Advanced Methods in Biostatistics

STAT 438

Winter 2022 (1221)[1]

Cameron Roopnarine[2] Yeying Zhu[3]

26th January 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

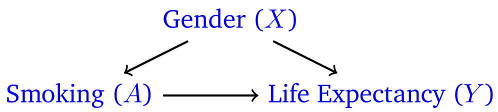
  1. – **Risk factor**: cigarette smoking.
     – **Outcome**: bladder cancer.
  2. – **Risk factor**: distance of home from hazardous waste site.
     – **Outcome**: respiratory disease.

- Three most popular observational studies:

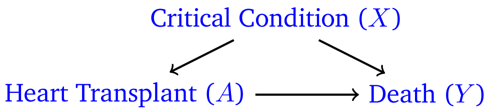
1. Cross-sectional studies.
2. Cohort studies.
3. Case-control studies.

- 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 $0 \leq \text{RR} < \infty$.

Remarks:

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

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

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

(2) If there exists absolutely no association between D and E, this results in $\text{RR} = 1$, that is, this will

happen when $\mathbb{P}(D \mid E) = \mathbb{P}(D \mid E^c)$.

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

(4) If 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,

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

- To make inference, we have, approximately,

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

where

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

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

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

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

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

For RR, we have

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

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

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

Using, Taylor's approximation, we have

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

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

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

Therefore,

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

Similarly,

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

Therefore,

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

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) $\text{ER} > 0$ means greater risk of D for E versus E^c.

(3) $\text{ER} < 0$ 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,

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

  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:

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

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

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

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,

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

- To make inference, we have approximately,

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

where

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

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.

<u>Notation</u>:

- 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

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

- $\alpha = \mathsf{log}(P_{x=0})$: 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

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

- $\alpha = \log\left(\frac{P_{x=0}}{1-P_{x=0}}\right)$: 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

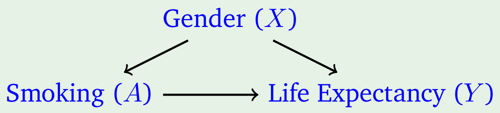
Observational Studies

- 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

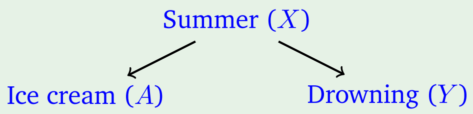
Differences in the outcome are not only due to the treatment, but also because of the masking effect of covariates (confounders).



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,

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

Week 4
24th to 28th January

2.8 Modeling of Propensity Scores

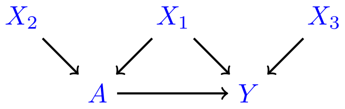
Estimation of Propensity Scores

- 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 $i = 1, \dots, n$ using data (A_i, X_i), $i = 1, \dots, n$.
2. Using $\widehat{\mathsf{ps}}(X_i)$, $i = 1, \dots, n$ 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, $S_1, \dots, S_K$

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

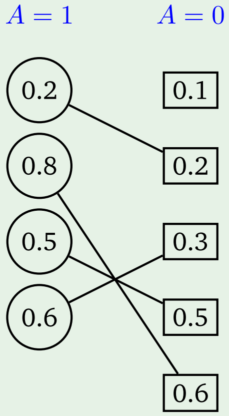
where for all $X \in S_k$, we have balance.

- $\mathbb{E}[Y^1 - Y^0 \mid X \in S_k]$ 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,

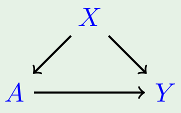
$$\begin{aligned}
\mathbb{E}\left[\frac{(1-A)Y}{1-\mathsf{ps}(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{(1-A)Y}{1-\mathsf{ps}(X)} \,\middle|\, Y^0, X\right]\right] \\
&\vdots \\
&= \mathbb{E}[Y^0].
\end{aligned}$$

