Applied Regression III

CMED6020 – Session 5

Eric Lau (ehylau@hku.hk)

School of Public Health
The University of Hong Kong

22 Feb 2021

Session 5 learning objectives

After this session, students should be able to

- Identify and handle multicollinearity
- Account for confounding factors in regression model
- Assess potential effect modifiers in regression model
- Perform basic mediation analysis

Multicollinearity

Multicollinearity

- Collinearity refers to strong linear dependency (high correlation) between two predictor variables
  - e.g. personal income and household income
  - Collinear variables give similar information
- For perfect collinearity between two predictor variables X_1 and X_2
  - We can find constants *a*, *b* so that

$$X_1 = a + bX_2$$

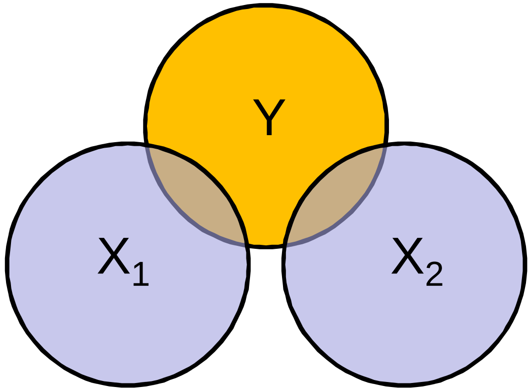
  - e.g. X_1 = degree in Celsius, X_2 = degree in Fahrenheit
- Usually correlation > 0.8 indicates severe collinearity problem
- Multicollinearity:
  - Predictor variable strongly depends on other predictor variables linearly
  - e.g. $X_4 = a + b_1X_1 + b_2X_2 + b_3X_3$ has a very good fit

Problems with multicollinearity

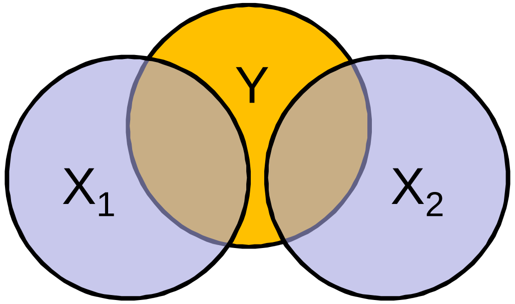
- Multicollinearity does not violate any of the assumptions of linear regression model
- But inflates standard error of the estimated coefficients
  - e.g. for perfectly correlated predictors $X_1 = a + bX_2$
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 = (\alpha + a\beta_1) + (b\beta_1 + \beta_2)X_2$
  - We can always find different sets of (α, β_1, β_2) to represent exactly the same relation
  - For highly correlated predictors, it will be difficult to distinguish their effect on the outcome variable

→ the uncertainty in those estimated coefficients will be large

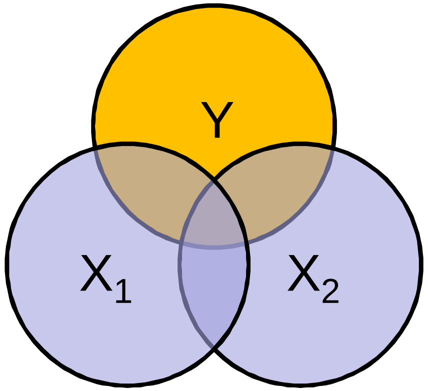
Venn diagram illustration of multicollinearity



