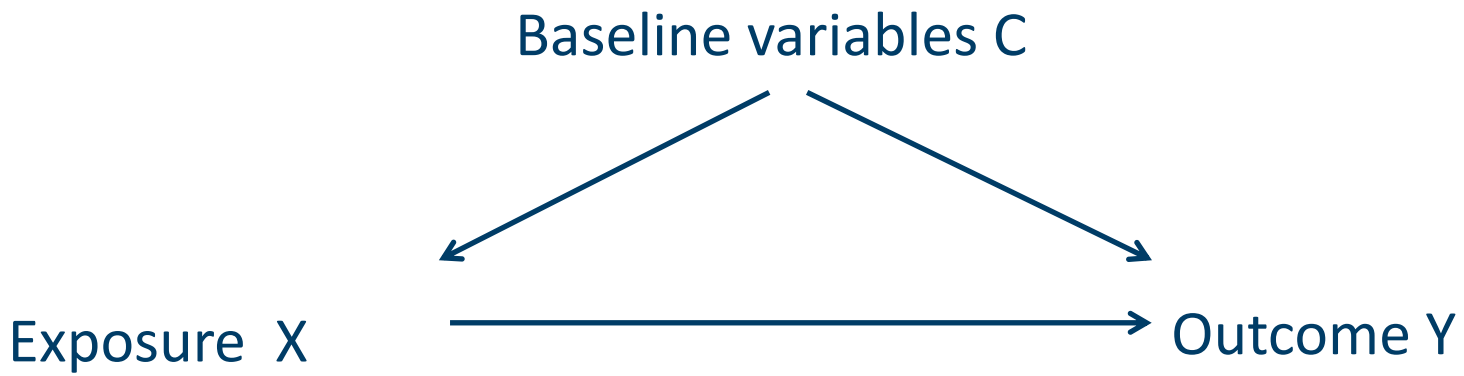
Leids Universitair Medisch Centrum

Causal inference I

Week 4: adjusting for confounding: Outcome regression and G-computation

Confounding in observational studies

Treatment groups are not comparable at baseline (start of study)



- Confounding may introduce spurious associations between exposure and observed outcome.
- Backdoor path in DAG
- No exchangeability: *Y(1), Y(0)* $\not\perp\!\!\!\perp$ *X*
  - X is not independent of $Y(x)$, for all x

Challenge: how to handle confounding

How to estimate causal effects from observational data?

1. Try to identify all confounders and adjust for them

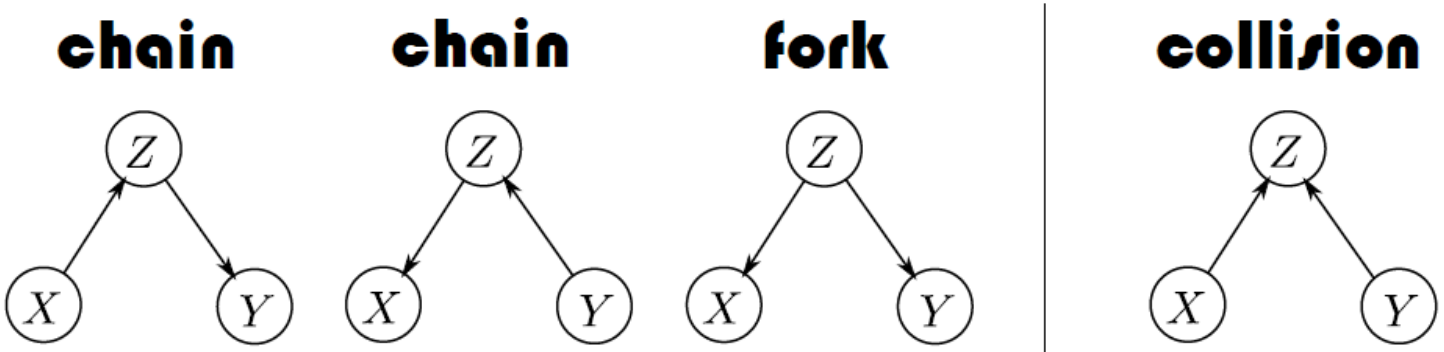
- Stratification
- Traditional regression methods.
- G-computation (Outcome regression + standardization)
- Propensity score methods (inverse probability weighing, matching, stratification)
- Double robust methods

