STAT6061/STAT5008 – Causal Inference

Part 1-3. Assumptions and Identification

An-Shun Tai

[1]*Department of Statistics*
National Cheng Kung University

[2]*Institute of Statistics and Data Science*
National Tsing Hua University

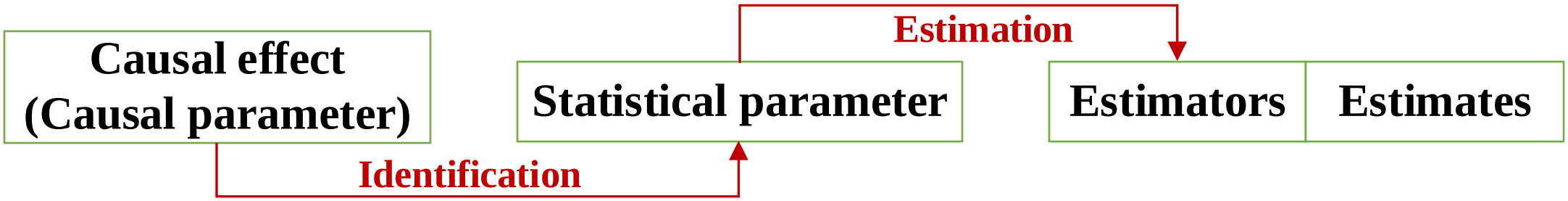
Identification and assumptions

- Causal effects (causal parameters) are functions of *potential outcomes*, whereas statistical parameters are functions of the distribution of *observed data*.

- *Identification* – the key step in causal inference – is the process of linking causal parameters to statistical parameters derived from observed data.

- However, there is no free lunch; the identification process requires certain assumptions, known as *identification assumptions*.

- The key identifying assumptions pertain to the *treatment assignment mechanism*.



Treatment assignment mechanism

- The fundamental bridge between the potential outcomes $(Y_i(0), Y_i(1))$ and observed outcome Y_i :

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

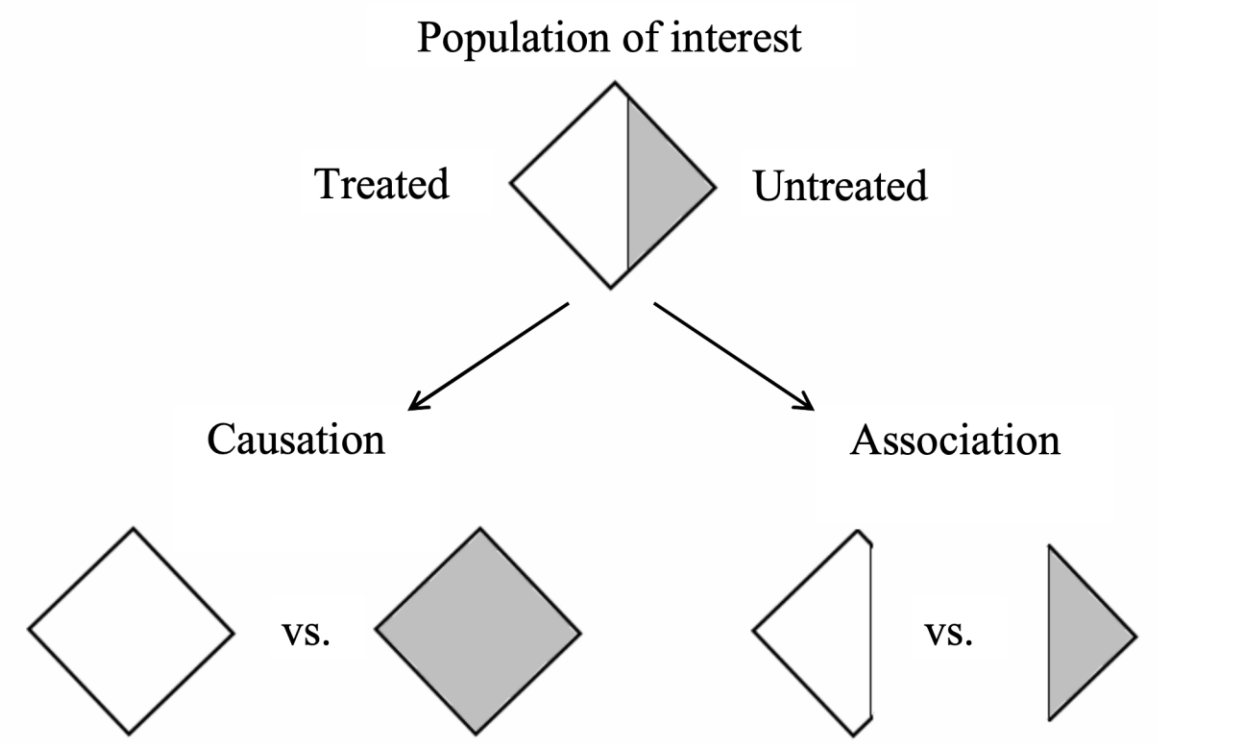
Note that, in this part, A_i is the treatment assignment indicator for unit i.