Independent predictors, weak fit



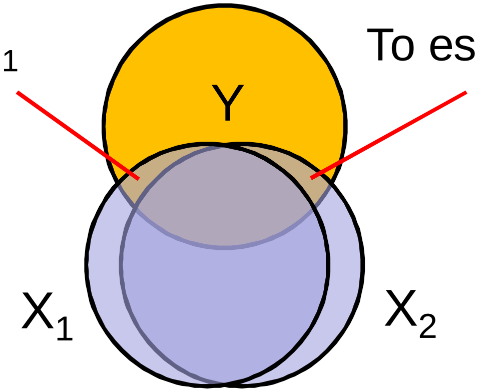
Independent predictors, strong fit



Mildly collinear predictors, moderate fit

To estimate β_1

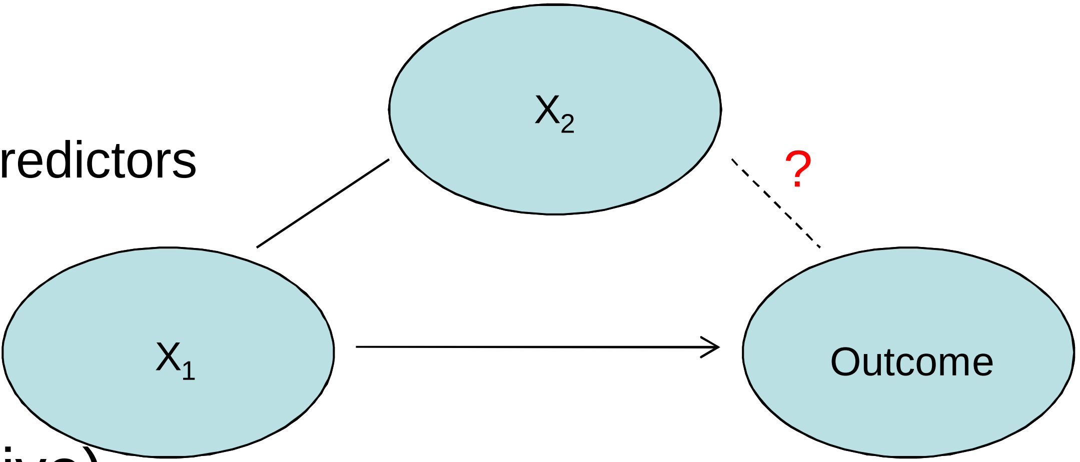
To estimate β_2



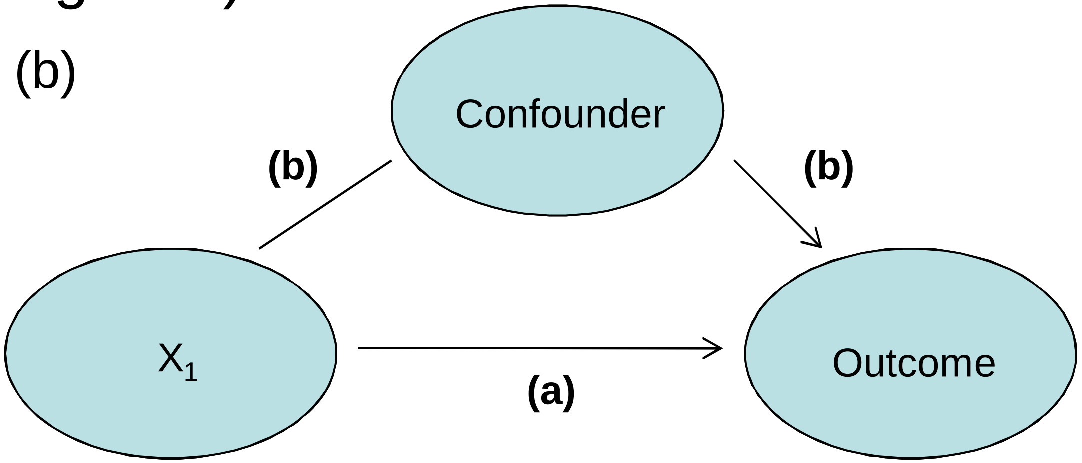
Strongly collinear predictors, moderate fit, larger uncertainty in the effect of X_1 or X_2

Multicollinearity versus confounding

- Uncommon for all predictors to be approximately independent
  - Especially when confounders are present
- Multicollinearity
  - Concern with correlation among predictors



- Confounding (positive or negative)
  - Observed association: (a) + (b)
  - To estimate: (a)
  - Exclusion of important confounders leads to biased estimation of (a)



