Causal inference

Statistical methods in Cancer Epidemiology using R

Janne Pitkäniemi

Faculty of Social Sciences, University of Tampere
Finnish Cancer Registry

janne.pitkaniemi@cancer.fi

March,29 2020

Contents

- 9.15-9.30 Previous practical recap
- 9.30-11.00 Causal inference
- Lectures are based on
    - Judea Pearl - “Causality”
    - Judea Pearl - “The Book of Why”

“. . . all approaches to causation are variants or abstractions of . . . structural theory . . . ”. Judea Pearl

Ladder of causal inference

- “ladder of causal inference” (Pearl J.)
  - association (seeing)
  - intervention (doing)
  - counterfactuals (imagining)
- We will discover how directed acyclic graphs describe conditional (in)dependencies;
- how the do-calculus describes interventions
- Structural Causal Models allow us to imagine what could have been.

What is a causal effect?

- interventionist position and say that a variable X has a causal influence on Y if changing X leads to changes in Y.

- This position is a very useful one in practice, but not everybody agrees with it (e.g., Cartwright, 2007).

Necessary, sufficient, contributory cause

- **Necessary** For x to be a necessary cause of y, the presence of y must imply the prior occurrence of x. The presence of x, however, does not imply that y will occur

- **Sufficient causes** For x to be a sufficient cause of y, the presence of x must imply the subsequent occurrence of y. However, another cause z may independently cause y. Thus the presence of y does not require the prior occurrence of x.

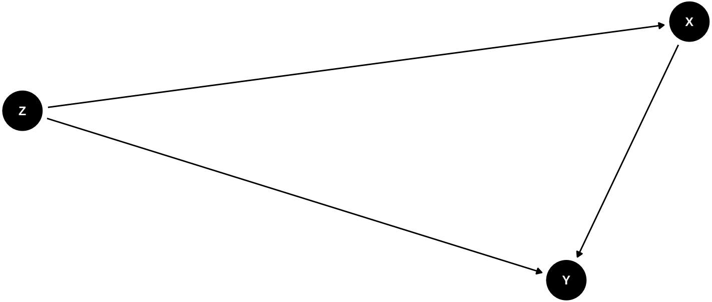
- **Contributory causes** For x to be a contributory cause of y, the presence of x must increase the likelihood of y. If the likelihood is 100%, then x is instead called sufficient. A contributory cause may also be necessary.

Model elements - Directed Acyclic Graphs

- Causal models have formal structures with elements with specific properties.
- Structural causal models (SCMs): DAGs that portray causal assumptions about a set of variables.
- In DAGs, it doesn't matter what form the relationship between two variables takes, only its direction.
- Directed arrows (E) and nodes (V) $G = (V, E)$
- **Acyclic**: no simultaneity, the future does not cause the past
- **Assumptions**:
  - Absence of variables: all common (observed and unobserved) causes of any pair of variables
  - Absence of arrows: zero causal effect

Model elements - Directed Acyclic Graphs

- **Chains** are straight line connections with arrows pointing from cause to effect. example of **chain** $Z \rightarrow X \rightarrow Y$
- **fork** $X \leftarrow Z \rightarrow Y$
- **inverted fork** $Z \rightarrow Y \leftarrow X$

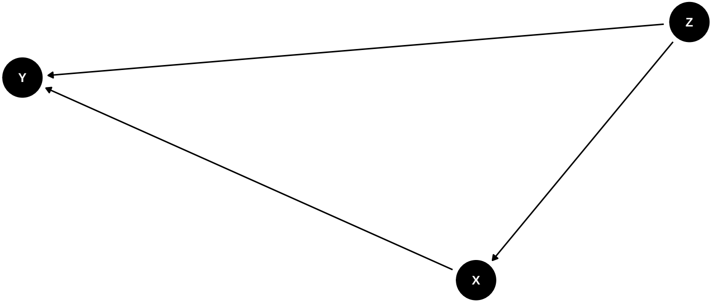


Model elements - Directed Acyclic Graphs

- Parents (Children): directly causing (caused by) a node
- Ancestors (Descendents): directly or indirectly causing (caused by) a node
- Path: an acyclic sequence of adjacent nodes
  - Causal path: all arrows pointing away from T and into Y
  - Non-causal path: some arrows going against causal order
- **Collider**: a vertex on a path with two incoming arrows
- A **mediator** node modifies the effect of other causes on an outcome (as opposed to simply affecting the outcome
- A **confounder** node affects multiple outcomes, creating a positive correlation among them

