\}$$

For RR, we have

$$\mathrm{Var}\big(\log(\widehat{\mathrm{RR}})\big) \approx \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.$$

Proof: Define $\hat{p}_a = \frac{a}{n_1}$ and $\hat{p}_c = \frac{c}{n_2}$. Assuming the exposed and unexposed groups are independent, we have

$$\begin{aligned}\mathrm{Var}\big(\log(\widehat{\mathrm{RR}})\big) &= \mathrm{Var}\big(\log(\hat{p}_a) - \log(\hat{p}_c)\big)\\ &= \mathrm{Var}\big(\log(\hat{p}_a)\big) - \mathrm{Var}\big(\log(\hat{p}_c)\big).\end{aligned}$$

Using, Taylor's approximation, we have

$$\begin{aligned}\mathsf{log}(\hat{p}_a) &\approx \mathsf{log}(p_a) + \frac{\mathrm{d}\,\mathsf{log}(p_a)}{\mathrm{d}p_a}(\hat{p}_a - p_a)\\ &= \mathsf{log}(p_a) + \frac{(\hat{p}_a - p_a)}{p_a}.\end{aligned}$$

Since $a \sim \text{BIN}(n_1, p_a)$,

$$\begin{aligned}\mathsf{Var}\big(\mathsf{log}(\hat{p}_a)\big) &\approx \frac{\mathsf{Var}(\hat{p}_a)}{p_a^2}\\ &= \frac{\mathsf{Var}\Big(\frac{a}{n_1}\Big)}{p_a^2}\\ &= \frac{n_1 p_a(1-p_a)}{n_1^2 p_a^2}\\ &= \frac{1-p_a}{n_1 p_a}.\end{aligned}$$

Therefore,

$$\widehat{\mathsf{Var}}\big(\mathsf{log}(\hat{p}_a)\big) = \frac{1-\hat{p}_a}{n_1\hat{p}_a} = \frac{b}{a(a+b)}.$$

Similarly,

$$\widehat{\mathsf{Var}}\big(\mathsf{log}(\hat{p}_c)\big) = \frac{d}{c(c+d)}.$$

Therefore,

$$\begin{aligned}\widehat{\mathsf{Var}}\big(\mathsf{log}(\widehat{\text{RR}})\big) &= \frac{b}{a(a+b)} + \frac{d}{c(c+d)}\\ &= \frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}.\end{aligned}$$

Remarks:

(1) Relative risk (or sometimes called **risk ratio**) is a common measure of the disease-exposure association from cohort studies.

(2) In general, the relative risk is *not* symmetric in the role of D and E, that is,

$$\frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)} \neq \frac{\mathbb{P}(E \mid D)}{\mathbb{P}(E \mid D^c)}.$$

1.4 Excess Risk

While RR is a relative measure of risk, it is sometimes of interest to look at absolute measures of risk. One such measure is *excess risk*.

The **excess risk** (ER) is:

$$\text{ER} = \mathbb{P}(D \mid E) - \mathbb{P}(D \mid E^c),$$

where $-1 \leq \text{ER} \leq 1$.

Remark:

(1) $\text{ER} = 0$ means no excess risk (null value).

(2) ER > 0 means greater risk of D for E versus E^c.

(3) ER < 0 means lower risk of D for E versus E^c.

ER Calculation

- Recall the table for a cohort study.

	Disease		
Risk Factor	Present (D)	Absent (D^c)	Total
Present (E)	a	b	n_1
Absent (E^c)	c	d	n_2

Then,

$$\widehat{\text{ER}} = \frac{a}{a+b} - \frac{c}{c+d} = \hat{p}_a - \hat{p}_c.$$

- To make inference, we have, approximately,

$$\widehat{\text{ER}} \sim \mathcal{N}\big(\text{ER}, \mathsf{Var}(\widehat{\text{ER}})\big),$$

where

$$\mathsf{Var}(\widehat{\text{ER}}) \approx \frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}.$$

- The (approximate) 95% confidence interval for ER is:

$$\widehat{\text{ER}} \pm 1.96\sqrt{\widehat{\mathsf{Var}}(\widehat{\text{ER}})}.$$

For ER, we have

$$\mathsf{Var}(\widehat{\text{ER}}) \approx \frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}.$$

Proof: Define $\hat{p}_a = \frac{a}{n_1}$ and $\hat{p}_c = \frac{c}{n_2}$. Note that $a \sim \text{BIN}(n_1, p_a)$ and $c \sim \text{BIN}(n_2, p_c)$. Hence,

$$\begin{aligned}
\mathsf{Var}(\widehat{\text{ER}}) &= \mathsf{Var}(\hat{p}_a - \hat{p}_c) \\
&= \mathsf{Var}(\hat{p}_a) + \mathsf{Var}(\hat{p}_c) \\
&= \mathsf{Var}\left(\frac{a}{n_1}\right) + \mathsf{Var}\left(\frac{c}{n_2}\right) \\
&= \frac{p_a(1-p_a)}{n_1} + \frac{p_c(1-p_c)}{n_2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widehat{\mathsf{Var}}(\widehat{\text{ER}}) &= \frac{\hat{p}_a(1-\hat{p}_a)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2} \\
&= \frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}.
\end{aligned}$$

1.5 Odds Ratio

The **odds** of disease for the *exposed group* is

$$\frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D^c \mid E)} = \frac{\mathbb{P}(D \mid E)}{1 - \mathbb{P}(D \mid E)}.$$

The **odds** of disease for the *unexposed group* is

$$\frac{\mathbb{P}(D \mid E^c)}{\mathbb{P}(D^c \mid E^c)} = \frac{\mathbb{P}(D \mid E^c)}{1 - \mathbb{P}(D \mid E^c)}.$$

The **odds ratio** for measuring the association of disease with the exposed versus unexposed groups is

$$\text{OR} = \frac{\mathbb{P}(D \mid E)/\,\mathbb{P}(D^c \mid E)}{\mathbb{P}(D \mid E^c)/\,\mathbb{P}(D^c \mid E^c)} = \frac{\mathbb{P}(D \mid E)/[1 - \mathbb{P}(D \mid E)]}{\mathbb{P}(D \mid E^c)/[1 - \mathbb{P}(D \mid E^c)]}.$$

Remarks:

- OR $= 1$ means no association between D and E.
- OR > 1 means greater odds of disease when E is present.
- OR < 1 means lower odds of disease when E is present.

OR Calculation

- For general study with binary disease and exposure (risk factor):

	Disease	
Risk Factor	Present (D)	Absent (D^c)
Present (E)	a	b
Absent (E^c)	c	d

Here,

$$\widehat{\text{OR}} = \frac{\mathbb{P}(D \mid E)/\,\mathbb{P}(D^c \mid E)}{\mathbb{P}(D \mid E^c)/\,\mathbb{P}(D^c \mid E^c)} = \frac{\left(\frac{a}{a+b}\right)/\left(\frac{b}{a+b}\right)}{\left(\frac{c}{c+d}\right)/\left(\frac{d}{c+d}\right)} = \frac{ad}{bc}.$$

- To make inference, we have approximately,

$$\log(\widehat{\text{OR}}) \sim \mathcal{N}\big(\log(\text{OR}), \text{Var}\big(\log(\widehat{\text{OR}})\big)\big),$$

where

$$\text{Var}\big(\log(\widehat{\text{OR}})\big) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

Remark: OR is symmetric in roles of D and E:

$$\frac{\mathbb{P}(E \mid D)/\,\mathbb{P}(E^c \mid D)}{\mathbb{P}(E \mid D^c)/\,\mathbb{P}(E^c \mid D^c)} = \frac{\left(\frac{a}{a+c}\right)/\left(\frac{c}{a+c}\right)}{\left(\frac{b}{b+d}\right)/\left(\frac{d}{b+d}\right)} = \frac{ad}{bc}.$$

Therefore, the OR for D associated with E is equal to the OR for E associated with D. It is this symmetry that makes OR a popular “risk” measure for case-control studies, where sampling is done on disease status, not risk factor status.

1.6 Comments

The various types of probabilities that may be of interest:

- Joint probabilities: $\mathbb{P}(D, E)$, $\mathbb{P}(D, E^c)$, $\mathbb{P}(D^c, E)$, and $\mathbb{P}(D^c, E^c)$.
- Marginal probabilities: $\mathbb{P}(D)$, $\mathbb{P}(E)$, $\mathbb{P}(D^c)$, and $\mathbb{P}(E^c)$.
- Conditional probabilities: $\mathbb{P}(D \mid E)$, $\mathbb{P}(D \mid E^c)$, $\mathbb{P}(E \mid D)$, and $\mathbb{P}(E \mid D^c)$.

Cross-sectional Study:

- All the above probabilities can be estimated by the observed proportions if the sampling is simple random sampling.

Cohort Study:

- $\mathbb{P}(D \mid E)$, $\mathbb{P}(D^c \mid E)$, $\mathbb{P}(D \mid E^c)$, and $\mathbb{P}(D^c \mid E^c)$ can be estimated.
- Marginal probabilities $\mathbb{P}(D)$, $\mathbb{P}(E)$, and joint probabilities such as $\mathbb{P}(D, E)$ cannot be estimated.
- RR, ER, and OR can be estimated.

Case-control Study:

- Only $\mathbb{P}(E \mid D)$, $\mathbb{P}(E^c \mid D)$, $\mathbb{P}(E^c \mid D^c)$, and $\mathbb{P}(E \mid D^c)$ can be estimated.
- RR and ER cannot be estimated.
- OR can be estimated. Furthermore, RR $\approx$ OR when the disease is rare.

If the disease is rare in a case-control study (i.e., $\mathbb{P}(D) \approx 0$), we have RR $\approx$ OR.

Proof:

$$\begin{aligned} \text{OR} &= \frac{\mathbb{P}(D \mid E)/\,\mathbb{P}(D^c \mid E)}{\mathbb{P}(D \mid E^c)/\,\mathbb{P}(D^c \mid E^c)} \\ &= \frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)} \frac{\overbrace{\mathbb{P}(D^c \mid E^c)}^{\approx 1}}{\underbrace{\mathbb{P}(D^c \mid E)}_{\approx 1}} \\ &\approx \frac{\mathbb{P}(D \mid E)}{\mathbb{P}(D \mid E^c)} \\ &= \text{RR}. \end{aligned}$$

1.7 Regression Models

- Linear model.
- Log-linear model.
- Probit model.
- Logistic regression model.

Notation:

- X: exposure variable of interest.
- D: disease status.
- P_x: $\mathbb{P}(D = 1 \mid X = x)$, that is, how the risk of disease changes according to the exposure variable.

1.7.1 Linear Model

$$P_x = \mathbb{P}(D = 1 \mid X = x) = \alpha + \beta x.$$

- $\alpha = P_{x=0}$: the baseline risk.
- $\beta = P_{x+1} - P_x$: excess risk with 1 unit increase in exposure.

Drawbacks:

(1) Possible to produce $\hat{P}_x < 0$ or $\hat{P}_x > 1$.

(2) Can't be directly applied to case-control data.

1.7.2 Log-Linear Model

$$\log(P_x) = \log\big(\mathbb{P}(D = 1 \mid X = x)\big) = \alpha + \beta x.$$

- $\alpha = \log(P_{x=0})$: the log baseline risk.
- β: log relative risk associated with 1 unit increase in exposure.

Drawbacks:

(1) Possible to produce $\hat{P}_x > 1$.

(2) Can't be directly applied to case-control data.

1.7.3 Probit Model

$$P_x = \mathbb{P}(D = 1 \mid X = x) = \Phi(\alpha + \beta x),$$

where $\Phi(u)$ is the cdf of a standard normal distribution.

- $\alpha = \Phi^{-1}(P_{x=0})$.
- $\beta > 0$: the risk increases as X increases.
  $\beta < 0$: the risk increases as X decreases.