Multicollinearity diagnostics

- Scatterplot between all predictor variables
  - More useful to identify pairwise collinearity
  - Less helpful to identify multicollinearity of multiple predictors
- Variance inflation factor (VIF)
  - $\mathrm{VIF}_{\mathrm{j}} = 1 / (1 – R_j^2)$
  - R_j^2 is the R^2 when predictor X_j is regressed on all other predictors
  - A value of VIF > 10 indicates potential multicollinearity problem

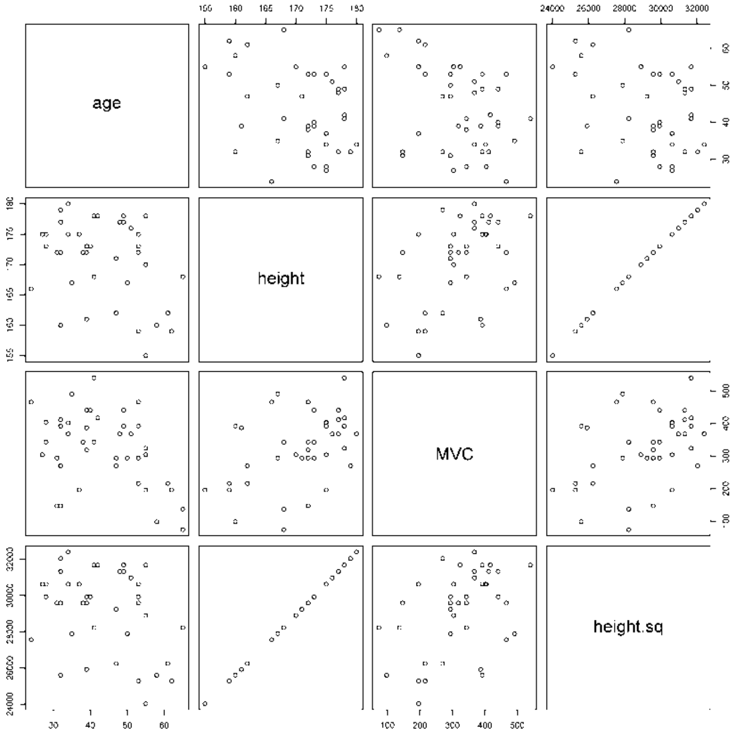
Dealing with multicollinearity

- Do nothing
  - Estimated coefficients are still unbiased
  - but inefficient estimation
  - Affect the main objectives of the analysis?
- Increase the sample size
- Polynomial terms and interactions: centering
  - By subtracting the mean values from the variable (useful for polynomial and interactions only)
  - i.e. $x \rightarrow x - \bar{x}$ $\qquad x^2 \rightarrow (x - \bar{x})^2$
- Drop one or more variables causing multicollinearity
  - Most appropriate when two variables basically measure the same thing
  - Problematic if the variable is an important confounder

MVC example

- Suppose we create a new variable ‘height.sq’ which is the square of the heights
- We fit the linear regression model including height.sq:

```
> mvc <- read.csv("http://web.hku.hk/~ehylau/mvc.csv")
> mvc$height.sq <- mvc$height^2
> pairs(mvc)
```



- height and height.sq highly correlated

MVC regression with highly correlated variables

```
> summary(lm(MVC~age+height+height.sq, data=mvc))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4391.6093 10766.3438   0.408   0.6857
age            -3.1422     1.4903  -2.108   0.0418 *
height        -52.3026   127.8036  -0.409   0.6847
height.sq       0.1712     0.3791   0.452   0.6542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.97 on 37 degrees of freedom
Multiple R-squared:  0.2653,    Adjusted R-squared:  0.2057
F-statistic: 4.453 on 3 and 37 DF,  p-value: 0.009074
```

VIF in R

- Install and load the “car” package

```
> require(car)
> mvc.lm3 <- lm(MVC~age+height+height.sq, data=mvc)
> vif(mvc.lm3)
        age      height   height.sq
   1.140494 2788.043855 2784.139173
```

- Very large VIF for the height terms
- For polynomial and interactions, VIF can be reduced by centering
- For other collinearity problem, centering does not help.