Confounding

- Sample 100,000
- Binary exposure, prevalence 30%
- Binary confounder, prevalence 30%
- Intercept =1.0
- OR(Y,X)=2.0
- OR(Y,Z)=10.0



Confounding

Not controlling for confounding

```
round(ci.exp(glm(y~x,family=binomial)),3)
```

```
            exp(Est.)  2.5% 97.5%
(Intercept)     1.649 1.624 1.675
x               1.857 1.801 1.914
```

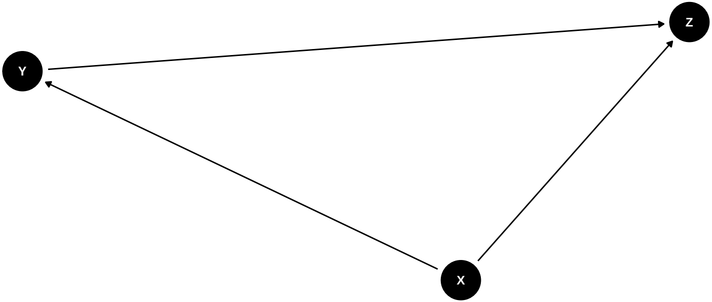
Control for confounding

```
round(ci.exp(glm(y~x+z,family=binomial)),3)
```

```
            exp(Est.)  2.5%  97.5%
(Intercept)     0.992 0.975  1.010
x               2.039 1.974  2.105
z              10.141 9.694 10.609
```

Falling for collider

- Sample 100,000
- Binary exposure, prevalence 30%
- Binary confounder, prevalence 30%
- Intercept =1.0
- OR(Y,X)=2.0
- OR(Y,Z)=10.0



Falling for collider

This is the analysis that you would do assuming that Z is confounder - **Biased**

```
round(ci.exp(glm(y~x+z,family=binomial)),3)
```

```
            exp(Est.)  2.5% 97.5%
(Intercept)     0.187 0.180 0.193
x               1.607 1.558 1.658
z               9.596 9.250 9.954
```

Ignoring the collider - you get **Unbiased** answer OR(Y,X)=2

```
round(ci.exp(glm(y~x,family=binomial)),3)
```

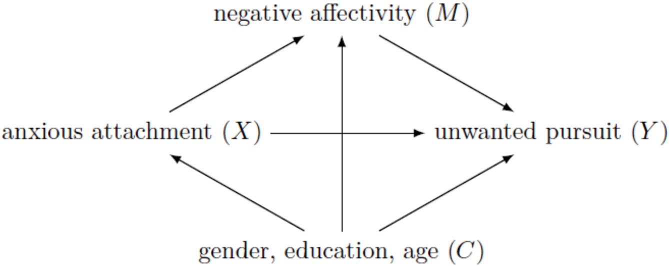
```
            exp(Est.)  2.5% 97.5%
(Intercept)     0.996 0.981 1.011
x               2.040 1.983 2.099
```

Mediation analysis

- A mediated effect is also called an indirect effect and it occurs when the effect of the independent variable on the dependent variable is — as the name says — mediated by another variable:
- A mediator must be endogenous: This means that the mediator cannot be the treatment or the conditions of the study
- The mediator itself must be dependent on these exogenous variables.
- A mediator (M) must reveal more about how the independent variable impacts the dependent variable: A mediator reveals something more about the process.

Mediation analysis

De Smet et al. (2012) and Loeys et al. (2013) proposed emotional distress or the amount of negative affectivity experienced during the breakup as a mediating variable for the effect of attachment style towards the ex-partner before the breakup on displaying unwanted pursuit behaviors after the breakup.