Today + next week

2. Pseudo-randomization/ natural experiments

- Instrumental variable analysis
- Regression discontinuity design

Last week: DAGs



- Graph in which causal assumptions on the relation between variables are visualized
- Use the “backdoor adjustment rules” to determine the set of variables {C} to adjust for in the analysis.

Today: we assume that the set of adjustment variables has been determined.

How to estimate causal effects?

Extending the causal assumptions to account for the adjustment set C

We assume:

1. Consistency
2. Conditional exchangeability

$$Y(1), Y(0) \perp\!\!\!\perp X \mid C$$

- Within levels of C, the groups are exchangeable
- $E(Y(x)|C) = E(Y(x)|C, X = 1) = E(Y(x)|C, X = 0)$
- No unmeasured confounding

3. Positivity

$0 < P(X = x|C = c) < 1$, for all values of x and c

- within each level and combination of the variables used to achieve exchangeability, there are exposed and unexposed subjects:

PART 1 Traditional methods : stratification and regression

part 4.2.3 and 4.2.4 R-causal.org

Example

NECOSAD study

- Observational cohort study
- Patients with kidney failure who start dialysis
- Exposure: dialysis via hemodialysis versus peritoneal
- Outcome: physical functioning after 6 months, measured with SF-36 scale (range 0-100)

Compare mean outcome between the groups

Exposure	n	Mean outcome	SD
hemodialysis (X=0)	675	50.6	29.7
Peritoneal dialysis (X=1)	325	62.0	25.0

Observed difference: 11.4 points (95% CI 7.9- 15.0)

A large difference

Compare groups at baseline

	hemodialysis (N=675)	peritoneal dialysis (N=325)	Overall (N=1000)
age start therapy			
Mean (SD)	61.1 (14.5)	53.3 (14.0)	58.6 (14.8)
sf 36 physical functioning PF at baseline			
Mean (SD)	47.7 (28.7)	60.0 (25.2)	51.7 (28.2)

Patients on hemodialysis are older and have lower physical functioning at baseline

No exchangeability

Age and physical functioning at baseline are confounders

Stratification

If there are only few levels of C

- Estimate the causal contrast separately within each confounder level and average the contrasts

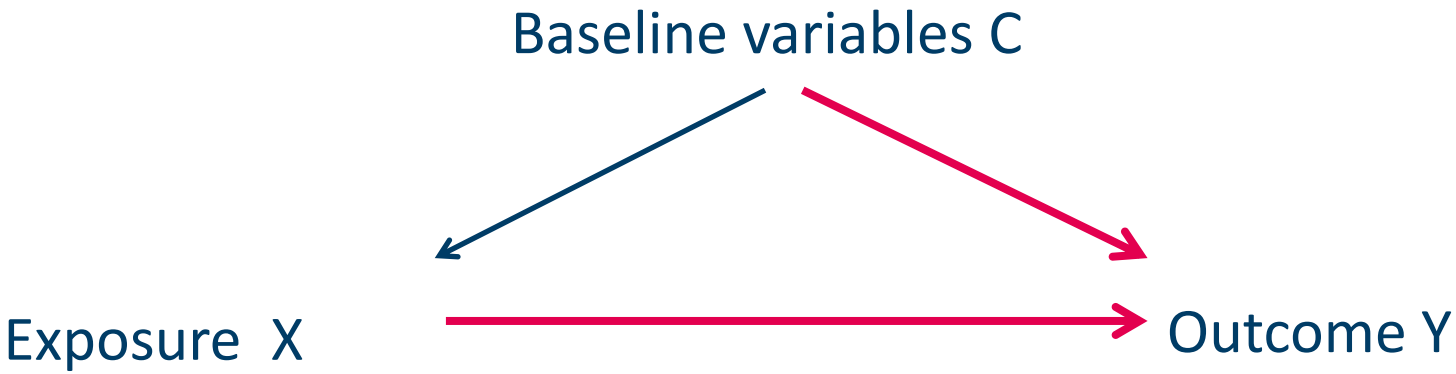
- If we could stratify the population by all relevant confounders C, individuals within each stratum of C would be as good as randomized, i.e. conditionally exchangeable.