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

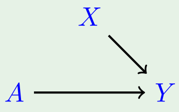
- Weighting scheme:
  - For those in the treatment group ($A_i = 1$), assign a weight of $w_i = 1/\widehat{\mathsf{ps}}(X_i)$.
  - For those in the control group ($A_i = 0$), assign a weight of $w_i = 1/(1 - \widehat{\mathsf{ps}}(X_i))$.
- By weighting, each subject is replicated w_i times. IPW creates 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 $b=1,\dots,B$ (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 $\hat{\tau}^{(b)}_{\text{IPW}_2}$, $b=1,\dots,B$.
- 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:

$$\begin{aligned}\hat{\tau}_{\text{DR}} &= \frac{1}{n}\sum_{i=1}^{n} \frac{A_i Y_i - [A_i - \widehat{\mathsf{ps}}(X_i)]\hat{m}_1(X_i)}{\widehat{\mathsf{ps}}(X_i)} \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - A_i)Y_i + [A_i - \widehat{\mathsf{ps}}(X_i)]\hat{m}_0(X_i)}{1 - \widehat{\mathsf{ps}}(X_i)} \\ &= \hat{\tau}_{1,\text{DR}} - \hat{\tau}_{0,\text{DR}},\end{aligned}$$

  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.

Example: Homelessness and Physical Health

Data Source: Kleinman, K and Horton, NJ. SAS and R: *Data Management, Statistical Analysis, and Graphics*. CRC Press.

The HELP (Health Evaluation and Linkage to Primary Care) study was a clinical trial for adult patients recruited from a detoxification unit. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care.

Secondary Analysis: Does homelessness affect physical health? http://sas-and-r.blogspot.com/2010/04/example-734-propensity-scores-and.html

Available data:

- `pcs`: measure of physical health via 36 question questionnaire (mean 48.05, range $[14.07, 74.81]$).
- `homeless`: binary indicator of homelessness (46.14% homeless).

- `age`: in years (mean 35.65, range $[19.00, 60.00]$).
- `female`: indicator of female sex (23.62% female).
- `i1`: alcohol intake per day (mean 17.9, median 13.0, range $[0, 142]$).
- `mcs`: mental component summary measure (mean 31.68, range $[6.76, 62.18]$).
- Linear model for the association between homelessness and physical health:

```
ds <- read.csv("help.csv")
attach(ds)
summary(lm(pcs ~ homeless))


Call:
lm(formula = pcs ~ homeless)

Residuals:
    Min      1Q  Median      3Q     Max 
-34.927  -7.903   0.644   8.387  25.805 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   49.001      0.688  71.220   <2e-16 ***
homeless      -2.064      1.013  -2.038   0.0422 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.75 on 451 degrees of freedom
Multiple R-squared:  0.009123,Adjusted R-squared:  0.006926 
F-statistic: 4.152 on 1 and 451 DF,  p-value: 0.04216
```

Unadjusted analysis: significant association between homelessness and physical health.

3.5.1 Estimation

- Fit a logistic model for the propensity score

```
glm1 <- glm(homeless ~ age + factor(female) + i1 + mcs, family = binomial)
PS <- glm1$fitted.values
summary(glm1)


Call:
glm(formula = homeless ~ age + factor(female) + i1 + mcs, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7603  -1.0271  -0.8211   1.2039   1.6332  

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.658259   0.518953  -1.268   0.2046
```

```
age              0.013659   0.012965   1.054   0.2921    
factor(female)1 -0.454530   0.237011  -1.918   0.0551 .  
i1               0.024878   0.005782   4.303 1.68e-05 ***
mcs             -0.009983   0.007768  -1.285   0.1987    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 625.28  on 452  degrees of freedom
Residual deviance: 592.54  on 448  degrees of freedom
AIC: 602.54

Number of Fisher Scoring iterations: 4
```