$

Mediation analysis - potential outcome

- Natural effect models are conditional mean models for nested counterfactuals $Y(x, M(x^*))$:

$$EY(x, M(x^*))|C = \beta_0 + \beta_1 x + \beta_2 x^* + \beta_3 C,$$

- $exp(\beta_1)$ captures the **natural direct effect** rate ratio (x, x+1)

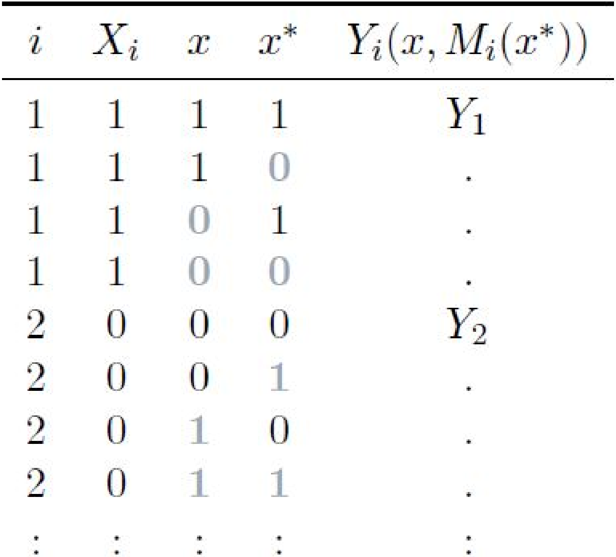
$$\frac{E(Y(x+1, M(x))|C)}{E(Y\ (x, M(x))|C)}$$

- $exp(\beta_2)$ captures the **natural indirect effect** rate ratio, corresponding to a one-unit increase in exposure level.

$$\frac{E(Y(x, M(x+1))|C)}{E(Y\ (x, M(x))|C)}$$

Mediation analysis - potential outcome

- expanding the data along unobserved (x, x^*) combinations



i	X_i	x	x^*	$Y_i(x, M_i(x^*))$
1	1	1	1	Y_1
1	1	1	0	.
1	1	0	1	.
1	1	0	0	.
2	0	0	0	Y_2
2	0	0	1	.
2	0	1	0	.
2	0	1	1	.
⋮	⋮	⋮	⋮	⋮

Mediation analysis - potential outcome

- imputation-based approach requires fitting a mean model for the outcome.

$$logit(Pr(Y = 1|X, M, C)) = \gamma_0 + \gamma_1 X + \gamma_2 M + \gamma_C,$$

Mediation analysis - potential outcome

- X: dichotomized version of anxious attachment level (attbin).
- M: negative affectivity (negaff) has been standardized
- Y: unwanted pursuit behavior (UPB),(=1) for the respondent has engaged in any unwanted pursuit behaviors

```
impFit <- glm(UPB ~ factor(attbin) + negaff + gender + educ + age, family = bin
expData <- neImpute(impFit)
head(expData, 4)
```

```
  id attbin0 attbin1        att attcat    negaff initiator gender educ age
1  1       1       1  1.0005617      M  0.840461    myself      F    M  41
2  1       0       1  1.0005617      M  0.840461    myself      F    M  41
3  2       0       0 -0.7085889      L -1.257465      both      M    M  42
4  2       1       0 -0.7085889      L -1.257465      both      M    M  42
        UPB
1 0.4916179
2 0.3841749
3 0.1870645
4 0.2629165
```

Mediation analysis

```
neMod1 <- neModel(UPB ~ attbin0 + attbin1 + gender + educ + age,
                  family = binomial("logit"), expData = expData, se = "robust")
summary(neMod1)
```

```
Natural effect model
with robust standard errors based on the sandwich estimator
---
Exposure: attbin
Mediator(s): negaff
---
Parameter estimates:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.92157    0.68922  -1.337    0.181
attbin01     0.40153    0.21345   1.881    0.060 .
attbin11     0.34069    0.08054   4.230 2.33e-05 ***
genderM      0.29399    0.22501   1.307    0.191
educM        0.34624    0.48167   0.719    0.472
educH        0.51428    0.48782   1.054    0.292
age         -0.01219    0.01194  -1.021    0.307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mediation analysis - results

- **Natural Direct effect:** for a subject with baseline covariate levels C, altering the level of anxious attachment from low (=0) to high (=1), while controlling negative affectivity at levels as naturally observed at any given level of anxious attachment x, increases the odds of displaying unwanted pursuit behaviors with a factor

