Chapter 8: Selection Bias

Contents

1 8.1 The Structure of Selection Bias (pp. 103-105) 1
- 1.1 Example: Folic Acid and Cardiac Malformations . . . 2
- 1.2 Bias from Conditioning on a Collider . . . 2

2 8.2 Examples of Selection Bias (pp. 105-109) 2
- 2.1 Self-Selection and Volunteer Bias . . . 3
- 2.2 Loss to Follow-Up . . . 3
- 2.3 Healthy Worker Bias . . . 3
- 2.4 Berkson's Bias (Hospital Selection Bias) . . . 3

3 8.3 Selection Bias and Confounding (pp. 109-111) 4
- 3.1 Structural Differences . . . 4
- 3.2 Can Both Occur Simultaneously? . . . 4
- 3.3 Confounding by Selection . . . 4

4 8.4 Selection Bias and Censoring (pp. 111-113) 4
- 4.1 Types of Censoring . . . 5
- 4.2 Selection Bias Due to Censoring . . . 5

5 8.5 How to Adjust for Selection Bias (pp. 113-115) 5
- 5.1 Methods to Adjust for Selection Bias . . . 5
- 5.2 When Can Selection Bias Be Eliminated? . . . 6

6 8.6 Selection Without Bias (pp. 115-116) 6
- 6.1 Unconditional Selection . . . 6
- 6.2 Conditional Selection . . . 7

7 Summary 7

8 References 8

An association created as a result of the process by which individuals are selected into the analysis is referred to as **selection bias**. Unlike confounding, this type of bias is not due to the presence of common causes of treatment and outcome, and can arise in both randomized experiments and observational studies. Like confounding, selection bias is just a form of lack of exchangeability between the treated and the untreated. This chapter provides a definition of selection bias and reviews the methods to adjust for it.

This chapter is based on Hernán and Robins (2020, chap. 8, pp. 103-116).

1 8.1 The Structure of Selection Bias (pp. 103-105)

The term "selection bias" encompasses various biases that arise from the procedure by which individuals are selected into the analysis. Here we focus on bias that would arise even if the treatment had a null effect on the outcome, i.e., **selection bias under the null**.

The structure of selection bias can be represented using causal diagrams. Figure 8.1 depicts a dichotomous treatment A, outcome Y, and their common effect C.

1.1 Example: Folic Acid and Cardiac Malformations

Suppose we study the effect of folic acid supplements A given to pregnant women shortly after conception on the fetus's risk of developing a cardiac malformation Y (1: yes, 0: no) during the first trimester of pregnancy.

Let C (1: yes, 0: no) indicate whether the pregnancy results in a live birth. Both treatment and outcome affect C:

- Folic acid A reduces the risk of spontaneous abortion (increases probability of live birth)
- Cardiac malformations Y increase the risk of spontaneous abortion (decrease probability of live birth)

Therefore, C is a **common effect** (collider) of A and Y: $A \rightarrow C \leftarrow Y$

Key insight: Even in a perfectly randomized study with no confounding, if we restrict our analysis to live births only ($C = 1$), we induce an association between A and Y through conditioning on their common effect C.

This is the fundamental structure of selection bias: conditioning on (or restricting to) a common effect of treatment and outcome.

1.2 Bias from Conditioning on a Collider

In the full population (not conditioning on C):

- Treatment A and outcome Y are marginally independent if there's no causal effect
- No association between A and Y: $A \perp\!\!\!\perp Y$

After restricting to live births ($C = 1$):

- Treatment A and outcome Y become associated
- Association is induced by conditioning on the collider C
- This association exists even if A has no causal effect on Y

Definition 1.1 (Selection Bias)**. Selection bias** occurs when conditioning on (or restricting the analysis to) a common effect of treatment and outcome, or conditioning on a variable affected by such a common effect.

This creates a non-causal association between treatment and outcome, even under the null hypothesis of no treatment effect.

Why does this create bias?

Within the selected subset (e.g., live births):

- Treated mothers who had a live birth are more likely to have fetuses without malformations (because treatment reduces abortion risk, so treated mothers with malformed fetuses can still have live births)
- Untreated mothers who had a live birth and whose fetuses had malformations must have been "lucky" in other ways to avoid abortion

This creates a spurious negative association between treatment and malformation in live births, even if treatment has no causal effect on malformations.

2 8.2 Examples of Selection Bias (pp. 105-109)

Selection bias can arise in many settings. Here we review several common examples.

2.1 Self-Selection and Volunteer Bias

Scenario: Volunteers for a study may differ systematically from non-volunteers.

Example: A study of the effect of exercise A on depression Y recruits volunteers.

If both:

- People who exercise are more likely to volunteer
- People without depression are more likely to volunteer

Then among volunteers, exercisers may appear less depressed even without a causal effect, simply because volunteers who don't exercise are selected for being unusually non-depressed (to offset their lower volunteering tendency).