- The difference in means

$$\begin{aligned} & \mathbb{E}(Y_i|A_i = 1) - \mathbb{E}(Y_i|A_i = 0) \\ = & \mathbb{E}(Y_i(1)|A_i = 1) - \mathbb{E}(Y_i(0)|A_i = 0) \\ = & \mathbb{E}(Y_i(1) - Y_i(0)|A_i = 1) + \{\mathbb{E}(Y_i(0)|A_i = 1) - \mathbb{E}(Y_i(0)|A_i = 0)\} \end{aligned}$$

- The first part is the ATT, while the second captures *hidden bias* arising from the *treatment assignment mechanism*, leading to characteristic differences between treated and untreated groups.

Treatment assignment mechanism



$\mathbb{E}\big(Y(1)\big)$ vs. $\mathbb{E}\big(Y(0)\big)$

ATT/ATE

$\mathbb{E}(Y|A=1)$ vs. $\mathbb{E}(Y|A=0)$

difference in means

Treatment assignment mechanism

- This demonstrates that a *well-designed* treatment assignment mechanism can eliminate hidden bias, i.e.,

$$\mathbb{E}(Y_i(0)|A_i = 1) - \mathbb{E}(Y_i(0)|A_i = 0) = 0,$$

allowing the naïve difference-in-means approach to accurately estimate the causal effect (ATE or ATT).

- How does causal inference differ from association inference?
  - ✓ Epidemiological perspective: addressing confounding variables/common causes
  - ✓ Clinical perspective: necessity of intervention
  - ✓ Statistical perspective: controlling for hidden bias
  - ✓ Experimental perspective: importance of treatment assignment mechanism

Probabilistic rule for treatment assignment mechanism

- The assignment mechanism is a probabilistic rule that determines the probabilities of all 2^N possible assignment vectors $\widetilde{\boldsymbol{A}} = (A_1, \ldots, A_N)$ for N units, given potential outcomes $(\widetilde{\boldsymbol{Y}}(0), \widetilde{\boldsymbol{Y}}(1))$ and covariates $(\widetilde{\boldsymbol{C}})$.

Definition (Treatment Assignment Mechanism) (Imbens and Rubin, 2015)
Given a population of N units, the assignment mechanism is a row-exchangeable function

$$\Pr\left(\widetilde{\boldsymbol{A}} \middle| \widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{Y}}(0), \widetilde{\boldsymbol{Y}}(1)\right)$$

taking on values in [0, 1], satisfying

$$\sum_{\widetilde{\boldsymbol{a}} \in \{0,1\}^N} \Pr\left(\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{a}} \middle| \widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{Y}}(0), \widetilde{\boldsymbol{Y}}(1)\right) = 1$$

for all $\widetilde{\boldsymbol{C}}$, $\widetilde{\boldsymbol{Y}}(0)$, *and* $\widetilde{\boldsymbol{Y}}(1)$.