Drawbacks:

(1) There is no natural interpretation of α and α in terms of association.

(2) Can't be directly applied to case-control data.

1.7.4 Logistic Regression Model

$$P_x = \mathbb{P}(D = 1 \mid X = x) = \frac{1}{1 + \exp\big\{-(\alpha + \beta x)\big\}}.$$

- $\alpha = \log\left(\frac{P_{x=0}}{1 - P_{x=0}}\right)$: the log odds of disease at baseline.
- β: log odds ratio associated with 1 unit increase in exposure.

Advantages:

(1) $0 < \hat{P}_x < 1$.

(2) $\exp\{\beta\}$: the odds ratio, which is symmetric with respect to D and E if both are binary.

(3) Can be applied to case-control data.

Remarks:

(1) “Correlation does not imply causation.”

(2) Regression models tell us correlational/associational relationship between the exposure and the disease outcome

(3) *Conclusion*: We need better tools to define causality

(4) *Solution*: Potential outcomes framework (Chapter 2).

Chapter 2

Causal Inference and Potential Outcomes

Week 3
17th to 21st January

2.1 Causal Inference

2.1.1 Introduction

Reference

- Hernán M.A., & Robins J.M. (2020). Causal Inference: What If. Boca Raton: Chapman Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Causal Inference

Two notions of causation:

- Causes of an effect/outcome.
- Effects of a cause.

Causes of an effect

- What are causes of lung cancer?
- What was the cause of outbreak of food poisoning?

Effects of a cause/intervention

- Does smoking cause lung cancer?
- Does mixed feeding cause obesity?
- How strong is the effect?

- We concentrate on effects of a cause/treatment/intervention.
- Fundamentally simpler question: search is for useful information rather than complete scientific understanding.
- Typical approach for estimating causal effects (which may be problematic): collect sample on treatments/exposures, outcomes, and other variables in population; Use standard statistical methods (such as multiple regression) to derive inferences about associations between observable variables.

A Note

- In pharmaceutical companies, people used to believe conducting randomized clinical trials is the only way to evaluate a newly developed drug. However, there is a shifting trend going on right now because of:
  - Difficult to find control subjects.
  - Compliance issue.
  - Exclusion criteria.
  - Cost issue.
- New trend: utilizing existing Electronic Health Records data to help find controls.
- The study is not randomized any more: observational study.

Draw Causality

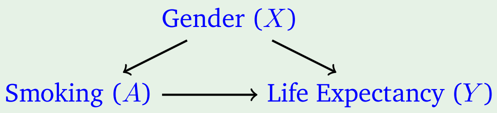
Observational Studies

- No control over which subjects have the exposure and which do not.
- Exposed and Unexposed groups may be quite different with respect to other subject characteristics
- It is sometimes useful to use these studies to look at the natural history of a disease, but any attempt to identify causality b/t a risk factor and outcome must be done w/ great caution.

2.1.2 Confounding

Confounding Issue in Observational Studies

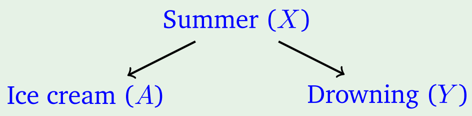
Differences in the outcome are not only due to the treatment, but also because of the masking effect of covariates (confounders).



Here, gender is known as a confounder. Very often, in real applications, the list of potential confounders could be very large, and even high-dimensional.

Another Example of Confounding

Researchers find when the consumption of ice cream increases, the death from drowning increases. Does eating ice cream lead to drowning?



Here, summer (hot weather) is a confounder.

Potential Outcomes Framework

- Useful to have more precise definitions of causal effects.
- Demystifies the process of going from association to causation.
- Allows explicit statements regarding what assumptions are necessary to justify causal inferences.
- Allows for more critical, better informed evaluation of causal claims.
- Helps determine when familiar methods useful or unfamiliar methods necessary.
- Motivates derivation and use of unfamiliar methods.

2.2 Potential Outcomes Framework

Definition of a Causal Effect

Suppose we have data on subjects $i = 1, \ldots, n$.

- $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{ip})^\top$: baseline covariates/potential confounders.
- A_i: treatment assignment/exposure status for subject i

$$A_i = \begin{cases} 1, & \text{if exposed/treated,} \\ 0, & \text{if unexposed/treated.} \end{cases}$$

- Y_i: observed outcome for subject i.

Counterfactuals/Potential outcomes

- Y_i^1: the potential outcome if subject i were treated/exposed.
- Y_i^0: the potential outcome if subject i were untreated/unexposed.

The **individual-level causal effect** for subject i is:

$$Y_i^1 - Y_i^0.$$

Causal Estimand

The **average causal effect** (ACE) is:

$$\text{ACE} = \mathbb{E}[Y_i^1 - Y_i^0] = \mathbb{E}[Y_i^1] - \mathbb{E}[Y_i^0],$$

where

- $\mathbb{E}[Y_i^1]$ is the mean potential outcome had all subjects in the population were treated/exposed, and
- $\mathbb{E}[Y_i^0]$ is the mean potential outcome had all subjects in the population were untreated/unexposed.

If Y is binary,

- ACE is causal excess risk (omit subscript i):

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1).$$

- **Causal relative risk**:
$$\frac{\mathbb{P}(Y^1 = 1)}{\mathbb{P}(Y^0 = 1)}.$$
- **Causal odds ratio**:
$$\frac{\mathbb{P}(Y^1 = 1)/\,\mathbb{P}(Y^1 = 0)}{\mathbb{P}(Y^0 = 1)/\,\mathbb{P}(Y^0 = 0)}.$$
- **Crude excess risk**:
$$\mathbb{P}(Y = 1 \mid A = 1) - \mathbb{P}(Y = 1 \mid A = 0).$$
- **Crude relative risk**:
$$\frac{\mathbb{P}(Y = 1 \mid A = 1)}{\mathbb{P}(Y = 1 \mid A = 0)}.$$
- **Crude odds ratio**:
$$\frac{\mathbb{P}(Y = 1 \mid A = 1)/\,\mathbb{P}(Y = 0 \mid A = 1)}{\mathbb{P}(Y = 1 \mid A = 0)/\,\mathbb{P}(Y = 0 \mid A = 0)}.$$

A Toy Example

Assume we have a population of 8 subjects:

	A	Y^0	Y^1
S_1	0	0	1
S_2	0	1	1
S_3	0	0	0
S_4	0	0	0
S_5	1	0	0
S_6	1	1	0
S_7	1	1	1
S_8	1	0	1

We get

$$\text{Causal excess risk (ACE)} = \mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1) = \frac{4}{8} - \frac{3}{8} = \frac{1}{8}.$$

For crude excess risk, we have

	A	Y^0	Y^1	Y
S_1	0	0	1	0
S_2	0	1	1	1
S_3	0	0	0	0
S_4	0	0	0	0
S_5	1	0	0	0
S_6	1	1	0	0
S_7	1	1	1	1
S_8	1	0	1	1

$$\text{Crude excess risk} = \mathbb{P}(Y = 1 \mid A = 1) - \mathbb{P}(Y = 1 \mid A = 0) = \frac{2}{4} - \frac{1}{4} = \frac{1}{4}.$$

Fundamental Problem of Causal Inference

For subject i, we only get to observe one of Y_i^1 and Y_i^0, that is,

$$Y_i = Y_i^1 A_i + Y_i^0(1 - A_i).$$

Remarks:

(1) In the literature, the above equality is often referred as the consistency assumption for causal inference

(2) For each subject i, one of the two potential outcomes is always missing.

(3) For this reason, many people believe causal inference is essentially a missing data problem.

2.3 Estimation

In **randomized studies**:

- $\mathbb{E}[Y \mid A = 1] = \mathbb{E}[Y^1 \mid A = 1] = \mathbb{E}[Y^1]$, and
- $\mathbb{E}[Y \mid A = 0] = \mathbb{E}[Y^0 \mid A = 0] = \mathbb{E}[Y^0]$.

Consequently, an unbiased estimate of ACE is:

$$\begin{aligned}
\widehat{\text{ACE}} &= \widehat{\mathbb{E}}[Y^1] - \widehat{\mathbb{E}}[Y^0] \\
&= \widehat{\mathbb{E}}[Y \mid A = 1] - \widehat{\mathbb{E}}[Y \mid A = 0] \\
&= \frac{\sum_{i=1}^n Y_i A_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n Y_i(1 - A_i)}{\sum_{i=1}^n (1 - A_i)},
\end{aligned}$$

where

- $\sum_{i=1}^n A_i = n_1$ is the number of treated/exposed subjects in the sample, and
- $\sum_{i=1}^n (1 - A_i) = n_0$ is the number of untreated/unexposed subjects in the sample.

In **observational studies**:

- $\mathbb{E}[Y \mid A = 1] = \mathbb{E}[Y^1 \mid A = 1] \neq \mathbb{E}[Y^1]$, and
- $\mathbb{E}[Y \mid A = 0] = \mathbb{E}[Y^0 \mid A = 0] \neq \mathbb{E}[Y^0]$,

where the inequalities are due to selection bias. Therefore, the estimator in randomized studies is biased for ACE in observational studies.

Assumptions for Causal Inference

2.4 Assumption 1

Assumption 1: Strongly Ignorable Treatment Assignment (SITA)

$$(Y^0, Y^1) \perp\!\!\!\perp (A \mid X).$$

Remarks:

- In observational studies, it means X includes all possible confounders (no unmeasured confounders).
- In randomized studies, we have $(Y^0, Y^1) \perp\!\!\!\perp A$.
- Within a subset of subjects with similar X, exposure/treatment can be viewed as if it were randomly assigned.
- This assumption cannot be verified on the observed data; more plausible as the size of X grows.
- If violated, instrumental variable approach can be used in some cases.

2.5 Assumptions 2–4

Assumption 2: Stable Unit Treatment Value Assumption (SUTVA)

$$(Y_i^0, Y_i^1) \perp\!\!\!\perp A_j \text{ for } i \neq j.$$

Remarks:

- Each subject's potential outcomes are not influenced by the actual treatment status of other subjects.
- Counter-example: infectious disease, family studies.
- If violated, divide the subjects into clusters.

Assumption 3: Common Support Condition (CSC)

$$0 < \mathbb{P}(A = 1 \mid X = x) < 1 \text{ for any } x.$$

Remarks:

- It means that Y^0 and Y^1 should both exist in principle.
- Can be violated if a particular group of subjects in the population always receive the treatment or never receive the treatment.
- If violated, re-define the population (exclude those subjects).

Assumption 4: Consistency

$$Y = Y^1 A + Y^0(1 - A).$$

Remarks:

- The observed outcome for a subject equals to the potential outcome under the actual treatment assignment the subject receives.
- Can be violated if different versions of treatment have different causal effects.

2.6 Propensity Scores

Motivation for Propensity Scores

The SITA assumption $(Y^0, Y^1) \perp\!\!\!\perp (A \mid X)$ gives us some ideas about how to estimate causal effects for observational studies.

- If we condition on X, we can estimate the causal effect as in a randomized study, which is relatively straightforward.
- However, if X contains a large number of covariates, conditioning on X is challenging (curse of dimensionality).
- Solution: propensity score methods

Propensity score is the conditional probability of being exposed/treated given baseline covariates:

$$\mathsf{ps}(x) = \mathbb{P}(A = 1 \mid X = x).$$

Also,

$$\mathsf{ps}(X) = \mathbb{P}(A = 1 \mid X).$$

Remarks:

- In simple randomized studies, $\mathsf{ps}(x) = 0.5$.
- In observational studies, $\mathsf{ps}(x)$ is unknown and must be estimated.