2.2 Loss to Follow-Up

Scenario: Individuals drop out of a study after treatment assignment but before outcome measurement.

Example: Sicker patients are more likely to drop out, and treatment affects sickness.

If the analysis is restricted to those who complete the study, selection bias may occur because completion status C is affected by both treatment A and outcome Y (or their common causes).

Loss to follow-up is one of the most common sources of selection bias in practice.

The bias can go in either direction depending on the pattern of dropout:

- If treatment makes people sicker and sicker people drop out more, the effect may appear less harmful than it truly is
- If treatment makes people healthier and healthier people drop out less, the effect may appear more beneficial than it truly is

2.3 Healthy Worker Bias

Scenario: Workers must be healthy enough to remain employed.

Example: Studying occupational exposures A on mortality Y among employed workers.

Employment status C is affected by both:

- Occupational exposures (some exposures make workers sick, leading to job loss)
- Health status (sick workers leave the workforce)

Restricting to currently employed workers conditions on C, inducing selection bias.

Result: Occupational cohorts often show lower mortality than the general population (the "healthy worker effect"), even for harmful exposures.

2.4 Berkson's Bias (Hospital Selection Bias)

Scenario: Hospital-based case-control studies.

Example: Studying the effect of smoking A on lung cancer Y using hospitalized patients.

If:

- Smokers with other conditions (e.g., heart disease) are more likely to be hospitalized
- Non-smokers with lung cancer are more likely to be hospitalized than healthy non-smokers

Then among hospitalized individuals, smoking and lung cancer may appear less associated than in the general population.

Berkson's bias (Berkson 1946) was one of the first recognized forms of selection bias.

Named after Joseph Berkson, who noted that hospital-based studies can produce spurious associations because hospitalization itself is influenced by multiple factors.

The bias arises because hospitalization C is a collider: various diseases and risk factors increase the probability of hospitalization.

3 8.3 Selection Bias and Confounding (pp. 109-111)

Selection bias and confounding are both forms of non-exchangeability, but they have different causal structures.

3.1 Structural Differences

Confounding:

- Due to common **causes** of treatment and outcome
- Structure: $A \leftarrow L \rightarrow Y$ (backdoor path)
- Present in the population before any selection

Selection bias:

- Due to conditioning on common **effects** of treatment and outcome
- Structure: $A \rightarrow C \leftarrow Y$ (collider)
- Created by the process of selecting individuals into the analysis

3.2 Can Both Occur Simultaneously?

Yes! A variable can be both a confounder and a source of selection bias.

Example: Consider a variable L that:

1. Is a common cause of A and Y (creates confounding)
2. Is conditioned on in the analysis (creates selection bias if L is also affected by A or Y)

Practical implication: Adjusting for a variable can simultaneously:

- Reduce confounding (if the variable is a confounder)
- Introduce selection bias (if the variable is a collider)

This is why careful causal analysis using DAGs is essential before deciding which variables to adjust for.

3.3 Confounding by Selection

In some causal diagrams, selection can induce confounding by opening backdoor paths that would otherwise be blocked.

Example: Suppose there is no confounding in the full population, but restricting the analysis to a subset creates confounding.

This occurs when:

- Selection S is a collider on a path between A and Y
- Conditioning on S opens that path, creating confounding

4 8.4 Selection Bias and Censoring (pp. 111-113)

Censoring is a specific form of selection where we fail to observe the outcome for some individuals.

4.1 Types of Censoring

Right censoring: Outcome is not observed because follow-up ends before the outcome occurs (e.g., study ends, patient drops out).

Left censoring: Outcome occurred before observation began.

Interval censoring: Outcome time is known only to occur within an interval.

4.2 Selection Bias Due to Censoring

Censoring causes selection bias if:

1. Censoring is affected by treatment, outcome, or their common causes
2. The analysis is restricted to uncensored observations

Example 4.1 (Censoring Creates Selection Bias)**.** Study the effect of AZT A on mortality Y in HIV-positive individuals.

Suppose:

- AZT reduces mortality (increases survival time)
- Sicker patients are more likely to drop out of the study
- We censor (exclude) individuals who drop out

Let C indicate whether an individual remains in the study until outcome measurement.

If C is affected by both A (through its effect on health/survival) and Y (sicker people with worse outcomes drop out more), then restricting to $C = 1$ creates selection bias.

Informative censoring: Censoring is said to be informative if it is associated with the outcome (conditional on treatment and measured covariates).

Informative censoring creates selection bias because it's equivalent to conditioning on a collider or descendant of a collider.