- The treatment assignment mechanism will be discussed in detail in Part 2.1.

Examples for two units

- Ignoring $\widetilde{\boldsymbol{C}}$, suppose $N = 2$. The $2^2 = 4$ possible assignment vectors $\widetilde{\boldsymbol{A}}$ are given by

$$\Omega = \{(0,0), (1,0), (0,1), (1,1)\}$$

Example 1 (clueless doctor).

$$Pr\left(\widetilde{\boldsymbol{A}} \middle| \widetilde{\boldsymbol{Y}}(0), \widetilde{\boldsymbol{Y}}(1)\right) = \frac{1}{4}, \qquad \widetilde{\boldsymbol{A}} \in \Omega$$

Example 2 (perfect doctor).

$$Pr\left(\widetilde{\boldsymbol{A}} \middle| \begin{matrix} Y_1(1) - Y_1(0) > 0, \\ Y_2(1) - Y_2(0) > 0 \end{matrix}\right) = \begin{cases} 1 & , \widetilde{\boldsymbol{A}} = (1,1) \\ 0 & , O.W. \end{cases}, \qquad Pr\left(\widetilde{\boldsymbol{A}} \middle| \begin{matrix} Y_1(1) - Y_1(0) > 0, \\ Y_2(1) - Y_2(0) \leq 0 \end{matrix}\right) = \begin{cases} 1 & , \widetilde{\boldsymbol{A}} = (1,0) \\ 0 & , O.W. \end{cases}$$

$$Pr\left(\widetilde{\boldsymbol{A}} \middle| \begin{matrix} Y_1(1) - Y_1(0) \leq 0, \\ Y_2(1) - Y_2(0) > 0 \end{matrix}\right) = \begin{cases} 1 & , \widetilde{\boldsymbol{A}} = (0,1) \\ 0 & , O.W. \end{cases}, \qquad Pr\left(\widetilde{\boldsymbol{A}} \middle| \begin{matrix} Y_1(1) - Y_1(0) \leq 0, \\ Y_2(1) - Y_2(0) \leq 0 \end{matrix}\right) = \begin{cases} 1 & , \widetilde{\boldsymbol{A}} = (0,0) \\ 0 & , O.W. \end{cases}$$

Perfect doctor (treatment for high blood pressure)

Science Table

Unit	Treatment (A=1)	Control (A=0)	Causal effect
Y_1	$Y_1(1) = 145$	$Y_1(0) = 150$	Improvement
Y_2	$Y_2(1) = 146$	$Y_2(0) = 145$	None
Y_3	$Y_3(1) = 145$	$Y_3(0) = 140$	None
Y_4	$Y_4(1) = 144$	$Y_4(0) = 140$	None
Y_5	$Y_5(1) = 145$	$Y_5(0) = 145$	None
Y_6	$Y_6(1) = 145$	$Y_6(0) = 160$	Improvement

ATE estimate = $145 - 146.7 < 0$

Observed outcomes (perfect doctor)

Unit	Treatment (A=1)	Control (A=0)	Treatment
Y_1	$Y_1(1) = 145$	$Y_1(0) = ?$	A=1
Y_2	$Y_2(1) = ?$	$Y_2(0) = 145$	A=0
Y_3	$Y_3(1) = ?$	$Y_3(0) = 140$	A=0
Y_4	$Y_4(1) = ?$	$Y_4(0) = 140$	A=0
Y_5	$Y_5(1) = ?$	$Y_5(0) = 145$	A=0
Y_6	$Y_6(1) = 145$	$Y_6(0) = ?$	A=1

Difference-in-means estimate = $145 - 142.5 > 0$

Perfect doctor (treatment for high blood pressure)