Centering in R

```
> mvc$ct.height <- scale(mvc$height, scale=F)
> mvc$ct.height.sq <- mvc$ct.height^2
> mvc.lm4 <- lm(MVC~age+ct.height+ct.height.sq, data=mvc)
> vif(mvc.lm4)
         age    ct.height ct.height.sq
    1.140494     1.603104     1.534462
```

Before centering:

```
> vif(mvc.lm3)
        age      height   height.sq
   1.140494 2788.043855 2784.139173
```

MVC example: regression after centering

```
> summary(mvc.lm4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  451.7998    67.0772   6.736  6.4e-08 ***
age           -3.1422     1.4903  -2.108   0.0418 *
ct.height      6.1506     3.0646   2.007   0.0521 .
ct.height.sq   0.1712     0.3791   0.452   0.6542
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99.97 on 37 degrees of freedom
Multiple R-squared:  0.2653,    Adjusted R-squared:  0.2057
F-statistic: 4.453 on 3 and 37 DF,  p-value: 0.009074
```

- Note that R^2 is the same as before

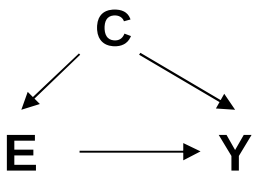
Other potential problems in regression models

- Multilevel structure
  - e.g. correlation within hospital or ward
  - Multi-level models
- Measurement error
  - Imprecise measurement in predictors will attenuate estimated coefficients toward zero

Confounding, effect modification and mediation

Background on confounding

- Common cause of exposure and outcome



- Strictly speaking, confounding is a causal concept
  - confounders cannot be identified from data
  - causal knowledge is needed
  - a DAG is an useful tool to clarify causal structure
- However, the causal structure may not be known completely in practice
  - Empirical selection may be needed
- Residual confounding in observational study

Issues of confounding

- Unable to account for confounding factors will result in biased estimate

Association of confounder with exposure is	Association of confounder with outcome is	Type of confounding	Expectation of change from unadjusted to adjusted estimate
Direct	Direct	Positive	Unadjusted > Adjusted
Direct	Inverse	Negative	Unadjusted < Adjusted
Inverse	Inverse	Positive	Unadjusted > Adjusted
Inverse	Direct	Negative	Unadjusted < Adjusted

(extracted from CMED6030)

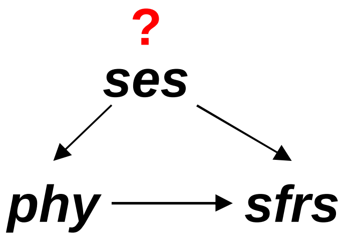
Minimizing confounding

- Automatic variable selection
  - Based on p-values of the variables
  - Based on AIC
- Relative change in estimate greater than 10%
- ‘Standard rules’ for confounding:
  - Check if C is associated with E
  - Check if C is associated with D
  - Check if C does not lie in the causal pathway between E and D

(extracted from CMED6030)

Example dataset: Cardiovascular risk in the elderly

- Dataset: examplecardio.csv
- 1000 subjects aged 60-90y
- Variables
  - *sfrs:* standardized Framingham risk score (10y cardiovascular risk)
  - *phy:* physical activity index, 0–100 (high)
  - *ses*: social economic status, 0–10 (high)
  - *age, bmi, male*
- Study objective: Assess the effect of physical activity on 10y cardiovascular risk



?

ses

phy ***sfrs***

‘Standard rules’ for confounding

- Check if *ses* is associated with *phy*

```
> summary(lm(phy~ses, data=cardio))
```

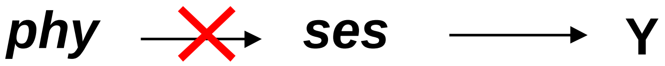
  - β = 6.373, p-value < 0.001

- Check if *ses* is associated with *sfrs*

```
> summary(lm(sfrs~ses, data=cardio))
```