Properties

Properties of Propensity Score

- Propensity score is a balancing score:
$$X \perp\!\!\!\perp (A \mid \mathsf{ps}(X))$$
- If the treatment is strongly ignorable given X, that is,
$$(Y^0, Y^1) \perp\!\!\!\perp (A \mid X),$$
then it is strongly ignorable given $\mathsf{ps}(x)$
$$(Y^0, Y^1) \perp\!\!\!\perp (A \mid \mathsf{ps}(X)).$$
- $\mathsf{ps}(x)$ is a scalar, free of dimension of X.
- It is a summary of the contribution of all baseline characteristics to the exposure/treatment assignment.

2.7 Properties of Propensity Score

The propensity score is a **balancing score**, that is,

$$X \perp\!\!\!\perp (A \mid \mathsf{ps}(X))$$

Proof: Rosenbaum and Rubin (1983).

$$\begin{aligned}
\mathbb{P}\big(A = 1 \,\big|\, \mathsf{ps}(X), X\big) &= \mathbb{P}(A = 1 \mid X) && \mathsf{ps}(X) \text{ is a function of } X \\
&= \mathsf{ps}(X).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\mathbb{P}\big(A = 1 \,\big|\, \mathsf{ps}(X)\big) &= \mathbb{E}\big[A \,\big|\, \mathsf{ps}(X)\big] && \text{since } A \text{ is binary} \\
&= \mathbb{E}\big[\mathbb{E}[A \mid \underbrace{X}_{C_1}] \,\big|\, \underbrace{\mathsf{ps}(X)}_{C_2}\big] && \text{LIE since } C_2 = f(C_1) \\
&= \mathbb{E}\big[\mathsf{ps}(X) \,\big|\, \mathsf{ps}(X)\big] \\
&= \mathsf{ps}(X).
\end{aligned}$$

Therefore,

$$\mathbb{P}\big(A = 1 \,\big|\, \mathsf{ps}(X), X\big) = \mathbb{P}\big(A = 1 \,\big|\, \mathsf{ps}(X)\big).$$

In other words, $X \perp\!\!\!\perp (A \mid \mathsf{ps}(X))$.

If $(Y^0, Y^1) \perp\!\!\!\perp (A \mid X)$, then

$$(Y^0, Y^1) \perp\!\!\!\perp (A \mid \mathsf{ps}(X)).$$

Proof:

$$\begin{aligned}
\mathbb{P}(A = 1 \mid Y^0, Y^1, \mathsf{ps}(X)) &= \mathbb{E}[A \mid Y^0, Y^1, \mathsf{ps}(X)] && \text{since } A \text{ is binary} \\
&= \mathbb{E}\big[\mathbb{E}[A \mid \underbrace{Y^0, Y^1, X}_{C_1}] \,\big|\, \underbrace{Y^0, Y^1, \mathsf{ps}(X)}_{C_2}\big] && \text{LIE since } C_2 = f(C_1) \\
&= \mathbb{E}\big[\mathbb{E}[A \mid X] \,\big|\, Y^0, Y^1, \mathsf{ps}(X)\big] && \text{SITA} \\
&= \mathbb{E}\big[\mathsf{ps}(X) \,\big|\, Y^0, Y^1, \mathsf{ps}(X)\big] \\
&= \mathsf{ps}(X) \\
&= \mathbb{P}\big(A = 1 \,\big|\, \mathsf{ps}(X)\big). && \text{from the previous result}
\end{aligned}$$

Therefore,

$$(Y^0, Y^1) \perp\!\!\!\perp (A \mid \mathsf{ps}(X)).$$

2.8 Modeling of Propensity Scores

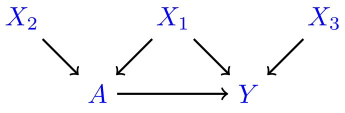
Estimation of Propensity Scores

- In observational studies, $\mathsf{ps}(x)$ is unknown and must be estimated.
- Logistic regression is the most common approach:

$$\mathsf{logit}\big(\mathsf{ps}(x)\big) = x^\top \beta.$$

  - For a subject with covariates x, $\widehat{\mathsf{ps}}(x) = \mathsf{expit}(x^\top \hat{\beta})$.
  - The goal of fitting a propensity score model is not interpretation (overfitting is okay).

Variable selection in PS model:



- X_1: real confounder.
- X_2: marginally related to the treatment.
- X_3: marginally related to the outcome.
- We should include X_1 and X_3 into the propensity score model.

Estimation of Propensity Scores

Remember when we estimate propensity scores, we model A (assuming binary) as a function of X. Therefore, we may employ non-parametric classification methods to estimate propensity scores:

- Classification and regression trees.
- Random forest.
- Generalized boosted model.
- Support vector machine.
- K nearest neighbours.

Chapter 3

Propensity Score-Based Methods

PS analysis is a two-step procedure:

1. Estimate propensity scores $\widehat{\mathsf{ps}}(X_i)$ for $i = 1, \ldots, n$ using data (A_i, X_i), $i = 1, \ldots, n$.
2. Using $\widehat{\mathsf{ps}}(X_i)$, $i = 1, \ldots, n$ to adjust the original sample and estimate causal effects:
   - Matching.
   - Stratification.
   - Inverse Probability Weighting (IPW).
   - Double-Robust Estimation.

3.1 Method 1: Matching

Basic idea: Consider matching strata, $S_1, \ldots, S_K$

$$\mathbb{E}[Y^1 - Y^0] = \sum_{k=1}^{K} \mathbb{E}[Y^1 - Y^0 \mid X \in S_k] \, \mathbb{P}(X \in S_k),$$

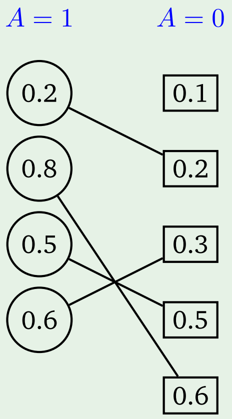
where for all $X \in S_k$, we have balance.

- $\mathbb{E}[Y^1 - Y^0 \mid X \in S_k]$ may be estimated as in a randomized study by using data in S_k.
- $\mathbb{P}(X \in S_k) \approx \frac{\text{number of subjects in } S_k}{n}$.
- Problem: if size of X is moderate or high-dimensional, it is hard to define the strata (curse of dimensionality).
- Solution: use the same stratified estimation strategy as before, but use $\mathsf{ps}(X)$ instead of X to stratify.

$$\mathbb{E}[Y^1 - Y^0] = \sum_{k=1}^{K} \mathbb{E}\big[Y^1 - Y^0 \bigm| \mathsf{ps}(X) \in S_k\big] \, \mathbb{P}\big(\mathsf{ps}(X) \in S_k\big).$$

- Lots of different matching algorithms available
- One example: 1 to 1 nearest available matching on estimated propensity scores:
  1. Randomly order the treated and untreated (control) subjects.
  2. Select the first treated subject and find the control subject with the closest propensity score.
  3. Both subjects are then removed from the pool and then repeat step 2 until all the treated subjects are matched.
  4. Fit an outcome model using the matched dataset.

A Toy Example



Once the matched dataset is formed, we have

$$\widehat{\text{ACE}} = \bar{Y}^{A=1,\text{matched}} - \bar{Y}^{A=0,\text{matched}}.$$

Different variations of matching algorithm:

- 1 to 1 versus 1 to M matching ($M = 3$ is a common choice).
- **With replacement** versus **without replacement** matching.
- **Matching without calipers** versus **matching within calipers** (only two closest subjects whose propensity score difference is within a prespecified caliper, say, 0.2, will be matched).
- Other variations.
- Advantage: matching based on propensity scores is far simpler than matching on even a modest number of risk factors simultaneously.
- Disadvantage: obtaining valid standard error of the causal estimator is challenging. In R, use `Matching` package.

3.2 Method 2: Stratification

Basic idea: create only a few strata; in each stratum, individuals have similar, but not identical, values of $\widehat{\text{ps}}(X)$.

Algorithm for Stratification

1. Divide the subjects into K (usually $K = 5$) strata on the basis of the quantiles of $\widehat{\text{ps}}(X_i)$, $i = 1, \ldots, n$.
2. The causal effect is estimated within each stratum as in a randomized study: For the j^{th} stratum, defined as S_j:

$$\widehat{\text{ACE}}^{(j)} = \frac{\sum_{i \in S_j} Y_i A_i}{\sum_{i \in S_j} A_i} - \frac{\sum_{i \in S_j} Y_i (1 - A_i)}{\sum_{i \in S_j} (1 - A_i)},$$

and

$$\widehat{\text{ACE}} = \frac{1}{K} \sum_{j=1}^{K} \widehat{\text{ACE}}^{(j)}.$$

- Advantage: Simpler than matching algorithms.

- Disadvantage:
    - It is possible to have no treated/untreated subjects in a particular stratum.
    - There is no good way to obtain valid standard error of $\widehat{\text{ACE}}$.

3.3 Method 3: Inverse Probability Weighting

$$\mathbb{E}\left[\frac{AY}{\mathsf{ps}(X)}\right] = \mathbb{E}[Y^1] \quad \text{and} \quad \mathbb{E}\left[\frac{(1-A)Y}{1-\mathsf{ps}(X)}\right] = \mathbb{E}[Y^0].$$

Proof: Lunceford & Davidian (2004).

$$\begin{aligned}
\mathbb{E}\left[\frac{AY}{\mathsf{ps}(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{AY}{\mathsf{ps}(X)} \,\middle|\, Y^1, X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{AY^1}{\mathsf{ps}(X)} \,\middle|\, Y^1, X\right]\right] && \text{since } AY^1 + (1-A)Y^0 = Y \\
&= \mathbb{E}\left[\frac{Y^1}{\mathsf{ps}(X)}\,\mathbb{E}[A \mid Y^1, X]\right] \\
&= \mathbb{E}\left[\frac{Y^1}{\mathsf{ps}(X)}\,\mathbb{E}[A \mid X]\right] && \text{SITA} \\
&= \mathbb{E}\left[\frac{Y^1}{\mathsf{ps}(X)}\,\mathsf{ps}(X)\right] \\
&= \mathbb{E}[Y^1].
\end{aligned}$$

Similarly,

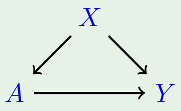
$$\begin{aligned}
\mathbb{E}\left[\frac{(1-A)Y}{1-\mathsf{ps}(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{(1-A)Y}{1-\mathsf{ps}(X)} \,\middle|\, Y^0, X\right]\right] \\
&\;\;\vdots \\
&= \mathbb{E}[Y^0].
\end{aligned}$$

We used SITA: $(Y^1, Y^0) \perp\!\!\!\perp (A \mid X)$, however we only need WITA: $Y^1 \perp\!\!\!\perp (A \mid X)$ and $Y^0 \perp\!\!\!\perp (A \mid X)$.

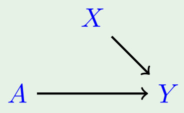
- Weighting scheme:
    - For those in the treatment group ($A_i = 1$), assign a weight of $w_i = 1/\widehat{\mathsf{ps}}(X_i)$.
    - For those in the control group ($A_i = 0$), assign a weight of $w_i = 1/(1 - \widehat{\mathsf{ps}}(X_i))$.
- By weighting, each subject is replicated w_i times. IPW creates a pseudo-population in which A and X are not associated (no confounding).

Example: IPW Weighting

- Before weighting:



- After weighting:



X and A are no longer confounded.