Methods to handle censoring:

- Inverse probability of censoring weighting (IPCW)
- Imputation methods
- Survival analysis methods (Chapter 17)

5 8.5 How to Adjust for Selection Bias (pp. 113-115)

Like confounding, selection bias can sometimes be adjusted for if appropriate data are available.

5.1 Methods to Adjust for Selection Bias

1. Inverse Probability of Selection Weighting

Create a pseudo-population where selection is independent of treatment and outcome.

For each individual in the selected sample, assign weight:

$$w^S = \frac{1}{Pr[S = 1|A, Y, L]}$$

where S indicates selection into the analysis and L are measured variables.

Assumptions required:

1. Selection probabilities $Pr[S = 1|A, Y, L]$ can be estimated (requires data on A, Y, L for at least some non-selected individuals, or strong modeling assumptions)
2. Positivity: $Pr[S = 1|A, Y, L] > 0$ for all relevant combinations

3. The variables L are sufficient to make S independent of unmeasured factors affecting the A-Y relationship

2. Standardization (Restriction and Conditioning)

If selection depends only on measured variables L:

- Estimate the association within strata of L
- Standardize to the population distribution of L

3. Stratification on Selection

If possible, collect data on both selected and non-selected individuals.

Estimate the effect separately in selected and non-selected subgroups.

If the causal effect is the same in both groups, we can identify the population average causal effect.

5.2 When Can Selection Bias Be Eliminated?

Selection bias can be eliminated if:

1. We measure all variables that determine selection and are associated with treatment and outcome
2. These variables suffice to make selection independent of the treatment-outcome relationship

Practical challenge: Often we cannot measure all determinants of selection.

For example:

- In studies with loss to follow-up, we may not know why people dropped out
- In volunteer studies, we may not know what motivated volunteers vs. non-volunteers
- This makes selection bias often more difficult to address than confounding

Prevention is better than cure: Design studies to minimize selection bias rather than trying to adjust for it statistically.

6 8.6 Selection Without Bias (pp. 115-116)

Not all selection creates bias. Selection is harmless under certain conditions.

6.1 Unconditional Selection

Selection does not create bias if the probability of selection does not depend on both treatment and outcome simultaneously.

Examples of harmless selection:

1. **Random sampling**: Each individual has the same probability of selection, independent of A and Y
2. **Selection based only on treatment**: Study only treated individuals (but selection doesn't depend on Y)
3. **Selection based only on outcome (case-control studies)**: If analyzed correctly, selection on outcome can be valid

Case-control studies: These studies select on outcome (all cases, sample of controls).

Despite this selection, case-control studies can validly estimate odds ratios under certain conditions:

- Selection of controls is independent of exposure, conditional on matching variables
- No selection bias for exposure-outcome association
- Can use logistic regression, with proper design and analysis

6.2 Conditional Selection

Selection based on variables that are not affected by treatment or outcome generally doesn't create bias.

Example: Selecting only women for a study of a treatment and outcome.

If sex is not affected by treatment or outcome, this selection doesn't bias the treatment-outcome association within women.

Generalizability vs. internal validity:

- Selection may limit **generalizability** (external validity) even when it doesn't create **selection bias** (internal validity)
- Estimating the effect only in women is internally valid but may not generalize to men
- This is different from selection bias, which creates a spurious association even within the selected population

7 Summary

This chapter examined **selection bias**, another threat to exchangeability.

Key concepts:

1. **Structure of selection bias**: Arises from conditioning on a common effect (collider) of treatment and outcome
2. **Sources of selection bias**:
    - Self-selection and volunteer bias
    - Loss to follow-up
    - Healthy worker bias
    - Berkson's bias (hospital-based studies)
3. **Selection bias vs. confounding**:
    - Confounding: common causes ($A \leftarrow L \rightarrow Y$)
    - Selection bias: common effects ($A \rightarrow C \leftarrow Y$)
    - Can occur simultaneously
4. **Censoring**: A specific type of selection where outcomes are unobserved
5. **Adjustment methods**:
    - Inverse probability of selection weighting
    - Standardization
    - Stratification (when possible)
6. **Selection without bias**: Not all selection creates bias (e.g., random sampling, case-control studies with proper analysis)

Practical implications:

1. **Study design**: Minimize selection bias through careful design
    - Maximize follow-up
    - Reduce loss to follow-up
    - Consider incentives for participation
2. **Analysis**: Use causal diagrams (DAGs) to identify potential selection bias
    - Identify colliders
    - Determine whether selection opens or blocks causal paths
    - Adjust only when appropriate
3. **Reporting**: Be transparent about:
    - Selection mechanisms
    - Loss to follow-up

- Potential for selection bias
- Methods used to address it

Looking ahead:

- **Chapter 9**: Measurement bias
- **Chapters 12-15**: Advanced methods including censoring and time-varying treatments

8 References

Hernán, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC. https://miguelhernan.org/whatifbook.