Science Table

Unit	Treatment (A=1)	Control (A=0)	Causal effect
Y_1	$Y_1(1) = 145$	$Y_1(0) = 150$	Improvement
Y_2	$Y_2(1) = 146$	$Y_2(0) = 145$	None
Y_3	$Y_3(1) = 145$	$Y_3(0) = 140$	None
Y_4	$Y_4(1) = 144$	$Y_4(0) = 140$	None
Y_5	$Y_5(1) = 145$	$Y_5(0) = 145$	None
Y_6	$Y_6(1) = 145$	$Y_6(0) = 160$	Improvement

ATE estimate = $145 - 146.7 < 0$

Observed outcomes (clueless doctor)

Unit	Treatment (A=1)	Control (A=0)	Treatment
Y_1	$Y_1(1) = 145$	$Y_1(0) = ?$	A=1
Y_2	$Y_2(1) = 146$	$Y_2(0) = ?$	A=1
Y_3	$Y_3(1) = 145$	$Y_3(0) = ?$	A=1
Y_4	$Y_4(1) = ?$	$Y_4(0) = 140$	A=0
Y_5	$Y_5(1) = ?$	$Y_5(0) = 145$	A=0
Y_6	$Y_6(1) = ?$	$Y_6(0) = 160$	A=0

Difference-in-means estimate = $145.3 - 148.3 < 0$

- In causal inference, the treatment assignment mechanism is essential for identifying causal effects.

Unconfounded assignment

Definition (Unconfounded Assignment Mechanism) (Imbens and Rubin, 2015)
An assignment mechanism is unconfounded if it does not depend on the potential outcomes:

$$P\left(\widetilde{\boldsymbol{A}}\middle|\widetilde{\boldsymbol{C}},\widetilde{\boldsymbol{Y}}(0),\widetilde{\boldsymbol{Y}}(1)\right)=P\left(\widetilde{\boldsymbol{A}}\middle|\widetilde{\boldsymbol{C}}\right)$$

for all $\widetilde{\boldsymbol{A}}$, $\widetilde{\boldsymbol{C}}$, $\widetilde{\boldsymbol{Y}}(0)$, *and* $\widetilde{\boldsymbol{Y}}(1)$.

- The treatment assignment mechanism in Example 1 is unconfounded, whereas the one in Example 2 is not.

- Commonly represented using conditional independence:

$$\{\widetilde{\boldsymbol{Y}}(0),\widetilde{\boldsymbol{Y}}(1)\}\perp\widetilde{\boldsymbol{A}}|\widetilde{\boldsymbol{C}}$$

Identification assumption: ignorability or exchangeability

Assumption (ignorability or exchangeability)

$$Y_i(a) \perp A_i | C_i \text{ for } a = 0, 1$$

Assumption (strong ignorability or full exchangeability)

$$\{Y_i(1), Y_i(0)\} \perp A_i | C_i$$

- The term **ignorability** (Rubin, 1978) in causal inference signifies that, when estimating causal effects, we can "ignore" the process by which units are assigned to treatments.
- **Exchangeability** means that, on average, swapping the treatment and control groups would not change the observed outcomes, ensuring their comparability.
- Show that $A_1 \perp B|C$ and $A_2 \perp B|C$ not imply $\{A_1, A_2\} \perp B|C$.

Identification assumption: ignorability or exchangeability

- The exchangeability assumption, also known as the unconfoundedness assumption, ensures that no unmeasured confounders influence both the treatment and the outcome.

- Outcome data-generating process:

$$Y(1) = g_1(C, \varepsilon_1); Y(0) = g_0(C, \varepsilon_0)$$

$$A = I(g_A(C, \varepsilon_A) \geq 0)$$

- $g_1(\cdot)$, $g_0(\cdot)$, and $g_A(\cdot)$ are general functions
- ε_1, ε_0, and ε_A are random error terms satisfying $\{\varepsilon_1, \varepsilon_0\} \perp \varepsilon_A$

This data-generating process guarantees the exchangeability and full exchangeability hold, i.e., $\{Y(1), Y(0)\} \perp A|C$

Identification assumption: ignorability or exchangeability

- Suppose there exists an unmeasured "common cause" U.