- To estimate $\mathbb{E}[Y^1]$, we take the weighted average of the observed Y in the treatment group, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\widehat{\mathsf{ps}}(X_i)}.$$

- To estimate $\mathbb{E}[Y^0]$, we take the weighted average of the observed Y in the control group, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\widehat{\mathsf{ps}}(X_i)}.$$

- The consistent (asymptotically unbiased) estimator for ACE is:

$$\hat{\tau}_{\text{IPW}_1}=\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i}{\widehat{\mathsf{ps}}(X_i)}-\frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\widehat{\mathsf{ps}}(X_i)}.$$

- A more efficient estimator for ACE is:

$$\hat{\tau}_{\text{IPW}_2}=\left(\sum_{i=1}^{n}\frac{A_i}{\widehat{\mathsf{ps}}(X_i)}\right)^{-1}\sum_{i=1}^{n}\frac{A_iY_i}{\widehat{\mathsf{ps}}(X_i)}-\left(\sum_{i=1}^{n}\frac{1-A_i}{1-\widehat{\mathsf{ps}}(X_i)}\right)^{-1}\sum_{i=1}^{n}\frac{(1-A_i)Y_i}{1-\widehat{\mathsf{ps}}(X_i)}.$$

- Properties (such as standard error) of IPW estimators should reflect the fact that $\mathsf{ps}(X_i)$'s are estimated.
  - In R, use `survey` or `geepack` packages to get standard errors of $\hat{\tau}$.
  - Bootstrap approach: A random re-sampling approach to measure the accuracy of a sample estimator.

Bootstrap Approach for IPW

- Sample $b=1,\ldots,B$ (say $B=500$) datasets of size n **with replacement** from the original data.
- For each bootstrapped sample, estimate $\hat{\tau}^{(b)}_{\text{IPW}_1}$ and $\hat{\tau}^{(b)}_{\text{IPW}_2}$, $b=1,\ldots,B$.
- Obtain

$$\widehat{\mathsf{Var}}(\hat{\tau}_{\text{IPW}_1})=\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}^{(b)}_{\text{IPW}_1}-\bar{\hat{\tau}}_{\text{IPW}_1})^2,$$

  where $\bar{\hat{\tau}}_{\text{IPW}_1}=\frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^{(b)}_{\text{IPW}_1}$.

- The same procedure applies to $\widehat{\mathsf{Var}}(\hat{\tau}_{\text{IPW}_2})$.

3.4 Method 4: Double-Robust Estimation

- Problem: If the propensity score model is incorrect, Matching, Stratification and IPW estimators will be biased.
- Solution: Combine IPW with the regression modelling approach to protect against model misspecification.

- “Double-robust” means $\hat{\tau}_{\text{DR}}$ is consistent for ACE, if one of the following is true:
  - The model for the propensity score, $\mathsf{ps}(X,\beta)$, is correctly specified:
  $$\mathsf{logit}\big(\mathsf{ps}(X,\beta)\big) = X^\top\beta.$$
  - The models for the outcome regression, $m_a(X;\gamma_a)$, $a=0,1$, are correctly specified:
  $$m_1(X;\gamma_1) = \mathbb{E}[Y \mid X, A=1] = X^\top\gamma_1.$$
  $$m_0(X;\gamma_0) = \mathbb{E}[Y \mid X, A=0] = X^\top\gamma_0.$$
- The consistency of the estimator does not require both sets of models to be correct.
- The DR estimator:
$$\begin{aligned}\hat{\tau}_{\text{DR}} = &\frac{1}{n}\sum_{i=1}^{n}\frac{A_iY_i - [A_i - \widehat{\mathsf{ps}}(X_i)]\hat{m}_1(X_i)}{\widehat{\mathsf{ps}}(X_i)}\\ &-\frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)Y_i + [A_i - \widehat{\mathsf{ps}}(X_i)]\hat{m}_0(X_i)}{1-\widehat{\mathsf{ps}}(X_i)}\\ &= \hat{\tau}_{1,\text{DR}} - \hat{\tau}_{0,\text{DR}},\end{aligned}$$
where $\hat{\tau}_{1,\text{DR}}$ estimates $\mathbb{E}[Y^1]$ and $\hat{\tau}_{0,\text{DR}}$ estimates $\mathbb{E}[Y^0]$.

Remark: The above estimator is also called augmented estimator since it can be viewed as taking the inverse weighted estimator and “augmenting” it by a second term.

Implementation of Double-Robust Estimation

1. Fit logistic regression model for the propensity score
$$\mathsf{logit}\big(\mathbb{P}(A_i = 1 \mid X_i = x_i)\big) = x_i^\top\beta, \qquad \widehat{\mathsf{ps}}_i = \mathsf{expit}(x_i^\top\hat{\beta}).$$
2. Fit regression model for the outcome using data from the treatment group only, predict for all subjects
$$\mathbb{E}[Y_i \mid X_i = x_i, A_i = 1] = x_i^\top\gamma_1, \qquad \hat{m}_1(x_i) = x_i^\top\hat{\gamma}_1.$$
3. Fit a regression model for the outcome (same form as above) using data from the control group only, predict for all subjects
$$\mathbb{E}[Y_i \mid X_i = x_i, A_i = 0] = x_i^\top\gamma_0, \qquad \hat{m}_0(x_i) = x_i^\top\hat{\gamma}_0.$$
4. Plug in predicted values into the expression for $\hat{\tau}_{\text{DR}}$.

- Properties (such as standard error) of the DR estimator should reflect the fact that $\mathsf{ps}(X_i)$’s are estimated.

- A formula for the theoretical standard error

  $$\widehat{\mathsf{Var}}(\hat{\tau}_{\text{DR}}) = \frac{1}{n^2}\sum_{i=1}^{n} \hat{I}_i^2,$$

  where

  $$\begin{aligned}\hat{I}_i = & \frac{A_iY_i - [A_i - \widehat{\mathsf{ps}}(X_i)]\hat{m}_1(X_i)}{\widehat{\mathsf{ps}}(X_i)} \\ & - \frac{(1-A_i)Y_i + [A_i - \widehat{\mathsf{ps}}(X_i)]\hat{m}_0(X_i)}{1-\widehat{\mathsf{ps}}(X_i)} \\ & - \hat{\tau}_{\text{DR}}.\end{aligned}$$

  Note: The formula only works well if both the propensity score and outcome regression models are correctly specified.

Bootstrap Approach for Double-Robust Estimation

- Sample $b = 1, \ldots, B$ (say $B = 500$) datasets of size n **with replacement** from the original data.
- For each bootstrapped sample, estimate $\hat{\tau}_{\text{DR}}^{(b)}$, $b = 1, \ldots, B$.
- Obtain

  $$\widehat{\mathsf{Var}}(\hat{\tau}_{\text{DR}}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\tau}_{\text{DR}}^{(b)} - \bar{\hat{\tau}}_{\text{DR}})^2,$$

  where $\bar{\hat{\tau}}_{\text{DR}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_{\text{DR}}^{(b)}$.

3.5 Case Study I: Propensity Score Analysis

Propensity Score Methods:

- Matching.
- Stratification.
- Inverse Probability Weighting.
- Double-Robust Estimation.

Chapter 4

Marginal Structural Models

Week 5
0131 to 4th February

Reference: Robins, James M., Miguel Angel Hernan, and Babette Brumback. “Marginal structural models and causal inference in epidemiology.” (2000): 550-560.

4.1 MSM

Definition

- In regression models, we assume
  $$\mathbb{E}[Y \mid A = a] = \beta_0 + \beta_1 a.$$
  LHS represents “average” response in members of the population who actually received treatment a.
- In MSMs, we assume
  $$\mathbb{E}[Y^a] = \beta_0^* + \beta_1^* a.$$
  LHS represents “average” response under treatment a in the entire population; MSMs model the potential outcomes directly.
- If A is unconfounded, $\beta_1^* = \beta_1$ because
  $$\mathbb{E}[Y \mid A = a] = \mathbb{E}[Y^a \mid A = a] = \mathbb{E}[Y^a].$$
- If A is confounded, $\beta_1^* \neq \beta_1$.
  - β_1^* cannot be directly estimated.
  - Idea: use inverse probability weighting (IPW) to create a pseudo-population in which A is unconfounded.

Parameters of Interest

If both A and Y are binary, recall

- **Crude excess risk**:
  $$\mathbb{P}(Y = 1 \mid A = 1) - \mathbb{P}(Y = 1 \mid A = 0).$$
- **Crude relative risk**:
  $$\frac{\mathbb{P}(Y = 1 \mid A = 1)}{\mathbb{P}(Y = 1 \mid A = 0)}.$$

- **Crude odds ratio**:
$$\frac{\mathbb{P}(Y=1 \mid A=1)/\,\mathbb{P}(Y=0 \mid A=1)}{\mathbb{P}(Y=1 \mid A=0)/\,\mathbb{P}(Y=0 \mid A=0)}.$$

For comparison,

- **Crude excess risk**:
$$\mathbb{P}(Y=1 \mid A=1) - \mathbb{P}(Y=1 \mid A=0).$$
- **Causal relative risk**:
$$\frac{\mathbb{P}(Y^1=1)}{\mathbb{P}(Y^0=1)}.$$
- **Causal odds ratio**:
$$\frac{\mathbb{P}(Y^1=1)/\,\mathbb{P}(Y^1=0)}{\mathbb{P}(Y^0=1)/\,\mathbb{P}(Y^0=0)}.$$

Causal ER, RR and OR can be expressed in terms of parameters of the following MSMs:

- $\mathbb{P}(Y^a=1) = \psi_0 + \psi_1 a \implies \text{Causal ER} = \psi_1$.
- $\log\bigl(\mathbb{P}(Y^a=1)\bigr) = \theta_0 + \beta_1 a \implies \text{Causal RR} = e^{\theta_1}$.
- $\operatorname{logit}\bigl(\mathbb{P}(Y^a=1)\bigr) = \beta_0 + \beta_1 a \implies \text{Causal OR} = e^{\beta_1}$.

Causal Diagram

TODO

Inverse Probability Weighting

The causal quantities can be consistently estimated in (b)–(c) by fitting the corresponding regression model ($Y \sim A$) with IPW weights:

$$w_i = \frac{1}{\mathbb{P}(A=A_i \mid X=X_i)}.$$

$$\hat{w}_i = \begin{cases} \dfrac{1}{\widehat{\mathsf{ps}}(X_i)}, & A_i = 1, \\ \dfrac{1}{1-\widehat{\mathsf{ps}}(X_i)}, & A_i = 0. \end{cases}$$

To avoid extreme weights, we may use stabilized weights:

$$sw_i = \frac{\mathbb{P}(A=A_i)}{\mathbb{P}(A=A_i \mid X=X_i)}.$$

- If we have a binary treatment and a binary outcome with a single time point, unstabilized and stabilized weights yield the same estimates;
- Otherwise, stabilized weights yield more efficient estimates.

4.2 Example 1: Why does IPW work?

	$X=1$		$X=0$	
	$A=1$	$A=0$	$A=1$	$A=0$
$Y=1$	108	24	20	40
$Y=0$	252	16	30	10
Total	360	40	50	50

- $\mathbb{P}(A=1 \mid X=1)=0.9$ and $\mathbb{P}(A=1 \mid X=0)=0.5$ both imply X and A are confounded.

Aggregated Data:

	$A=1$	$A=0$
$Y=1$	128	64
$Y=0$	282	26
Total	410	90

- Crude relative risk $=\mathbb{P}(Y=1 \mid A=1)/\,\mathbb{P}(Y=1 \mid A=0)=\frac{128/410}{64/90}=0.44$.