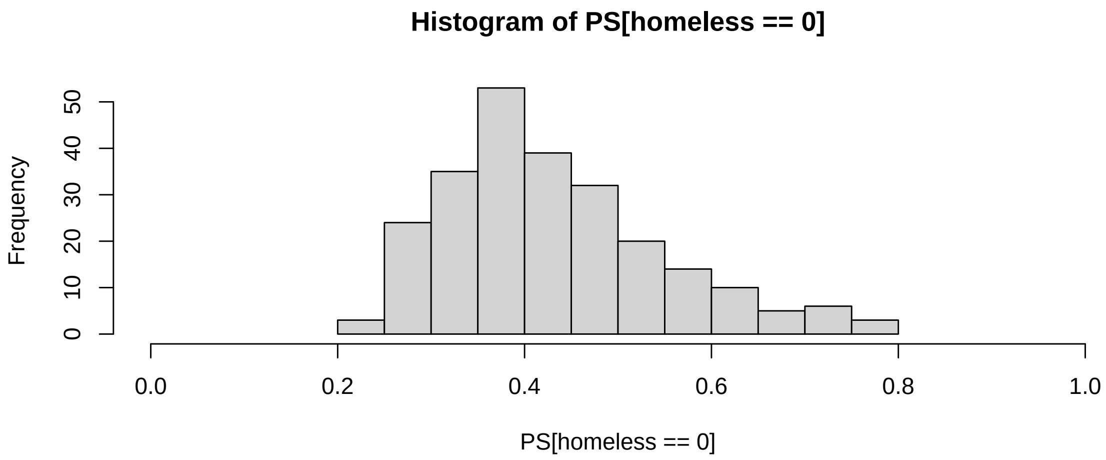
- Which variables should be included in the PS model? Real confounders and variables that are predictive of the outcome.

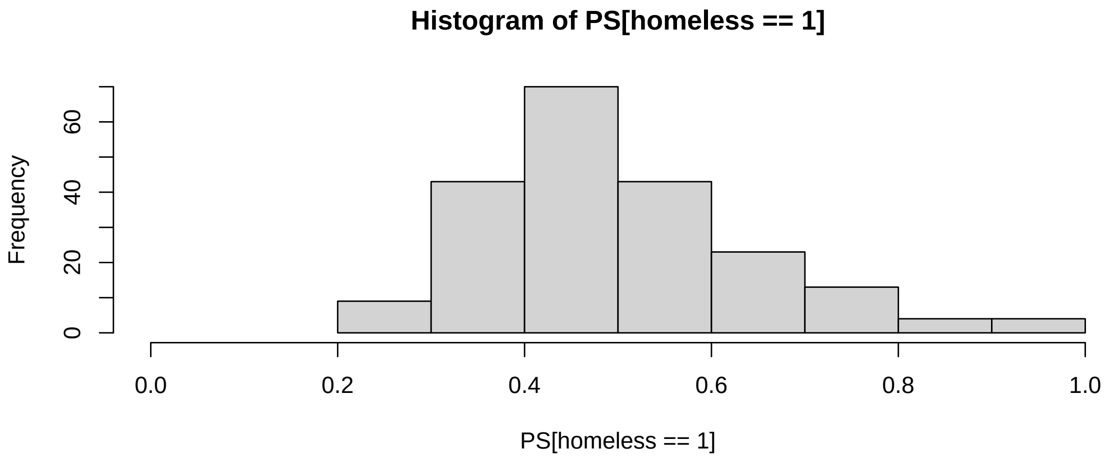
```
summary(lm(pcs ~ age))$coef

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 59.4549837 2.33874676 25.421728 4.125943e-89
age         -0.3199256 0.06411778 -4.989655 8.651612e-07

summary(lm(pcs ~ female))$coef

             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 48.986239  0.5732719 85.450274 1.076730e-280
female      -3.969879  1.1795541 -3.365576  8.291456e-04

summary(lm(pcs ~ i1))$coef

              Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 49.9424609 0.66766179 74.802036 2.407295e-256
i1          -0.1057625 0.02487199 -4.252273  2.573816e-05

summary(lm(pcs ~ mcs))$coef

               Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 45.10957488 1.34341383 33.578317 9.114672e-125
mcs          0.09278014 0.03931041  2.360192  1.869034e-02
```

- Check that there is a reasonable amount of overlap in the propensity scores between the two groups:

```
summary(PS[homeless == 0])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2137  0.3474  0.4040  0.4297  0.4980  0.7876 

summary(PS[homeless == 1])

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2635  0.4003  0.4739  0.4984  0.5768  0.9643 
```

- Mean propensity to homelessness is larger in the homeless group.
- There are no non-homeless subjects with a propensity score > 0.8.