See week 2 exercise 3 and week 3, the Simpson paradox

However

- In our example: many levels of C (two continuous confounders)

Controlling for confounding with outcome modelling



Model the relation between the outcome *Y*, exposure/treatment *X* and confounders *C*

Linear regression (outcome is continuous)

Unadjusted model

```
> lin.model1 <-lm(sf_phys6~exposure, data=kidney)
> summary(lin.model1)
 Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   50.586      1.089  46.435  < 2e-16 ***
Exposure                      11.418      1.911   5.975  3.2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The expected physical functioning after 6 months is 11.4 points higher for exposure = 1(peritoneal dialysis) compared to exposure = 0 (hemodialysis) .

Very small p-value, Highly statistically significant

Linear regression with adjustment for {C}

Adjusted model

```
> lin.model2 <-lm(sf_phys6~exposure+age+sf_phys0, data=kidney)
> summary(lin.model2)


Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             40.48968    3.37938  11.981  < 2e-16 ***
Exposure                 0.57333    1.38438   0.414    0.679
age                     -0.35029    0.04543  -7.711 3.02e-14 ***
sf_phys0                 0.66071    0.02362  27.972  < 2e-16 ***
---
```

The expected physical functioning after 6 months is 0.57 points higher for exposure = 1(peritoneal dialysis) compared to exposure = 0 (hemodialysis) after adjustment for age and physical functioning at baseline.

A small difference. Not statistically significant

What do we observe?

Without adjustment: a strong effect of exposure on the outcome (Peritoneal seems to perform better

After adjustment for confounders: a small, not significant effect of exposure on the outcome. Both dialysis forms seems to perform similar

Of note:

- We assumed conditional exchangeability, given age and physical functioning at start
- We assumed that the linear model fits (the relation between C, X and Y is well specified)

Traditional regression approach

- Fit a regression model (linear for a continuous outcome, logistic for a binary outcome, or another GLM model)
- Use the exposure and the confounder set C as independent variables in the model
- Report the coefficient for exposure, representing the exposure effect conditional on C

Some drawbacks of this traditional approach

- A standard regression model assumes that the effect of exposure on the outcome is the same for all subjects
  - No interaction between exposure and other covariates in the model