We can calculate causal relative risk by conditioning on X (assuming the ignorability assumption holds):

$$\begin{aligned}\mathbb{P}(Y^1=1) &= \mathbb{P}(Y=1 \mid A=1, X=1)\,\mathbb{P}(X=1)+\mathbb{P}(Y=1 \mid A=1, X=0)\,\mathbb{P}(X=0)\\ &= \frac{108}{360}\frac{4}{5}+\frac{20}{50}\frac{1}{5}\\ &= 0.32.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}(Y^0=1) &= \mathbb{P}(Y=1 \mid A=0, X=1)\,\mathbb{P}(X=1)+\mathbb{P}(Y=1 \mid A=0, X=0)\,\mathbb{P}(X=0)\\ &= \frac{24}{40}\frac{4}{5}+\frac{40}{50}\frac{1}{5}\\ &= 0.64.\end{aligned}$$

Therefore, Causal Relative Risk $=\mathbb{P}(Y^1=1)/\,\mathbb{P}(Y^0=1)=0.32/0.64=0.5$.

Another approach is performing Inverse Probability Weighting:

X	A	Y	n (observed)	$\mathbb{P}(A \mid X)$	w	N (pseudo)
1	1	1	108	0.9	1.11	120
1	1	0	252	0.9	1.11	280
1	0	1	24	0.1	10	240
1	0	0	16	0.1	10	160
0	1	1	20	0.5	2	40
0	1	0	30	0.5	2	60
0	0	1	40	0.5	2	80
0	0	0	10	0.5	2	20

Pseudopopulation Created by IPW with Binary A Stratified by the Confounder X:

	$X=1$		$X=0$	
	$A=1$	$A=0$	$A=1$	$A=0$
$Y=1$	120	240	40	80
$Y=0$	280	160	60	20
Total	400	400	100	100

- $\mathbb{P}(A=1 \mid X=1)=\mathbb{P}(A=1 \mid X=0)=0.5$ imply X and A are confounded.

Aggregated Data from the Pseudopopulation:

	$A = 1$	$A = 0$
$Y = 1$	160	320
$Y = 0$	340	180
Total	500	500

- Crude relative risk $= \mathbb{P}(Y = 1 \mid A = 1)/\,\mathbb{P}(Y = 1 \mid A = 0) = \frac{160/500}{320/500} = 0.5$.

4.3 Example 2: Fit MSM for Continuous Treatments

Early Dieting in Girls Study: a longitudinal study that aims to examine parental influences on daughters' growth and development from ages 5–15.

- Reference: Zhu, Y., Coffman, D. L., Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. Journal of causal inference, 3(1), 25-40.
- Treatment (M2WTCON):
  - mother's overall weight concern, calculated as the average score of five likert-scale (1-5) questions
  - measured at girls' age 7
  - which can be regarded as a continuous treatment variable
- Outcome (earlydiet): whether the daughter diets between ages 7 and 11, which is binary.
- Baseline confounders (50 covariates in total):
  - family history of diabetes and obesity, family income, etc
  - daughter's disinhibition, daughter's body esteem, etc
  - mother's perception of mother's current size and mother's satisfaction with daughter's current body, etc.

The following is one sample question in the weight concern questionnaire:

1. How afraid are you of gaining 3 pounds:
   (1) Not afraid
   (2) Slightly afraid
   (3) Moderately afraid
   (4) Very afraid
   (5) Terrified

- Fit a marginal structural model
$$\mathsf{logit}\big(\mathbb{P}(Y^a = 1)\big) = \beta_0^* + \beta_1^* a,$$
where $a \in [a_1, a_2]$ and β_1^* is the causal log odds ratio for 1 unit increase in the treatment (dosage) A.
- The causal effect β_1^* can be consistently estimated by fitting the corresponding regression model
$$\mathsf{logit}\big(\mathbb{P}(Y = 1)\big) = \beta_0 + \beta_1 a,$$
using IPW where
$$sw_i = \frac{f(A_i)}{f(A_i \mid X_i)},\ i = 1, \ldots, n,$$
where $f(\,\cdot\,)$is the unconditional density and $f(\,\cdot \mid \cdot\,)$ is the conditional density.

- Problem: when X is of even moderate dimension, estimate the conditional density is challenging.
- Solution: use normal approximation to estimate the unconditional and conditional density.
- Using normal density, we can estimate $f(A_i)$ by:

  $$\hat{f}(A_i) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left\{-\frac{(A_i - \hat{\mu})^2}{2\hat{\sigma}^2}\right\},$$

  where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and SD of $A_1, \ldots, A_n$.
- Assume

  $$A_i = X_i^\top \beta + \varepsilon_i,$$

  where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Let $r_i = A_i - X_i^\top \hat{\beta}$, we can approximate $f(A_i \mid X_i)$ by $\hat{f}(r_i \mid X_i)$:

  $$\hat{f}(A_i \mid X_i) = \hat{f}(r_i \mid X_i) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left\{-\frac{(A_i - X_i\hat{\beta})^2}{2\tilde{\sigma}^2}\right\},$$

  where $\tilde{\sigma}^2$ is the mean-squared error (MSE).

R Code, not public.

Week 6 | Monday

7th February

Review of chapters 1–4.

Week 6 | Wednesday

9th February

Chapter 5 Part I

Slide 8

$$\begin{aligned}
\mathsf{Cov}(Y, Z) &= \mathsf{Cov}(\beta_0 + \beta_1 A + v, Z) \\
&= \mathsf{Cov}(\beta_0, Z) + \mathsf{Cov}(\beta_1 A, Z) + \mathsf{Cov}(v, Z) \\
&= \mathsf{Cov}(\beta_1 A, Z) \\
&= \beta_1 \mathsf{Cov}(A, Z),
\end{aligned}$$

where $\mathsf{Cov}(\beta_0, Z) = 0$ since β_0 is a scalar, and $\mathsf{Cov}(v, Z) = 0$ by the exogeneity assumption. Re-arranging,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\widehat{\mathsf{Cov}}(Y, Z)}{\widehat{\mathsf{Cov}}(A, Z)} \\
&= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^n (A_i - \bar{A})(Z_i - \bar{Z})}
\end{aligned}$$

Bias of $\hat{\beta}$: By the Weak Law of Large Numbers (WLLN),

$$\hat{\beta}_1 \xrightarrow[n\to\infty]{\mathbb{P}} \frac{\mathsf{Cov}(Y, Z)}{\mathsf{Cov}(A, Z)}.$$

RHS:

$$\begin{aligned}\frac{\mathsf{Cov}(\beta_0+\beta_1 A+v,Z)}{\mathsf{Cov}(A,Z)} &= \frac{\beta_1\,\mathsf{Cov}(A,Z)+\mathsf{Cov}(v,Z)}{\mathsf{Cov}(A,Z)}\\ &= \beta_1+\frac{\mathsf{Cov}(v,Z)}{\mathsf{Cov}(A,Z)}\\ &= \beta_1+\frac{\mathsf{Corr}(v,Z)\sigma_v\sigma_Z}{\mathsf{Corr}(A,Z)\sigma_A\sigma_Z}\\ &= \beta_1+\underbrace{\frac{\mathsf{Corr}(v,Z)\sigma_v}{\mathsf{Corr}(A,Z)\sigma_A}}_{\text{asymptotic bias}}.\end{aligned}$$

If Z is a valid instrument, then $\mathsf{Corr}(v,Z)=0$, and so

$$\hat\beta_1 \xrightarrow[n\to\infty]{\mathbb{P}} \beta_1.$$

- $\mathsf{Corr}(v,Z)$ shows how bad the instrument is; we want this to be small.
- $\mathsf{Corr}(A,Z)$ shows us how weak the instrument is; we want this to be large.

Note that if OLS is used here ($Y\sim A$),

$$\hat\beta_{1,\text{OLS}}=\frac{\widehat{\mathsf{Cov}}(Y,A)}{\hat\sigma_A^2}\xrightarrow[n\to\infty]{\mathbb{P}}\frac{\mathsf{Cov}(Y,A)}{\sigma_A^2}.$$

RHS:

$$\begin{aligned}\frac{\mathsf{Cov}(\beta_0+\beta_1 A+v,A)}{\sigma_A^2} &= \frac{\beta_1\,\mathsf{Cov}(A,A)+\mathsf{Cov}(v,A)}{\sigma_A^2}\\ &= \beta_1+\frac{\mathsf{Cov}(v,A)}{\sigma_A^2}\\ &= \beta_1+\underbrace{\frac{\mathsf{Corr}(v,A)\sigma_V}{\sigma_A}}_{\text{asymptotic bias}}.\end{aligned}$$

If $\mathsf{Corr}(v,A)\neq 0$, then $\hat\beta_{1,\text{OLS}}\overset{\mathbb{P}}{\not\to}\beta_1$. If

$$\frac{\mathsf{Corr}(v,Z)}{\mathsf{Corr}(Z,A)}<\mathsf{Corr}(v,A),$$

then two-stage least squares (TSLS) is better than OLS (less asymptotic bias).

Week 7 | Monday

14th February ♡

Chapter 5 Part II

Slide 3

Under the assumption $\mathbb{E}[v^2] = \sigma^2$, we can derive the variance of $\hat{\beta}_1$:

$$\mathsf{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (A_i - \bar{A})} \times \frac{1}{R^2_{AZ}},$$

where R^2_{AZ} is the *coefficient of determination* in a regression of A on Z. Similarly,

$$\mathsf{Var}(\hat{\beta}_{1,\text{OLS}}) = \frac{\sigma^2}{\sum_{i=1}^n (A_i - \bar{A})}.$$

Remarks:

(1) $\mathsf{Var}(\hat{\beta}_1) > \mathsf{Var}(\hat{\beta}_{1,\text{OLS}})$ since $0 < R^2_{AZ} < 1$, so two stage least squares (TSLS) is less efficient than ordinary least squares (OLS).

(2) If Z is a weak instrument, then R^2_{AZ} is very small and $\hat{\beta}_1$ is highly inefficient.

Slide 4–5

Why does TSLS work?

(1) $\hat{A}$ is the projection of A onto the space spanned by Z, and $Z \perp\!\!\!\perp v$, implies $\hat{A} \perp\!\!\!\perp v$;

(2) $Y_i = \beta_0 + \beta_1 \hat{A}_i + \eta_i$, where $\hat{A}_i \perp\!\!\!\perp \eta_i$ implies OLS can be applied in the second stage.

Slide 6

(1) **Bootstrap approach.**

(i) Sample $b = 1, \ldots, B$ datasets of size n with replacement.

(ii) Apply TSLS to the bootstrapped sample, estimate $\hat{\beta}_1^{(b)}$, $b = 1, \ldots, B$.

(iii) Hence, the variance is given by

$$\mathsf{Var}(\hat{\beta}_1) = \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\beta}_1^{(b)} - \bar{\hat{\beta}}_1)^2,$$

where $\bar{\hat{\beta}}_1 = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_1^{(b)}$.

(2) **Derive the theoretical standard error.**

In matrix form, we have

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad A = \begin{bmatrix} 1 & A_1 \\ \vdots & \vdots \\ 1 & A_n \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & Z_1^\top \\ \vdots & \vdots \\ 1 & Z_n^\top \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \gamma = \begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}.$$

Hence,

$$\hat{\beta}_{\text{TSLS}} = (\hat{A}^\top \hat{A})^{-1} \hat{A}^\top Y,$$

where $\hat{A} = P_Z A = Z(Z^\top Z)^{-1} Z^\top A$. Note that P_Z is often called a *projection matrix*. Therefore,

$$\widehat{\mathsf{Var}}(\hat{\beta}_{\text{TSLS}}) = \hat{\sigma}^2 (\hat{A}^\top \hat{A})^{-1},$$

where the **theoretical standard variance** is given by

$$\hat{\sigma}^2 = \frac{(Y - A^\top \hat{\beta}_{\text{TSLS}})^\top (Y - A^\top \hat{\beta}_{\text{TSLS}})}{n-2}.$$

Slide 14

- H_0: A is exogenous, that is, $\text{Cov}(A, v) = 0$.
- If A is endogenous, that is, $\text{Cov}(v_i, \varepsilon_i) \neq 0$.
- If Z is a valid instrument, a consistent estimator of ε_i is $\hat{\varepsilon}_i$, which implies $\text{Cov}(v_i, \hat{\varepsilon}_i) \neq 0$, that is, $\text{Cov}(Y_i, \hat{\varepsilon}_i) \neq 0$.