3.5.2 Matching

In R, the `Matching` library provides tools for matching and analysis.

```
library(Matching)
rr <- Match(Y = pcs, Tr = homeless, X = PS, M = 1, replace = TRUE,
  estimand = "ATE")
summary(rr)


Estimate...  -1.2774
AI SE......  1.2311
T-stat.....  -1.0376
p.val......  0.29945

Original number of observations..............  453
Original number of treated obs...............  209
```

```
Matched number of observations...............  453
Matched number of observations  (unweighted).  536

# average causal effect estimate
rr$est

          [,1]
[1,] -1.277365

# standard error of ACE estimate
rr$se

[1] 1.231057

# p-value for testing ACE=0
pnorm(rr$est/rr$se) * 2

          [,1]
[1,] 0.2994486
```

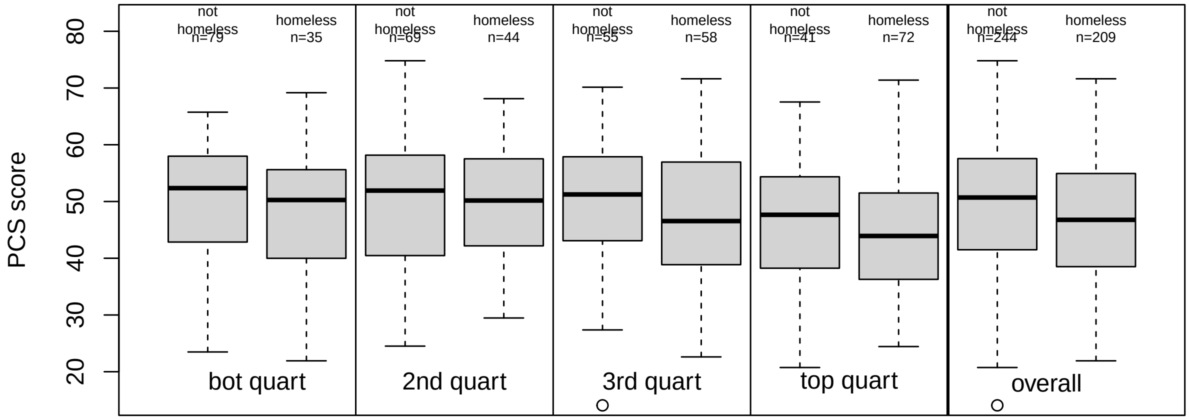
- The causal estimate of -1.28 in the matched comparison is not statistically significant ($p = 0.30$).
- Note that the specific results depend on the particular options that are selected for the matching.

3.5.3 Stratification

We will stratify the dataset into four ($K = 4$) approximately equally sized groups.

```
breakvals <- fivenum(PS)
strata <- cut(PS, breaks = breakvals, labels = c("bot quart",
  "2nd quart", "3rd quart", "top quart"), include.lowest = TRUE)
strata[1:5]

[1] 3rd quart top quart 2nd quart bot quart 3rd quart
Levels: bot quart 2nd quart 3rd quart top quart

table(strata)

strata
bot quart 2nd quart 3rd quart top quart
      114       113       113       113
```

Boxplots of the PCS scores for homeless and non-homeless by the four strata of propensity scores (http://sas-and-r.blogspot.com/search?q=7.36):



- The difference between the median PCS scores is smaller within each of the quartiles of the propensity scores than the difference between medians overall.
- Proportion of subjects in the two groups varies by strata.
- Use a t-test in each stratum to test whether difference in mean PCS is significant between homeless and non-homeless subjects.
- Mean pcs is lower in 3 out of 4 strata but not statistically significant.
- Little credible evidence for a health cost ascribable to homelessness.

```
stratdf <- data.frame(pcs, homeless, strata)
out <- by(stratdf, strata, function(mydataframe) {
  with(mydataframe, t.test(pcs[homeless == 0], pcs[homeless ==
    1]))
})
# bot quart: t = 0.80603, df = 58.564, p-value = 0.4235

# 2nd quart: t = -0.10106, df = 101.08, p-value = 0.9197

# 3rd quart: t = 0.82302, df = 110.92, p-value = 0.4123

# top quart: t = 0.92219, df = 74.798, p-value = 0.3594

lm(pcs ~ homeless, data = ds[strata == "bot quart", ])$coef

(Intercept)    homeless
   49.88714    -1.79067

lm(pcs ~ homeless, data = ds[strata == "2nd quart", ])$coef

(Intercept)    homeless
 49.3708866   0.2041892
```

```
lm(pcs ~ homeless, data = ds[strata == "3rd quart", ])$coef

(Intercept)    homeless
   49.44396    -1.67437

lm(pcs ~ homeless, data = ds[strata == "top quart", ])$coef

(Intercept)    homeless
  46.075841   -1.985963
```