- Standard regression yields a conditional effect
  - often in causal inference we want marginal effects
  - For certain regression models (such as the logistic model) conditional and marginal effects differ, even if there would not be any confounding (non collapsibility of the odds ratio)

- We may not be interested in the parameters of the model but in a different causal contrast
  - For example, we have a binary outcome and are interested in a causal risk difference instead of an odds ratio

PART 2 G-computation: outcome regression with standardization

Standardization with outcome modelling

Estimate $E\big(Y(x)\big)$, average potential outcome if treatment X was set to x

$$E\big(Y(x)\big) =$$

$$= E_C E(Y(x)|C) \qquad \text{Double expectation rule}$$

$$= E_C E(Y(x)|C, X = x) \qquad \text{Conditional exchangebility}$$

$$= E_C E(Y|C, X = x) \qquad \text{Consistency}$$

We estimate $E\big(Y(x)\big)$ in two steps

1. model $E(Y|C, X = x)$
2. Estimate $E_C E(Y|C, X = x)$

Simplest situation. One binary C

$$E\big(Y(x)\big) =$$
$$E_C E(Y|C, X = x) =$$

$$E(Y \,|\, C = 0, X = x)P(C = 0) + E(Y|C = 1, X = x)P(C = 1)$$

Simplest situation. One binary C

$$E\big(Y(x)\big) =$$
$$E(Y|C=0, X=x)P(C=0) +$$
$$E(Y|C=1, X=x)P(C=1)$$

Let x=0

$\hat{E}\ (Y \mid C=0, X=0)$= 1/4

$\hat{E}\ (Y \mid C=1, X=0)$= 2/4

$P(C=0) =$ 6/16 =0.375

$P(C=1) =$ 10/16=0.625

$\hat{E}\ (Y(0))$ = 1/4*0.375+ 2/4*0.625=0.40625

Same for

$\hat{E}\ (Y(1))$ = 0.7083(check it yourself)

	C	X	Y
1	0	0	0
2	0	0	1
3	0	0	0
4	0	0	0
5	0	1	0
6	0	1	1
7	1	0	1
8	1	0	1
9	1	0	0
10	1	0	0
11	1	1	1
12	1	1	1
13	1	1	0
14	1	1	1
15	1	1	1
16	1	1	1

Average treatment effect (ATE)

$\hat{E}\ (Y(0))$ =0.40625

$\hat{E}\ (Y(1))$ = 0.7083

ATE = 0.302 (causal risk difference)

30.2% difference in outcome Y=1, if everyone receives X=1 versus X=0

Many confounders, how to estimate $E\big(Y(x)\big)$?

Standardization with outcome modelling

Step 1. Model $E(Y|C,X)$

- Fit a regression of the outcome on the exposure and relevant covariates, using the observed data set.
- Interactions between exposure and other covariates can be included

Step 2. Calculate for each individual: $\widehat{\mathrm{E}}(Y_i|C = c_i,\ X = x)$

- Use the model to predict for each observation, the expected potential outcome for exposure level x

Step 3 Estimate $E_C E(Y|C,X = x)$

- marginalize $\mathrm{E}(Y|X{=}x,C)$ over the distribution of C
- Can be approximated by averaging over the observed values of C, i.e., its empirical distribution.

G-computation in the example

Step 1. Model $E(Y|C,X)$