- Outcome data-generating process changes to

$$Y(1) = g_1(C, U, \varepsilon_1); Y(0) = g_0(C, U, \varepsilon_0)$$

$$A = I(g_A(C, U, \varepsilon_A) \geq 0)$$

→ The exchangeability and full exchangeability $\{Y(1), Y(0)\} \perp A|C$ do not hold in general.

- This approach is known as the *Non-Parametric Structural Equation Model (NPSEM).*

Identification assumption: SUTVA (1)

No interference assumption: Unit *i*'s potential outcomes do not depend on other units' treatments. This is sometimes called the no-interference assumption.

➢ Common scenarios of violating no interference assumption

- Spillover Effects: When treatment affects untreated individuals (e.g., herd immunity in vaccination studies).
- Peer/Network Effects: Influence within social or professional networks (e.g., students sharing knowledge).
- Clustered Treatment Assignment: Group-level treatment leads to within-group interference (e.g., community-wide policies).
- Market or Environmental Effects: Indirect effects on untreated units due to system-wide changes (e.g., wage policy shifts).

Identification assumption: SUTVA (2)

Consistency assumption: There are no other versions of the treatment. Equivalently, we require that the treatment level be well defined or have no ambiguity at least for the outcome of interest.

- Common scenarios of violating no interference assumption.
  - Treatment Variability: Different ways of delivering the same treatment produce different effects (e.g., drug formulations).
  - Misclassification of Treatment: The assigned treatment does not match the received treatment (e.g., non-adherence in clinical trials).
  - Undefined Treatment Condition: Ambiguity in defining treatment levels (e.g., "exposure to pollution" with no clear threshold).

Identification assumption: SUTVA (3)

Stable Unit Treatment Value Assumption, SUTVA: Both no interference assumption and consistency assumption hold.

- SUTVA is essential for defining well-behaved potential outcomes and ensuring that causal effects can be estimated without ambiguity.
  - Ensures that causal effects are well-defined and comparable.
  - Prevents bias in causal inference by avoiding spillover effects or treatment inconsistencies.
  - Allows for valid interpretation of estimated treatment effects.

- Mathematical Representation

No interference assumption: $Y_i(a_1, a_2, \dots a_i, \dots a_n) = Y_i(a_i)$

Consistency assumption: $Y_i(a) = Y_i$, if $A_i = a$

Identification assumption: positive/overlap (1)

Positive assumption, also known as the overlap assumption, ensures that every unit has a nonzero probability of receiving either treatment or control.

- Formally, for all covariate values C,

$$0 < \Pr(A = 1|C) < 1$$

This means that there is sufficient overlap in the distributions of treated and untreated groups.

- Why is positivity important?
  - Ensures that comparisons between treatment and control groups are valid.
  - Prevents extreme extrapolation beyond observed data.
  - Required for methods like inverse probability weighting (IPW) and propensity score matching.

Identification assumption: positive/overlap (2)

Positive assumption, also known as the overlap assumption, ensures that every unit has a nonzero probability of receiving either treatment or control.

- Formally, for all covariate values C,

$$0 < \Pr(A = 1|C) < 1$$

This means that there is sufficient overlap in the distributions of treated and untreated groups.

- When is positivity violated?
  - Some groups always receive treatment or never receive treatment (e.g., a policy only applied to a specific population).
  - Perfect predictors of treatment assignment exist (e.g., age > 65 always leads to treatment).
  - Sparse data regions where certain covariate values only appear in one group.

Identification

- Identification is the process of linking causal parameters to statistical parameters derived from observed data.

$$\mathbb{E}(Y(a)) = \int \mathbb{E}(Y(a)|C = c)\Pr(C = c)dc$$

(Law of iterated expectations)

$$= \int \mathbb{E}(Y(a)|C = c, A = a)\Pr(C = c)dc$$

(Exchangeability assumption)

$$= \int \mathbb{E}(Y|C = c, A = a)\Pr(C = c)dc$$

(Consistency assumption)