3.5.4 Inverse Probability Weighting

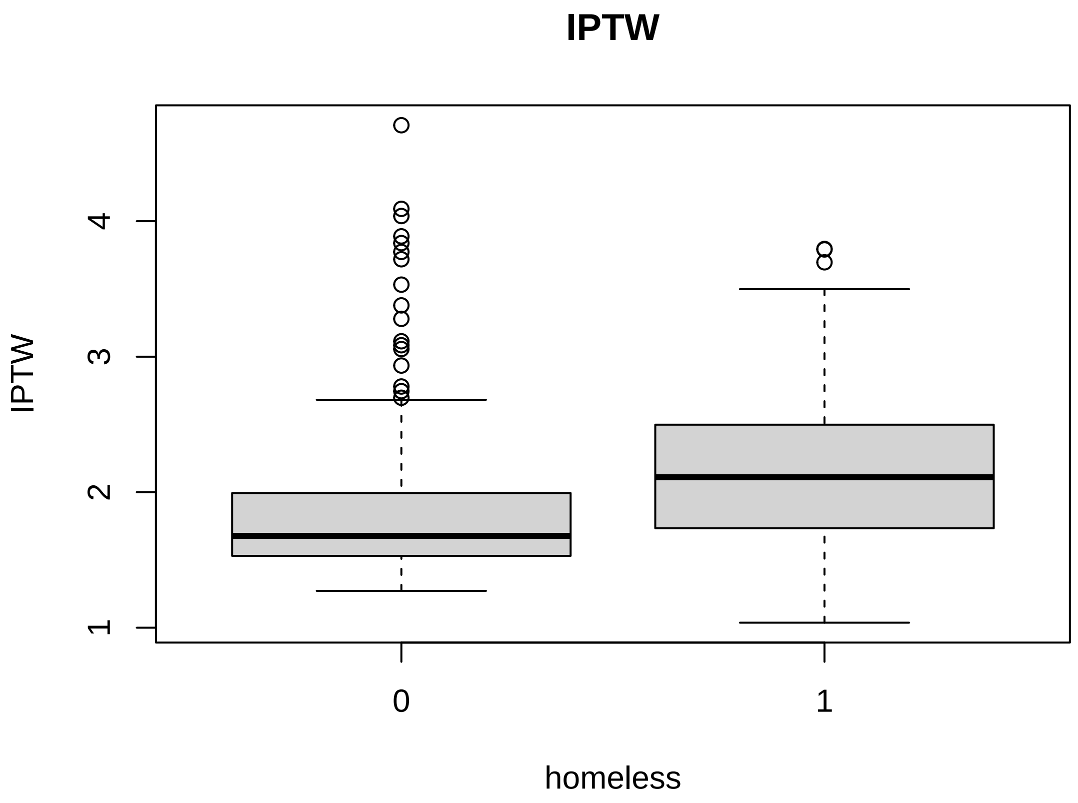
```
ACE.1 <- 1/length(PS) * (sum(homeless * pcs/PS) - sum((1 - homeless) *
  pcs/(1 - PS)))
ACE.1

[1] -1.652115

ACE.2 <- sum(homeless * pcs/PS)/sum(homeless/PS) - sum((1 - homeless) *
  pcs/(1 - PS))/sum((1 - homeless)/(1 - PS))
ACE.2

[1] -1.302807
```

Obtain Standard Error Using Bootstrap

```
ACE.1 <- rep(NA, 10000)
ACE.2 <- rep(NA, 10000)
for (i in 1:10000) {
  select <- sample(1:length(pcs), size = length(pcs), replace = TRUE)
  ndata <- ds[select, ]
  ndata$ps <- glm(homeless ~ age + factor(female) + i1 + mcs,
    data = ndata, family = binomial)$fitted.values
  ACE.1[i] <- 1/length(ndata$PS) * (sum(ndata$homeless * ndata$pcs/ndata$PS) -
    sum((1 - ndata$homeless) * ndata$pcs/(1 - ndata$PS)))
  ACE.2[i] <- sum(ndata$homeless * ndata$pcs/ndata$PS)/sum(ndata$homeless/ndata$PS) -
    sum((1 - ndata$homeless) * ndata$pcs/(1 - ndata$PS))/sum((1 -
      ndata$homeless)/(1 - ndata$PS))
}
# sd(ACE.1) = 4.776752 sd(ACE.2) = 1.067576
```