```
# outcome regression with interactions
lin.model.int<-lm(sf_phys6~exposure*age+exposure*sf_phys0,
 data=kidney)
```

Step 2. Calculate for each individual: $\widehat{\mathrm{E}}(Y_i|C = c_i, X = 1)$

```
# predict outcome under peritoneal
kidney.1 <- kidney #dataset where everyone receives
 hemodialysis
kidney.1$exposure <- 1
EYhat1<-predict(lin.model.int, newdata=kidney.1)
```

Same for $\widehat{\mathrm{E}}(Y_i|C = c_i, X = 0)$

G-computation in the example

Step 3 Estimate $E_C E(Y|C, X = x)$

- Can be approximated by averaging over the observed values of C, i.e., its empirical distribution.

```
# average over the population
 EY1<-mean(EYhat1)
 EY0<-mean(EYhat0)
 # Calculate ATE
 ATE <- EY1-EY0
 c(EY0,EY1,ATE)
 54.1603603 55.0434977  0.8831374
```

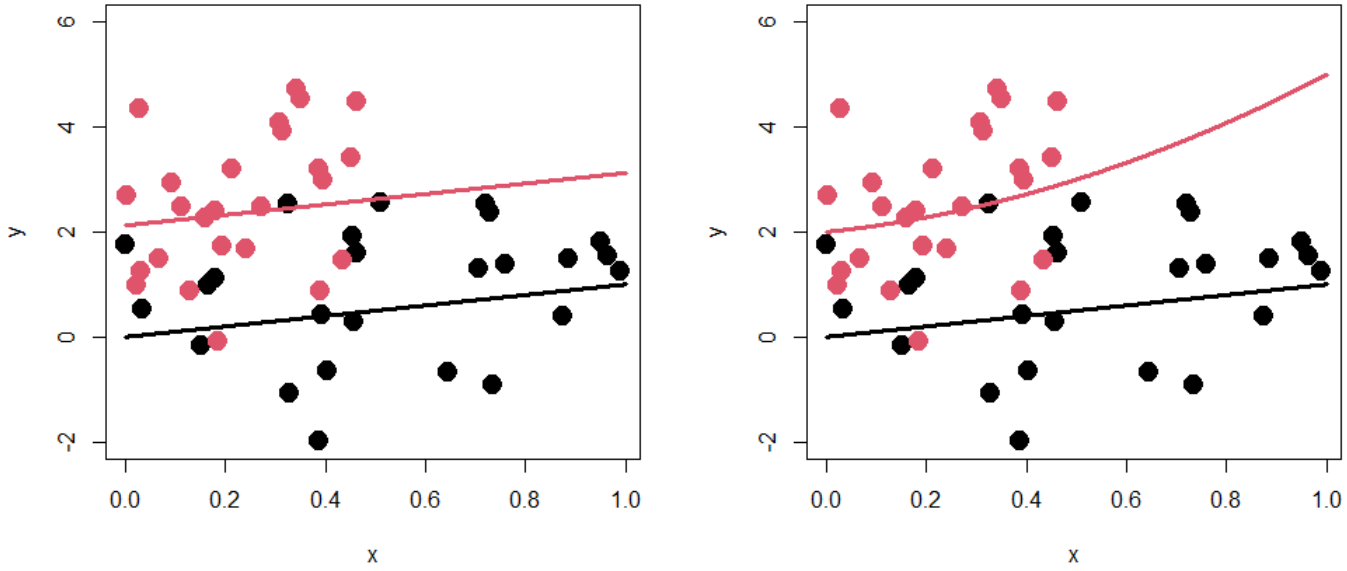
Assumptions

- Assumptions of positivity, consistency and conditional exchangeability should hold
- Additional assumption: the outcome model is correctly specified
  - This could be checked from the data and adapted accordingly.
- Problem of non positivity
  - If there is little overlap in the values of some C between the X=1 and X=0 groups : extrapolation problems

Non positivity

Poor overlap

- Little overlap between exposed/non exposed
- Regression models extrapolate over regions with little/no data
- No warning about lack of fit
- Too little information to be able to choose between models



Choice of outcome model

We often use regression models:

- For a binary outcome: outcome model may be a logistic model
- For time to event data: a survival model
- For counts: a Poisson model

Alternatively, machine learning methods can be used to model the effect of X and C on Y, but be aware:

- Do not use machine learning for confounder selection
- Think about sample splitting
- More in Hernán MA, Robins JM (2024). Causal Inference: What If. Chapter 18.3

Choice of population

We may be interested in the causal effect in the total population. The ATE

$$E\big(Y(1)\big) - E\big(Y(0)\big)$$

Alternatively, we could be interested in the effect in the subgroup currently being exposed:

$$E(Y(1) - Y(0) \mid X = 1)$$

This difference is called the ATT, the average treatment effect in the treated.

"What has been gained by those currently receiving treatment?

Is calculated by

Step 3 Averaging the estimated individual potential outcomes in the subgroup of those currently being treated.

Choice of population

Or one could be interested in a subset defined by other variable

$$E(Y(1) - Y(0) | age < 26)$$

“Should we give the treatment to young people?”

Which is calculated by :

Step 3 Averaging the estimated individual potential outcomes in the subgroup of those <26 years

Standard errors

- Standard errors and confidence intervals
  - Analytically
  - Bootstrap

More to read: Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique by Snowden, Rose, and Mortimer, Am J Epidemiol. 2011;173(7):731–738

The group assignment

How is it going?

How is the writing going?