WEEK 7 | WEDNESDAY

16th February

Case Study IV Analysis

Slide 22

$$\begin{aligned}\hat{\beta}_{\text{IVW}} &= \frac{\sum_{k=1}^K A_k Y_k / \sigma_{Yk}^2}{\sum_{k=1}^K A_k^2 / \sigma_{Yk}^2} \\ &= \frac{\displaystyle\sum_{k=1}^K \frac{Y_k}{A_k} \frac{A_k^2}{\sigma_{Yk}^2}}{\displaystyle\sum_{k=1}^K \frac{A_k^2}{\sigma_{Yk}^2}}.\end{aligned}$$

<ins>Remarks</ins>:

- The Y_k / A_k is known as the *Wald estimator*,

$$\begin{aligned}\frac{Y_k}{A_k} &= \frac{\widehat{\text{Cov}}(Y, Z_k)/\widehat{\text{Cov}}(Z_k, Z_k)}{\widehat{\text{Cov}}(A, Z_k)/\widehat{\text{Cov}}(Z_k, Z_k)} \\ &= \frac{\widehat{\text{Cov}}(Y, Z_k)}{\widehat{\text{Cov}}(A, Z_k)}.\end{aligned}$$

- The A_k^2 / σ_{Yk}^2 is derived by

$$\widehat{\text{Var}}\left(\frac{Y_k}{A_k}\right) = \frac{\sigma_{Yk}^2}{A_k^2}.$$

- In the simple case (one genetic), we get

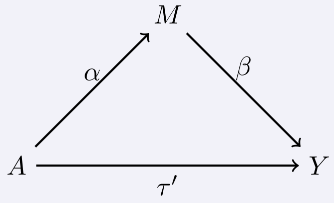
$$\hat{\beta}_{\text{IVW}} = \frac{Y_k}{A_k}.$$

WEEK 8 | MONDAY

27th February

Chapter 6 Part I

Slide 7



- τ': direct effect ($A \to Y$);
- $\alpha\beta$: indirect effect ($A \to M \to Y$);
- τ: total effect, that is, $\tau = \alpha\beta + \tau'$;
- M is the **intermediate** variable.

Slide 8

$$Y = \beta_1 + \tau A + \varepsilon_3 \tag{4.1}$$
$$Y = \beta_2 + \tau' A + \beta M + \varepsilon_2 \tag{4.2}$$
$$M = \beta_3 + \alpha A + \varepsilon_3 \tag{4.3}$$

Plug in (3) into (2),

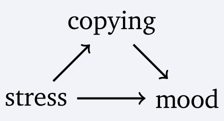
$$\begin{aligned} Y &= \beta_2 + \tau' A + \beta(\beta_3 + \alpha A + \varepsilon_3) + \varepsilon_2 \\ &= \underbrace{\beta_2 + \beta\beta_3}_{\beta_1} + \underbrace{(\tau' + \alpha\beta)}_{\tau} A + \underbrace{\beta\varepsilon_3 + \varepsilon_2}_{\varepsilon_1}, \end{aligned}$$

which yields $\tau = \tau' + \alpha\beta \implies \tau - \tau' = \alpha\beta$.

- $\tau - \tau'$: subtraction approach.
- $\alpha\beta$: product approach.

Slide 10

Why can the direct and indirect effect be opposite? (for the following diagram)



Slide 11

When testing the mediation effect, we can conduct the following hypothesis:

$$H_0\colon \alpha\beta = 0,$$

which is a *composite* hypothesis. That is, we will need to test

- $\alpha = 0$, $\beta \neq 0$;

- $\alpha \neq 0$, $\beta = 0$;
- $\alpha = 0$, $\beta = 0$.

To get the test statistic, we use the Delta method.

- $\mathbb{E}[X] \approx \hat{\mu}$ is consistent;
- $\mathsf{Var}(X) \approx \mathsf{se}^2(X)$.
- To get $\mathsf{Var}(g(X))$, using the first-order Delta method, we get

$$\mathsf{Var}(g(X)) \approx \left[g'(\hat{\mu})\right]^2 \mathsf{se}^2(X),$$

  assuming $g(\,\cdot\,)$ is differentiable.

In the bivariate case,

- $g(X, Y) = XY$, so that

$$\nabla g(X, Y) = \begin{bmatrix} \frac{\partial g}{\partial X} \\ \frac{\partial g}{\partial Y} \end{bmatrix} = \begin{bmatrix} Y \\ X \end{bmatrix}.$$

- We know that

$$\begin{aligned} \mathsf{Var}(\hat{\alpha}) &\approx \mathsf{se}^2(\hat{\alpha}), \\ \mathsf{Var}(\hat{\beta}) &\approx \mathsf{se}^2(\hat{\beta}). \end{aligned}$$

  Hence, using the first-order Delta method, we get

$$\begin{aligned} \mathsf{Var}(\hat{\alpha}\hat{\beta}) &\approx \begin{bmatrix} \hat{\beta} & \hat{\alpha} \end{bmatrix} \begin{bmatrix} \mathsf{se}^2(\hat{\alpha}) & 0 \\ 0 & \mathsf{se}^2(\hat{\beta}) \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} \\ &= \hat{\alpha}^2 \mathsf{se}^2(\hat{\beta}) + \hat{\beta}^2 \mathsf{se}^2(\hat{\alpha}). \end{aligned}$$

Chapter 6 Part II

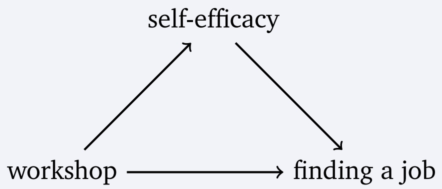
Slide 5

$$\begin{aligned} \text{CDE}_1 &= \mathbb{E}[Y^{1,1} - Y^{0,1}] \\ &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 - \alpha_0 - \alpha_2 \\ &= \alpha_1 + \alpha_3. \end{aligned}$$

Week 8 | Wednesday

2nd March

Slide 3



If M is binary,

- $\text{CDE}_1 = \mathbb{E}[Y^{1,1} - Y^{0,1}]$;
- $\text{CDE}_0 = \mathbb{E}[Y^{1,0} - Y^{0,0}]$.

If there is no interaction between A and M, then

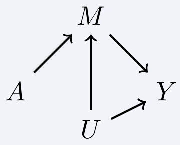
$$\text{CDE} = \text{CDE}_1 = \text{CDE}_0.$$

Slide 5

Since we have two time points now and using the diagram on **slide 4**, we see that

$$\begin{aligned} sw_i &= sw_i(1) \times sw_i(2) \\ &= \frac{\mathbb{P}(A = A_i)}{\mathbb{P}(A = A_i \mid X = X_i)} \times \frac{\mathbb{P}(M = M_i \mid A = A_i)}{\mathbb{P}(M = M_i \mid A = A_i, X = X_i, W = W_i)}. \end{aligned}$$

Slide 7



- A: treatment being assigned;
- M: treatment actually received;
- Y: outcome;
- U: age.

We lose the ability to draw a causal inference between A and Y.

Slide 10

Using the conditional probabilities, the total effect is

$$\mathbb{E}[Y^1 - Y^0] = \text{CACE}p_C + \text{DACE}p_D + \text{AACE}p_A + \text{NACE}p_M.$$

- The exclusion restriction implies AACE $= \text{NACE} = 0$.
- The monotocity assumption implies $p_D = 0$.
- The third assumption implies $\mathbb{E}[Y^1 - Y^0] = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$.

Re-arranging,

$$\widehat{\text{CACE}} = \frac{\widehat{\mathbb{E}}[Y^1 - Y^0]}{\hat{p}_C} = \frac{\widehat{\mathbb{E}}[Y \mid A = 1] - \widehat{\mathbb{E}}[Y \mid A = 0]}{\hat{p}_C}.$$

Looking at the quantity

$$\mathbb{P}(M = 1 \mid A = 1) - \mathbb{P}(M = 1 \mid A = 0),$$

we see that

- $\mathbb{P}(M = 1 \mid A = 1)$ are the always takers and compliers, and
- $\mathbb{P}(M = 1 \mid A = 0)$ are the always takers (and defiers; although there is no defiers in this case).

Hence,

$$\hat{p}_C = \widehat{\mathbb{E}}[M \mid A = 1] - \widehat{\mathbb{E}}[M \mid A = 0],$$

which yields the same formula for $\widehat{\text{CACE}}$ as **slide 11**.

Slide 18

If $A_i = 1$, we need to first impute M_i^0 with

$$M \sim A + X,$$

and then impute Y_i^{1,M_i^0}, Y_i^{0,M_i^1}, Y_i^{0,M_i^0} with

$$Y \sim M + A + X.$$

Case Study III: Mediation Analysis

Slide 6

The output line-by-line is:

- NIE_0;
- NIE_1;
- NDE_0;
- NDE_1;
- TE;
- NIE_0/TE;
- NIE_1/TE,

and rest are unimportant.

Week 9 | Monday

7th March

Midterm Project

- Binary Y: ACE_2 is the coefficient for A in `lm(Y~A,weights=IPTW)` which gives the causal excess risk.
- Binary Y: `glm(Y~A,weights=IPTW,family=binomial)` gives the causal log odds ratio (coefficient for A).
- Count Y: use the `poisson` family, which gives causal log relative risk (coefficient for A).
- DR estimator log odds ratio is the coefficient for A: `glm(Y~A+X,weights=IPTW,family=binomial)`, which is consistent if either this glm is correctly specified or if the PS model is correctly specified.

Case Study III: Mediation Analysis

Slide 12

$$
\begin{aligned}
sw_i &= sw_i(1) \times sw_i \\
&= \underbrace{\frac{\mathbb{P}(A = A_i)}{\mathbb{P}(A = A_i \mid X = X_i)}}_{\text{w1}} \underbrace{\frac{\mathbb{P}(M = M_i \mid A = A_i)}{\mathbb{P}(M = M_i \mid A = A_i, X = X_i)}}_{\text{w2}}.
\end{aligned}
$$

For example, in the slides `w1.num` gives $\mathbb{P}(A = A_i)$. We talk about categorical M, but if M is continuous, we replace the probability by the density (similar to A2).

Chapter 7 Part I

Covered slides 1–23.

Week 9 | Wednesday

9th March

Chapter 7 Part I

Suppose we are interested in evaluating Yeying. Some examples of missingness:

- MCAR: student forgot to come to class;
- MAR: student did not come to class because the weather was bad;
- MNAR: student did not come to class because he thinks that the professor is bad at teaching.

Chapter 7 Part II

Slide 10

Suppose that Y has missing values. For

$$\frac{\mathrm{Cov}(X, Y)}{\mathrm{Cov}(X, X)},$$

we can calculate $\mathrm{Cov}(X, X)$ because there is no missingness in X. However, we have an unknown sample size for Y, so pairwise deletion might not be appropriate.

Covered slide up to and including 29.

Week 10 | Monday

14th March

Chapter 7 Part II

Finished the module.

Chapter 7 Part III

Slide 8

$\tilde{\gamma} \sim \text{MVN}(\hat{\gamma}, V)$.

Covered all the slides.