Equivalently: Calculating The Inverse Weights

```
IPTW <- rep(0, length(PS))
IPTW[homeless == 1] <- 1/PS[homeless == 1]
IPTW[homeless == 0] <- 1/(1 - PS[homeless == 0])
boxplot(IPTW ~ homeless, main = "IPTW")
```



Equivalently: Fit a Weighted Regression

```
library(survey)
design.ps <- svydesign(ids = ~1, weights = ~IPTW, data = ds)
msm1 <- svyglm(pcs ~ homeless, design = design.ps)
summary(msm1)$coef

             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 48.623563  0.7519392 64.664225 2.971552e-230
homeless    -1.302807  1.0744770 -1.212504  2.259544e-01
```

Note: the standard error from the survey package provides the robust standard error and correct p-value.

3.5.5 Double-Robust Estimation

1. Fit a logistic model for the propensity score.

```
glm1 <- glm(homeless ~ age + factor(female) + i1 + mcs, family = binomial)
PS <- glm1$fitted.values
summary(glm1)$coef

                    Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)     -0.658258820 0.518953077 -1.268436 2.046423e-01
age              0.013659005 0.012965104  1.053521 2.921024e-01
```

```
factor(female)1 -0.454530491 0.237010899 -1.917762 5.514120e-02
i1               0.024878024 0.005781542  4.303009 1.684943e-05
mcs             -0.009983455 0.007768247 -1.285162 1.987357e-01
```

2. Fit regression model for the outcome using data from the treatment group only, predict for all subjects.

```
out1 <- lm(pcs ~ age + factor(female) + i1 + mcs, subset = (homeless ==
  1))
m1 <- predict(out1, ds)
summary(out1)$coef

                   Estimate Std. Error   t value     Pr(>|t|)
(Intercept)     55.79947502 3.44908558 16.178049 2.007994e-38
age             -0.27753512 0.08846710 -3.137156 1.958233e-03
factor(female)1 -4.01357667 1.78089142 -2.253690 2.527913e-02
i1              -0.06712578 0.03089995 -2.172359 3.098149e-02
mcs              0.11537037 0.05955119  1.937331 5.408570e-02
```

3. Fit a regression model for the outcome (same form as above) using data from the control group only, predict for all.

```
out0 <- lm(pcs ~ age + factor(female) + i1 + mcs, subset = (homeless ==
  0))
m0 <- predict(out0, ds)
summary(out0)$coef

                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)     59.6411734 3.79405564 15.7196359 1.027021e-38
age             -0.2676526 0.09564903 -2.7982780 5.556571e-03
factor(female)1 -4.1990531 1.56235123 -2.6876499 7.701442e-03
i1              -0.1034846 0.04562826 -2.2679935 2.422283e-02
mcs              0.0397022 0.05052110  0.7858537 4.327316e-01
```

4. Plug in predicted values into the expression for $\hat{\tau}_{\text{DR}}$.

```
DR.est <- mean((homeless * pcs - (homeless - PS) * m1)/PS) -
  mean(((1 - homeless) * pcs + (homeless - PS) * m0)/(1 - PS))
DR.est

[1] -1.211244
```

5. Use the same bootstrap technique to obtain standard error as in IPW.

Summary

Recall estimates of the ACE from other approaches:

- Naive unadjusted: -2.06 (1.013).
- Matching: -1.28 (1.231).
- Stratification: $(-1.79, 0.20, -1.67, -1.99)$ [average difference -1.31].
- IPW: -1.30 (1.075).
- Double-Robust: -1.21 (1.02).

Checking Balance

How do we know if a propensity score based method did a good job or not?