Identification, causal effects

- Causal effect on the difference scale (ATE):

$$\mathbb{E}\big(Y(1)\big) - \mathbb{E}\big(Y(0)\big) = \int \mathbb{E}(Y|C=c, A=1)\Pr(C=c)dc - \int \mathbb{E}(Y|C=c, A=0)\Pr(C=c)dc$$

- Causal effect on the ratio scale:

$$\frac{\mathbb{E}\big(Y(1)\big)}{\mathbb{E}\big(Y(0)\big)} = \frac{\int \mathbb{E}(Y|C=c, A=1)\Pr(C=c)dc}{\int \mathbb{E}(Y|C=c, A=0)\Pr(C=c)dc}$$

- Causal effect on the odds ratio scale:

$$\frac{P(Y(1)=1)[1-P(Y(0)=1)]}{P(Y(0)=1)[1-P(Y(1)=1)]}$$

$$= \frac{\int \Pr(Y=1|A=1, C=c)\Pr(C=c)\,dc \times [1 - \int \Pr(Y=1|A=0, C=c)\Pr(C=c)\,dc]}{\int \Pr(Y=1|A=0, C=c)\Pr(C=c)\,dc \times [1 - \int \Pr(Y=1|A=1, C=c)\Pr(C=c)\,dc]}$$

Link to standard statistical models

Linear model: $Y = \alpha_0 + \alpha_A A + \alpha_C C + \varepsilon$

Logistic model: $\text{logit}\{\Pr(Y = 1)\} = \beta_0 + \beta_A A + \beta_C C$

Log-Linear model: $\log(Y) = \gamma_0 + \gamma_A A + \gamma_C C$

Do these parameters (i.e., α_A, β_A, and γ_A) represent the causal effects of interest?

Far better **an approximate answer to the *right* question**, which is often vague, than **an *exact* answer to the wrong question**, which can always be made precise.

- John Tukey -

Link to standard statistical models: Linear model

For linear model:

$$Y = \alpha_0 + \alpha_A A + \alpha_C C + \varepsilon$$

Causal effect on the difference scale (ATE):

$$\begin{aligned}\mathbb{E}\big(Y(1)\big) - \mathbb{E}\big(Y(0)\big) &= \int \mathbb{E}(Y|C=c, A=1)\Pr(C=c)dc - \int \mathbb{E}(Y|C=c, A=0)\Pr(C=c)dc \\ &= \int \{\alpha_0 + \alpha_A + \alpha_C c\}\Pr(C=c)dc - \int \{\alpha_0 + \alpha_C c\}\Pr(C=c)dc \\ &= \alpha_A\end{aligned}$$

Causal effect on the ratio scale:

$$\frac{\mathbb{E}\big(Y(1)\big)}{\mathbb{E}\big(Y(0)\big)} = \frac{\int \mathbb{E}(Y|C=c, A=1)\Pr(C=c)dc}{\int \mathbb{E}(Y|C=c, A=0)\Pr(C=c)dc} = \frac{\int \{\alpha_0 + \alpha_A + \alpha_C c\}\Pr(C=c)dc}{\int \{\alpha_0 + \alpha_C c\}\Pr(C=c)dc} = \frac{\alpha_0 + \alpha_A + \alpha_C \mathbb{E}(C)}{\alpha_0 + \alpha_C \mathbb{E}(C)}$$

- Under the given identification assumptions, α_A can be interpreted as ATE but not the causal effect on the ratio scale

Link to standard statistical models: Log-Linear model

For log-linear model:

$$\log(Y) = \gamma_0 + \gamma_A A + \gamma_C C$$

Causal effect on the difference scale (ATE):

$$\mathbb{E}\big(Y(1)\big) - \mathbb{E}\big(Y(0)\big) = \int \mathbb{E}(Y|C=c, A=1)\Pr(C=c)dc - \int \mathbb{E}(Y|C=c, A=0)\Pr(C=c)dc$$

$$= \int exp(\gamma_0 + \gamma_A + \gamma_C c)\Pr(C=c)dc - \int exp(\gamma_0 + \gamma_C c)\Pr(C=c)dc$$