Theoretical Part

Notations:

- Y: disease status;
- A: exposure status;
- X: a single covariate (for simplicity);
- Missing indicator:
$$R = \begin{cases} 1, & \text{if } X \text{ is missing,} \\ 0, & \text{if } X \text{ is observed;} \end{cases}$$
- Missing mechanism:
$$q(y, a, x) = \mathbb{P}(R = 1 \mid Y = y, A = a, X = x).$$
  - If q does not depend on x or y, we have missing completely at random (MCAR).
  - If q does not depend on x, we have missing at random (MAR).
  - Otherwise, we have missing not at random (MN AR).
- Fit logistic regression with missing data. Given
$$\mathbb{P}(Y = 1 \mid A = a, X = x) = f(\beta_0 + \beta_A a + \beta_X x),$$
where $f(t) = \mathsf{expit}(t) = \mathsf{exp}\{t\}/\big(1 + \mathsf{exp}\{t\}\big)$. We can show that
$$\mathbb{P}(Y = 1 \mid A = a, X = x, R = 0) = f\left(\beta_0 + \mathsf{log}\left(\frac{1 - q(1, a, x)}{1 - q(0, a, x)}\right) + \beta_A a + \beta_X x\right). \tag{1}$$
Our missing data equation is:
$$\mathbb{P}(Y = 1 \mid A = a, X = x, R = 1) = f\left(\beta_0 + \mathsf{log}\left(\frac{q(1, a, x)}{q(0, a, x)}\right) + \beta_A a + \beta_X x\right).$$
Hence, for the missing data, our model is
$$\mathbb{P}(Y = 1 \mid A = a, R = 1) = \int_x f\left(\beta_0 + \mathsf{log}\left(\frac{q(1, a, x)}{q(0, a, x)}\right) + \beta_A a + \beta_X x\right) \mathrm{d}F(x \mid A = a, R = 1). \tag{2}$$

1. Complete data analysis:
   - If MCAR, $q(1, a, x) = q(0, a, x)$. Hence,
$$\mathbb{P}(Y = 1 \mid A = a, X = x) = \mathbb{P}(Y = 1 \mid A = a, X = x, R = 0).$$
   In other words, the model we get for the missing data is consistent for the true data.

- If $q(y,a,x) = q(a,x)$ (in other words, the missingness does not depend on the disease outcome), then
$$\mathbb{P}(Y = 1 \mid A = a, X = x) = \mathbb{P}(Y = 1 \mid A = a, X = x, R = 0).$$
Note: in case-control studies, the missingness often depends on the disease outcome. In cohort studies, the missingness does not depend on the disease outcome.
- If $q(y,a,x) = q(y)q(a,x)$ (i.e., the missingness does not jointly depend on Y and (X,A)), one can still get valid (consistent) estimates of β_A and β_X. However, the intercept will change to
$$\beta_0^\star = \log\left(\frac{q(1)}{q(0)}\right).$$

2. Missing indicator approach:
   - Add the missing indicator R to the regression model.
   - This approach does not work in the general case even when MCAR. Why? Suppose for a contradiction we assume MCAR, that is, $q(y,a,x) = q$. Then,
   $$\mathbb{P}(Y = 1 \mid A = a, X, R = 0) = f(\beta_0 + \beta_A a + \beta_X x + \beta_R 0) = f(\beta_0 + \beta_A a + \beta_X x). \tag{a}$$
   The missing values are given by
   $$\mathbb{P}(Y = 1 \mid A = a, X = c, R = 1) = f(\beta_0 + \beta_A a + \beta_X c + \beta_R) = f(\beta_0^\star + \beta_A a), \tag{b}$$
   where $\beta_0^\star = \beta_0 + \beta_X c + \beta_R$.
     - β_A in (a) corresponds to the adjusted OR of the exposure variable.
     - β_A in (b) corresponds to the unadjusted OR of the exposure variable.
     - $\exp\{\hat{\beta}_A\}$ obtained from this approach lies between the adjusted and unadjusted odds ratios. Therefore, β_A is an *inconsistent* estimator of the odds ratio.

Next time, we will prove (1).

WEEK 10 | WEDNESDAY
16th March

Theoretical Part

No slides today, writing for 1h 20min. Recall that
$$q(y,a,x) = \mathbb{P}(R = 1 \mid Y = y, A = a, X = x).$$

Given
$$\mathbb{P}(Y = 1 \mid A = a, X = x) = f(\beta_0 + \beta_a + \beta_X x) = f(\beta_0 + \beta_A a + \beta_X x)$$
where $f(t) = \text{expit}(t)$, we can show (1):

$$\begin{aligned}
&\mathbb{P}(Y = 1 \mid A = a, X = x, R = 0) \\
&= \frac{\mathbb{P}(R = 0, Y = 1 \mid A = a, X = x)}{\mathbb{P}(R = 0 \mid A = a, X = x)} \\
&= \frac{\mathbb{P}(R = 0 \mid Y = 1, A = a, X = x)\,\mathbb{P}(Y = 1 \mid A = a, X = x)}{\mathbb{P}(R = 0, Y = 1 \mid A = a, X = x) + \mathbb{P}(R = 0, Y = 0 \mid A = a, X = x)} \\
&= \frac{\mathbb{P}(R = 0 \mid Y = 1, A = a, X = x)\,\mathbb{P}(Y = 1 \mid A = a, X = x)}{\mathbb{P}(R = 0 \mid Y = 1, A = a, X = x)\,\mathbb{P}(Y = 1 \mid A = a, X = x) + \mathbb{P}(R = 0 \mid Y = 0, A = a, X = x)\,\mathbb{P}(Y = 0 \mid A = a, X = x)} \\
&= \frac{[1 - q(1,a,x)]f(\beta_0 + \beta_A a + \beta_X x)}{[1 - q(1,a,x)]f(\beta_0 + \beta_A a + \beta_X x) + [1 - q(0,a,x)][1 - f(\beta_0 + \beta_A a + \beta_X x)]} \\
&= f\left(\beta_0 + \log\left(\frac{1 - q(1,a,x)}{1 - q(0,a,x)}\right) + \beta_A a + \beta_X x\right).
\end{aligned}$$

1. Complete data analysis.
2. Missing indicator approach.
3. Single imputation method.
   - Replace each missing value by $\bar{X}$.
   - A better way is to replace the missing value by $\bar{X} \mid A$, that is, the mean of the observed X given A.
   - Why? For simplicity, assume $q(y, a, x) = q(a)$. By (1), we have

   $$\mathbb{P}(Y = 1 \mid A = a, X = x, R = 0) = \mathbb{P}(Y = 1 \mid A = a, X = x).$$

   By (2), we have

   $$\begin{aligned} \mathbb{P}(Y = 1 \mid A = a, R = 1) &= \int_x f(\beta_0 + \beta_A a + \beta_X x) \, \mathrm{d}F(x \mid A = a,\, R = 1) \\ &= \int_x f(\beta_0 + \beta_A a + \beta_X x) \, \mathrm{d}F(x \mid A = a) \text{ since } R \perp\!\!\!\perp X \mid A \\ &= f\left(\beta_0 + \beta_A a + \int_x \beta_X x \, \mathrm{d}F(x \mid A = a)\right) \text{ if } f \text{ as an approx. linear function} \\ &= f\big(\beta_0 + \beta_A a + \beta_X \, \mathbb{E}[X \mid A = a]\big) \end{aligned}$$

4. Inverse probability weighting.
   - Weigh each subject with complete data by

   $$\frac{1}{1 - \hat{q}(y_i, a_i, x_i)}, \; i = 1, \ldots, n, \; R_i = 0.$$

   - For logistic regression, the score function for MLE is

   $$S_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} S_\beta(y_i, a_i, x_i),$$

   where

   $$S_\beta(y, a, x) = \frac{\partial}{\partial \beta} \{ y \log(f_\beta(a, x)) + (1 - y) \log(1 - f_\beta(a, x)) \}.$$

   - For incomplete data, $S_\beta(y_i, a_i, x_i)$ is unknown for any subject with $R_i = 1$.
   - We use the Horvitz-Thompson Estimator: solve β by letting $\tilde{S}_n(\beta) = 0$, where

   $$\begin{aligned} \tilde{S}_n(\beta) &= \frac{1}{n} \sum_{i=1, R_i = 0}^{n} \frac{S_\beta(y_i, a_i, x_i)}{1 - \hat{q}(y_i, a_i, x_i)} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{S_\beta(y_i, a_i, x_i)(1 - R_i)}{1 - \hat{q}(y_i, a_i, x_i)}. \end{aligned}$$

   - Works well if MAR and q is properly modelled (i.e., use q instead of $\hat{q}$). Why? We will show that

   $$\mathbb{E}\left[\frac{S_\beta(Y, A, X)(1 - R)}{1 - q(Y, A, X)}\right] = \mathbb{E}\big[S_\beta(Y, A, X)\big],$$

   that is, we will have a consistent estimator. Note that MAR implies
     - q does not depend on X;
     - $R \perp\!\!\!\perp X \mid Y, A$.

$$\begin{aligned}
\mathbb{E}\left[\frac{S_\beta(Y,A,X)(1-R)}{1-q(Y,A,X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{S_\beta(Y,A,X)(1-R)}{1-q(Y,A)}\,\middle|\, Y,A\right]\right] \text{ since } R \perp\!\!\!\perp X \mid Y,A \\
&= \mathbb{E}\left[\mathbb{E}\big[S_\beta(Y,A,X \mid Y,A)\big]\frac{\mathbb{E}\big[(1-R)\mid Y,A\big]}{1-q(Y,A)}\right] \text{ since } R \perp\!\!\!\perp X \mid Y,A \\
&= \mathbb{E}\left[\mathbb{E}\big[S_\beta(Y,A,X \mid Y,A)\big]\frac{1-q(Y,A)}{1-q(Y,A)}\right] \\
&= \mathbb{E}\big[S_\beta(Y,A,X)\big].
\end{aligned}$$

5. Multiple imputation.

   - Perform imputation m times.
   - Rubin's rule gives

     $$\hat{\beta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j,$$

     $$\begin{aligned}
     \widehat{\mathsf{Var}}(\hat{\beta}) &= \frac{1}{m}\sum_{j=1}^{m}s_j^2 + \left(1+\frac{1}{m}\right)\frac{1}{m-1}\sum_{j=1}^{m}(\hat{\beta}_k-\hat{\beta})^2 \\
     &= \underbrace{\frac{1}{m}\sum_{j=1}^{m}s_j^2}_{\bar{U}} + \underbrace{\frac{1}{m-1}\sum_{j=1}^{m}(\hat{\beta}_j-\hat{\beta})^2}_{B} + \underbrace{\frac{1}{m(m-1)}\sum_{j=1}^{m}(\hat{\beta}_j-\hat{\beta})^2}_{(1/m)B}.
     \end{aligned}$$

     - $\bar{U}$: within imputation variance;
     - B: between imputation variance;
     - $(1/m)B$: simulation error (i.e., extra variability as a consequence of imputing the missing data using a finite number of imputations instead of an infinite number of times).
   - Define

     $$T = \bar{U} + \left(1+\frac{1}{m}\right)B, \qquad \lambda = \frac{B+B/m}{T},$$

     where λ is known as the proportion of variation attributed to the missing data. Furthermore,

     $$r = \frac{B+B/m}{\bar{U}},$$

     where r is known as the relative increase in variance due to the missing data.
   - ν: degrees of freedom.
     - Rubin (1978):

       $$\nu_{\text{old}} = (m-1)\left(1+\frac{1}{r^2}\right) = \frac{m-1}{\lambda^2}.$$

       Problem: If $\lambda = 0$ (i.e., we have complete data), then ν_{old} is indeterminate (which is clearly inappropriate).
     - Banard and Rubin (1999) (used in the mice package):

       $$\nu_{\text{com}} = n-k,$$

       where n is the sample size, and k is the number of parameters. Also,

       $$\nu_{\text{obs}} = \frac{\nu_{\text{com}}}{\nu_{\text{com}}+3}\nu_{\text{com}}(1-\lambda).$$

- The newly proposed degrees of freedom is then given by

$$\nu_{\text{new}} = \frac{\nu_{\text{old}}\nu_{\text{obs}}}{\nu_{\text{old}} + \nu_{\text{obs}}},$$

  where we see that $\nu_{\text{new}} \leq \nu_{\text{com}}$, and if $\lambda = 0$, then $\nu_{\text{new}} = \nu_{\text{com}}$.
- H_0: $\beta = \beta_0$, H_A: $\beta \neq \beta_a$:

$$T = \frac{\hat{\beta} - \beta_0}{\sqrt{\widehat{\mathsf{Var}}(\beta)}} \dot{\sim} t(\nu_{\text{new}}),$$

  under H_0. Furthermore, $T^2 \sim F(1, \nu_{\text{new}})$ under H_0 (which is equivalent, but used in some software).