```
exp(Est.)      2.5%     97.5%
1.4941107 0.9833185 2.2702377
```

- **Natural indirect effect**: Altering levels of negative affectivity as observed at low anxious attachment scores to levels that would have been observed at high anxious attachment scores, while controlling their anxious attachment score at any given level x, increases the odds of displaying unwanted pursuit behaviors with a factor

```
exp(Est.)      2.5%     97.5%
 1.405922  1.200624  1.646326
```

Survival model with long term survivors

- In standard survival analysis sufficient follow-up assumed
- Fraction of the study subjects will never experice the event of interest
    - fraction of patients treated will be cured
    - fraction of population non-susceptible (immune) to event

Survival model with long term survivors

- Let D be partially latent variable indicating if subject is susceptible, cured $D = 1$ and $D = 0$ otherwise

- Then the probability of an event for a subject is the product of probability of beeing susceptible and event at time t

$$P(D = 1 \mid X_i) f(t \mid D == 1, X_i)$$

- It is convienient to specify survivor function

$$S(t \mid D == 1, X_i) = P(T > t \mid D == 1, X_i)$$

Survival model with long term survivors

- Susceptibility can be modelled with any parametric function for binary rv. f.ex logistic

$$P(D = 1 \mid X_i) = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)}$$

- Time-to-event with any parametric function exponential, weibull as well as proportional hazards
- Problems: identifiability between susceptiblity intercept and time-to-event parameters, need more censored observations
- Maller and Zhou presented a testing procedure for susceptiblity fraction 0.
- restrict to problems we consensus is that there is group of non-susceptibles in the population
- separate modelling more informative of the problem if the groups exist

Survival model with long term survivors

```r
library(smcure);
library(survival)
data("e1684");
attach(e1684);
head(e1684)
```

```
  TRT FAILTIME FAILCENS         AGE SEX
1   1  1.15068        1 -11.0359437   0
2   1  0.62466        1  -5.1290437   0
3   0  1.89863        0  23.1859563   1
4   0  0.45479        1  11.1448563   1
5   0  2.09041        1 -13.3208437   0
6   1  9.38356        0   0.9421563   0
```

```r
#Kaplan Meier estimate of S,CDF
fit <- survfit(Surv(FAILTIME,FAILCENS)~TRT,data = e1684)

#LTS model
pd <- smcure(Surv(FAILTIME,FAILCENS)~TRT,cureform=~TRT,
             data=e1684,model="ph",Var = FALSE)
```

```
Program is running..be patient... done.
Call:
smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT, cureform = ~TRT,
    data = e1684, model = "ph", Var = FALSE)
```

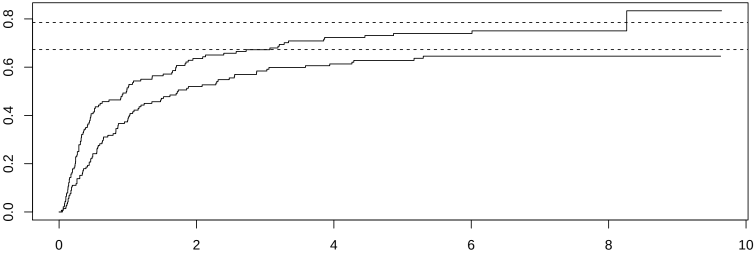
Survival model with long term survivors

```
Call:
smcure(formula = Surv(FAILTIME, FAILCENS) ~ TRT, cureform = ~TRT,
    data = e1684, model = "ph", Var = FALSE)

Cure probability model:
              Estimate
(Intercept)  1.2957164
TRT         -0.5747481


Failure time distribution model:
      Estimate
TRT -0.1318355
```

Survival model with long term survivors

```
# extract suscceptible proportions
lp1<-coef(res$logistfit)[1]
lp2<-sum(coef(res$logistfit))
p1<-exp(lp1)/(1+exp(lp1))
p2<-exp(lp2)/(1+exp(lp2))
```

Survival model with long term survivors



- Proportion of immunes in TRT==0 is 0.2148868
- Proportion of immunes in TRT==1 is 0.3271798
- HR for TRT==1 vs TRT==0 for non-immunes is 0.8764852