- Recall $X \perp\!\!\!\perp \big(A \mid \mathsf{ps}(X)\big)$. If the propensity score model is correct, it should balance the covariates after matching/stratification/IPW.
- Check balance. In IPW,

$$w_i = \begin{cases} \dfrac{1}{\widehat{\mathsf{ps}}(X_i)}, & \text{if } A_i = 1, \\ \dfrac{1}{1-\widehat{\mathsf{ps}}(X_i)}, & \text{if } A_i = 0. \end{cases}$$

Absolute Standardized Mean Difference (ASMD)

$$\mathsf{ASMD}^{(j)} = \frac{|\bar{X}^w_{j,1} - \bar{X}^w_{j,0}|}{\mathsf{sd}(X_{j,1})},\ j = 1, \ldots, p,$$

$$\text{ASMD mean} = \sum_{j=1}^{p} \frac{\mathsf{ASMD}^{(j)}}{p},$$

where

- $\bar{X}^w_{j,1} = \frac{\sum w_i A_i X_{ij}}{\sum w_i A_i}$ is the weighted mean of X_j in the treatment group;
- $\bar{X}^w_{j,0} = \frac{\sum w_i (1-A_i) X_{ij}}{\sum w_i (1-A_i)}$ is the weighted mean of X_j in the control group;
- $\mathsf{sd}(X_{j,1})$ is the unweighted standard deviation of X_j in the treatment group.

Instead, we can also look at

$$\text{ASMD max} = \max_j \mathsf{ASMD}^{(j)} \quad \text{or} \quad \text{ASMD mean} = \operatorname*{median}_j \mathsf{ASMD}^{(j)}.$$

Kolmogorov-Smirnov (KS) Statistic

$$\mathsf{KS}^{(j)} = \sup\big|F^w_{1,n_1}(X_{j,1}) - F^w_{0,n_0}(X_{j,0})\big|,$$

$$\text{KS mean} = \sum_{j=1}^{p} \frac{\mathsf{KS}^{(j)}}{p},$$

where

- $F^w_{1,n_1}(X_{j,1})$ is the weighted empirical cdf of X_j in the treatment group;
- $F^w_{0,n_0}(X_{j,0})$ is the weighted empirical cdf of X_j in the control group.

- Usually, if ASMD < 0.2 (or 0.1), we say that the covariates are balanced across treatment groups and conclude the propensity score based method did a good job.
- The `std.diff` function posted on LEARN computes ASMD for each covariate. For categorical variables with K levels, K indicator variables will be generated for each level, and ASMD values will be calculated for each level.
- `std.diff(u,z,w)`:
  - `u`: the covariate;

  - z: treatment indicator;
  - w: weights to be applied.

- Checking balance with the IPW approach:

```
std.diff(age, homeless, IPTW)

[1] 0.09585512

std.diff(factor(female), homeless, IPTW)

[1] 0.09554836 0.09554836

std.diff(i1, homeless, IPTW)

[1] 0.2343231

std.diff(mcs, homeless, IPTW)

[1] 0.07989432

# An equivalent way:
x <- data.frame(age, factor(female), i1, mcs)
sapply(x, std.diff, homeless, IPTW)

$age
[1] 0.09585512

$factor.female.
[1] 0.09554836 0.09554836

$i1
[1] 0.2343231

$mcs
[1] 0.07989432
```

- Checking balance with the matching approach:

```
match <- c(rr$index.treated, rr$index.control)
x.matched <- x[match, ]
homeless.matched <- homeless[match]
w <- rep(1, length(match))
sapply(x.matched, std.diff, homeless.matched, w)

$age
[1] 0.03229325

$factor.female.
[1] 0.1205278 0.1205278

$i1
```

```
[1] 0.05847934

$mcs
[1] 0.05000506
```

Summary

Steps in a propensity score analysis:

1. Estimate the propensity score using data (A_i, X_i), $i = 1, \ldots, n$.
2. Select a method for propensity score adjustment:
   - Matching.
   - Stratification.
   - Inverse Probability Weighting (IPW).
   - Double-Robust Estimation.
3. Assess the balance between the treated and control groups:
   - If the covariates are balanced, go to the final step.
   - If the covariates are not balanced, the fitted propensity score model is not good enough; refine the model.
4. Fit an outcome model between the treatment and the outcome variable using the estimated propensity scores.