WEEK 11 | MONDAY

21st March

6. Maximum Likelihood Estimation

- Assume (X_i, Y_i), $i = 1, \ldots, n$.

(1) Y is fully observed; the first m components of X are observed, but the last $(n - m)$ components of X are missing.

(2) The joint pdf/pmf of (X, Y) is $f(x, y)$.

(3) The parameter(s) of interest is θ.

Research Question: We want to estimate some (or all) components of θ.

- Then, the joint (observed) likelihood for θ is based on the observed data is:

$$\mathcal{L}(x, y; \theta) = \prod_{i=1}^{m} f(x_i, y_i; \theta) \prod_{j=m+1}^{n} g(y_i; \theta), \tag{4.4}$$

  where $g(y_i; \theta) = \sum_x f(x, y; \theta)$ if X is discrete, or $g(y_i) = \int_x f(x, y; \theta)\, \mathrm{d}x$ if X is continouous.
- θ is solved by maximizing the joint likelihood $\mathcal{L}(x, y; \theta)$.

Missing Data for Contingency Table (MLE)

Suppose we have a cross-sectional study with the focus on the association between mother's marital status and babies low birthweight status.

	Birthweight (Y)		
Martial Status (X)	Low	Normal	
Unmarried	12 (p_{11})	68 (p_{12})	80
Married	5 (p_{21})	95 (p_{22})	100
	17	163	180

- We further assume the babies birthweight is missing for 5 married mothers, and 15 unmarried mothers.
- If a complete case analysis is conducted, then the likelihood is:

$$\mathcal{L} = p_{11}^{12} p_{12}^{68} p_{21}^{5} p_{22}^{95},$$

under the constraint $p_{11}+p_{12}+p_{21}+p_{22}=1$. We get

$$\hat{p}_{ij}=\frac{n_{ij}}{n^\star},$$

where $n^\star=\sum_i\sum_j n_{ij}$. Hence,

$$\hat{p}_{11}=\frac{12}{180}=0.063,\quad \hat{p}_{12}=0.378,\quad \hat{p}_{21}=0.028,\quad \hat{p}_{22}=0.533.$$

- If the information is missing completely at random, then we have a simple random sample, and by maximizing our likelihood we will have an asymptotically unbiased estimator by the properties of the likelihood function. By maximizing the likelihood, the MLE of $\theta=(p_{11},\ldots,p_{22})$ is biased if the missing mechanism is not MCAR.
- If we also consider modelling the missing outcome, then we know that

$$p_{11}=\mathbb{P}(\text{unmarried},\text{low}),$$

$$p_{12}=\mathbb{P}(\text{unmarried},\text{normal}),$$

so that $\mathbb{P}(\text{unmarried})=p_{11}+p_{12}$. Therefore, our new likelihood is:

$$\mathcal{L}=p_{11}^{12}p_{12}^{65}p_{21}^{5}p_{22}^{95}(p_{11}+p_{12})^{15}(p_{21}+p_{22})^{5},$$

subject to $p_{11}+p_{12}+p_{21}+p_{22}=1$. By maximizing this likelihood, under regularity conditions of MLE and MAR assumption, the MLE of θ is asymptotically unbiased (consistent), efficient and normally distributed. Maximizing this likelihood is not easy, and requires a Newton Raphson method. We will try sequentially.

$$\hat{p}_{ij}=\widehat{\mathbb{P}}(X=i,Y=j)=\widehat{\mathbb{P}}(X=i)\,\widehat{\mathbb{P}}(Y=j\mid X=i).$$

That is,

$$\hat{p}_{11}=\widehat{\mathbb{P}}(X=1,Y=1)=\widehat{\mathbb{P}}(X=1)\,\widehat{\mathbb{P}}(Y=1\mid X=1)=\frac{80+15}{200}\times\frac{12}{80}=0.071.$$

Similarly,

$$\hat{p}_{12}=\widehat{\mathbb{P}}(X=1)\,\widehat{\mathbb{P}}(Y=2\mid X=1)=\frac{80+15}{200}\times\frac{68}{80}=0.404.$$

$$\hat{p}_{21}=\widehat{\mathbb{P}}(X=2)\,\widehat{\mathbb{P}}(Y=1\mid X=2)=\frac{100+5}{200}\times\frac{5}{100}=0.026.$$

$$\hat{p}_{22}=\widehat{\mathbb{P}}(X=2)\,\widehat{\mathbb{P}}(Y=2\mid X=2)=\frac{100+5}{200}\times\frac{95}{100}=0.499.$$

Expectation-Maximization (EM) Algorithm

The EM algorithm is a general iterative approach to getting maximum likelihood estimates with missing data. It always follows (next lecture we do it in R):

(1) Fill in the missing values by their estimated values. Now, we have complete data.

(2) Estimate the parameters for this “complete” dataset.

(3) Use the estimated parameters to re-estmate the missing values (called an E-step). We fill in the missing values with their *expected* values, given the observed data (**E**xpectation-step).

(4) Re-estimate the parameters from this updated “complete” dataset (M-step). We *maximize* the

likelihood to estimate the parameters (**M**aximization-step).

Iterate between steps (3) and (4) until convergence of the parameter estimates.

Expectation-Maximization (Bivariate Normal Data)

Consider

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim \text{BVN}(\mu, \Sigma),$$

where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix},$$

where Σ is a symmetric matrix. We observe $(y_1, \ldots, y_n)$ and $(x_1, \ldots, x_m)$, where $m < n$. The goal is to estimate the mean μ with two approaches:

(1) Direct computation of MLE;

(2) EM algorithm.

In the BVN example, these two approaches are equivalent. Some properties from STAT 330: If $\begin{bmatrix} Y \\ X \end{bmatrix} \sim$ BVN(μ, Σ), then

- $Y \sim \mathcal{N}(\mu_1, \sigma_{11}^2)$ and $X \sim \mathcal{N}(\mu_2, \sigma_{22}^2)$;
- Conditional distribution:

$$X \mid Y = y \sim \mathcal{N}\Bigg(\underbrace{\mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(y - \mu_1),}_{\mathbb{E}[x_i|y_i]} \quad \underbrace{\sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2}}_{\mathbb{E}[x_i^2|y_i] - \mathbb{E}[x_i|y_i]^2} \Bigg).$$

Week 12 | Monday

28th March

Direct Computation of MLE

$$\begin{aligned}
\mathcal{L}(x, y; \theta) &= \prod_{i=1}^{m} \underbrace{f(x_i, y_i; \theta)}_{g(y_i;\theta)h(x_i|y_i;\theta)} \prod_{i=m+1}^{n} g(y_i; \theta) \\
&= \prod_{i=1}^{n} g(y_i; \theta) \prod_{i=1}^{m} h(x_i \mid y_i; \theta) \\
&= \prod_{i=1}^{n} g(y_i; \mu_1, \sigma_{11}^2) \prod_{i=1}^{m} h(x_i \mid y_i; \mu, \Sigma) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma_{11}^2}}\right)^n \text{exp}\left\{-\frac{\sum_{i=1}^n (y_i - \mu_1)^2}{2\sigma_{11}^2}\right\} \left(\frac{1}{\sqrt{2\pi(\sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2})}}\right)^m \text{exp}\left\{-\frac{\sum_{i=1}^m (x_i - \mu_2 - \frac{\sigma_{12}}{\sigma_{11}}(y_i - \mu_1))^2}{2(\sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2})}\right\}.
\end{aligned}$$

Ignoring constants, we have:

$$\ell(x,y;\mu,\Sigma) = -\frac{n}{2}\mathsf{log}(\sigma_{11}^2) - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i-\mu_1)^2}{\sigma_{11}^2} - \frac{m}{2}\mathsf{log}\left(\sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2}\right) - \frac{1}{2}\sum_{i=1}^{m} \frac{[x_i - \mu_2 - \frac{\sigma_{12}}{\sigma_{11}}(y_i-\mu_1)]^2}{\sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2}}. \tag{3}$$

Solving MLE, we obtain

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$
$$\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}_1,$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^m (y_i - \bar{y}^*)(x_i - \bar{x}^*)}{\sum_{i=1}^m (y_i - \bar{y}^*)^2},$$
$$\hat{\beta}_0 = \bar{x}^* - \hat{\beta}_1\bar{y}^*.$$

where the $*$ is for complete data, that is,

$$\bar{x}^* = \frac{1}{m}\sum_{i=1}^{m} x_i,$$
$$\bar{y}^* = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

EM Algorithm

Idea: sometimes, maximizing the "observed" likelihood (3) is not easy and it's easier to maximize the "complete" likelihood if we can fill in the missing values first. Under bivariate normal assumption,

$$L_C(x,y;\mu,\Sigma) = \prod_{i=1}^{n} f(x_i,y_i;\mu,\Sigma).$$
$$\ell_C(x,y;\mu,\Sigma) = -\frac{n}{2}\mathsf{log}(\mathsf{det}(\Sigma)) - \frac{1}{2}\sum_{i=1}^{n}\left(\begin{bmatrix} y_i \\ x_i \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\Sigma^{-1}\left(\begin{bmatrix} y_i \\ x_i \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right).$$

E-step: the sufficient statistic(s) for bivariate normal is given by

$$S_1 = \sum_{i=1}^{n} y_i, \qquad S_2 = \sum_{i=1}^{n} x_i, \qquad S_{11} = \sum_{i=1}^{n} y_i^2, \qquad S_{22} = \sum_{i=1}^{n} x_i^2, \qquad S_{12} = \sum_{i=1}^{n} x_i y_i.$$

We can impute the values by computing

$$\mathbb{E}[x_i \mid y_i;\mu,\Sigma] = \beta_0 + \beta_1 y_i,$$
$$\mathbb{E}[x_i^2 \mid y_i;\mu,\Sigma] = (\beta_0 + \beta_1 y_i)^2 + \sigma_{22\cdot 1},$$
$$\mathbb{E}[x_i y_i \mid y_i;\mu,\Sigma] = (\beta_0 + \beta_1 y_i)y_i.$$

where

$$\beta_0 = \mu_2 - \frac{\sigma_{12}}{\sigma_{11}}\mu_1, \qquad \beta_1 = \frac{\sigma_{12}}{\sigma_{11}}, \qquad \sigma_{22\cdot 1} = \sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2}.$$

Given the input of μ, Σ, we can obtain $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_{22\cdot 1}$.

M-step:

$$
\begin{aligned}
\hat{\mu}_1 &= \frac{S_1}{n} = \frac{\sum_{i=1}^n y_i}{n}, \\
\hat{\mu}_2 &= \frac{S_2}{n} = \frac{\sum_{i=1}^n x_i}{n}, \\
\hat{\sigma}_{11}^2 &= \frac{S_{11}}{n} - \hat{\mu}_1^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}, \\
\hat{\sigma}_{22}^2 &= \frac{S_{22}}{n} - \hat{\mu}_2^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \\
\hat{\sigma}_{12} &= \frac{S_{12}}{n} - \hat{\mu}_1\hat{\mu}_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}.
\end{aligned}
$$

Repeat the E-step and M-step until all the parameter estimates converge.