Causal effect on the ratio scale:

$$\frac{\mathbb{E}\big(Y(1)\big)}{\mathbb{E}\big(Y(0)\big)} = \frac{\int \mathbb{E}(Y|C=c, A=1)\Pr(C=c)dc}{\int \mathbb{E}(Y|C=c, A=0)\Pr(C=c)dc} = \frac{\int exp\{\gamma_0 + \gamma_A + \gamma_C c\}\Pr(C=c)dc}{\int exp\{\gamma_0 + \gamma_C c\}\Pr(C=c)dc} = e^{\gamma_A}$$

- Under the given identification assumptions, e^{γ_A} represents the causal effect on the ratio scale but not on the difference scale, whereas γ_A can be interpreted as the causal effect on the log-ratio scale.

Link to standard statistical models: Logistic model

For logistic model:

$$\text{logit}\{\Pr(Y=1)\} = \beta_0 + \beta_A A + \beta_C C$$

Causal effect on the odds ratio scale:

$$\frac{\Pr(Y(1)=1)[1-\Pr(Y(0)=1)]}{\Pr(Y(0)=1)[1-\Pr(Y(1)=1)]}$$

$$= \frac{\int \Pr(Y=1|A=1,C=c)\Pr(C=c)\,dc \times [1-\int \Pr(Y=1|A=0,C=c)\Pr(C=c)\,dc]}{\int \Pr(Y=1|A=0,C=c)\Pr(C=c)\,dc \times [1-\int \Pr(Y=1|A=1,C=c)\Pr(C=c)\,dc]}$$

$$= \frac{\int expit(\beta_0+\beta_A+\beta_C c)\Pr(C=c)\,dc \times [1-\int expit(\beta_0+\beta_C c)\Pr(C=c)\,dc]}{\int expit(\beta_0+\beta_C c)\Pr(C=c)\,dc \times [1-\int expit(\beta_0+\beta_A+\beta_C c)\Pr(C=c)\,dc]}$$

where $expit(x) = e^x/(1+e^x)$.

- Therefore, β_A (or any function of β_A) cannot be used to represent the causal (log) OR.

Link to standard statistical models: Logistic model

For logistic model:

$$\text{logit}\{\Pr(Y = 1|A, C)\} = \beta_0 + \beta_A A + \beta_C C$$

- However, under the rare disease assumption (typically less than 10%), β_A can be interpreted as the approximate causal log odds ratio.

Under the rare disease assumption,

1. The causal odds ratio (OR) can be approximated by the causal risk ratio (RR):

$$\frac{\Pr(Y(1) = 1)[1 - \Pr(Y(0) = 1)]}{\Pr(Y(0) = 1)[1 - \Pr(Y(1) = 1)]} = \frac{\Pr(Y(1) = 1)}{\Pr(Y(0) = 1)}$$

2. The logistic function can be approximated by the log function:

$$\text{logit}\{\Pr(Y = 1|A, C)\} \approx \log\{\Pr(Y = 1|A, C)\}$$

Link to standard statistical models: Logistic model

For logistic model:

$$\text{logit}\{\Pr(Y = 1|A, C)\} = \beta_0 + \beta_A A + \beta_C C$$

- However, under the rare disease assumption (typically less than 10%), β_A can be interpreted as the approximate causal log odds ratio.

Causal effect on the odds ratio scale ≈

$$\frac{\Pr(Y(1) = 1)}{\Pr(Y(0) = 1)} = \frac{\int \Pr(Y = 1|A = 1, C = c) \Pr(C = c)\, dc}{\int \Pr(Y = 1|A = 0, C = c) \Pr(C = c)\, dc}$$

$$\approx \frac{\int exp(\beta_0 + \beta_A + \beta_C \mathrm{c}\,) \Pr(C = c)\, dc}{\int exp(\beta_0 + \beta_C \mathrm{c}\,) \Pr(C = c)\, dc} = e^{\beta_A}$$

References

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34-58.