  - β = -0.142, p-value < 0.001

- Check if *ses* does not lie in the casual pathway between ses and sfrs



phy **ses** **Y**

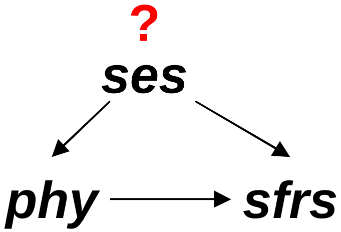
- *ses* satisfies all three conditions, so *ses* is selected as a confounder

Relative change in estimate (practice)

- Model 1:

- Model 2:

- Relative change in estimate:



Relative change in estimate (practice)

- Model 1:

  $\text{sfrs} = a + b_1 * \textit{phy} + \varepsilon$

- Model 2:

  $\text{sfrs} = a + b'_1 * \textit{phy} + b_2 * \textit{ses} + \varepsilon$

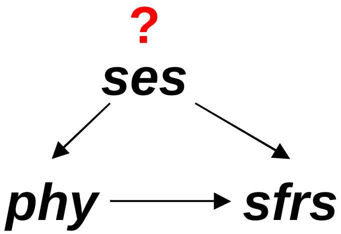
- Relative change in estimate:

  $b_1 = -0.0348$

  $b'_1 = -0.0445$

  relative change = 28% > 10%

→ *ses* is selected as a confounder



Age as a confounder?

- Model 1:

$$sfrs = a + b_1 * \textit{phy} + \varepsilon$$

- Model 2:

$$sfrs = a + b'_1 * \textit{phy} + b_2 * \textit{age} + \varepsilon$$

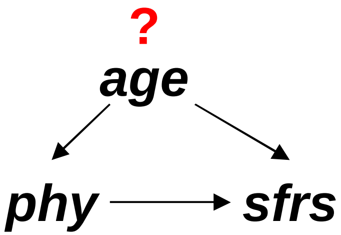
- Relative change in estimate:

$b_1 = -0.0348$

$b'_1 = -0.0350$

relative change = 0.7% < 10%

→ *age* is not selected as a confounder



Automatic variable selection

- Start with a model with the full set or a subset of the candidate variables
  - must include the exposure and outcome
  - may include higher order terms (e.g. interaction or polynomial terms)
- In each step, add each single term among the candidate variables not included in the model, or drop each term in the existing model
- Compute the AIC for each potential model
- Choose the model with lowest AIC
- Repeat until there is no more change

- Similar automatic variable selection can be done using p-values

Automatic variable selection in R

```
> require(MASS)
> cf2 <- lm(sfrs~phy+ses, data=cardio)
> stepAIC(cf2)
```

```
Start:  AIC=-789.42
sfrs ~ phy + ses

        Df  Sum of Sq     RSS      AIC
<none>                  451.39  -789.42
- ses    1      50.12   501.51  -686.13
- phy    1     458.42   909.81   -90.52
```

Drawback of automatic variable selection

- Multiple testing: incorrect P-values
- Biased towards variables with large absolute estimated coefficients
- Overfitting

“Authors should avoid stepwise methods of model building, except for the narrow application of hypothesis generation for subsequent studies. Stepwise methods include forward, backward, or combined procedures for the inclusion and exclusion of variables in a statistical model based on predetermined P value criteria. Better strategies than P value driven approaches for selecting variables are those that use external clinical judgment…”

Annals of Internal Medicine

General strategy for addressing confounding

- Based on a priori subject-matter knowledge of the causal structure, include appropriate confounder(s) in the model
- Empirical selection of confounders from the data
  - Model selection by AIC (backward selection is preferred to reduce multiple testing)
  - Change in estimate of the exposure / intervention
  - ‘Standard rules’ for confounding

Interaction effect

- Effect of exposure on outcome modified by a third factor (effect modifier)
- To identify population subgroups with higher (synergism) or lower (antagonism) from intervention
- Can be evaluated by stratified analysis on the potential effect modifier
- Can be evaluated by fitting regression model with an interaction term

Including interaction terms in regression model in R

Term	Interpretation	Example
A+B	Include both A and B	height + age
A-B	Exclude B from A	height*age – height:age
	Dropping the intercept	height + age - 1
A:B	Interaction of A and B	height:age
A*B	A + B + A:B	height*age

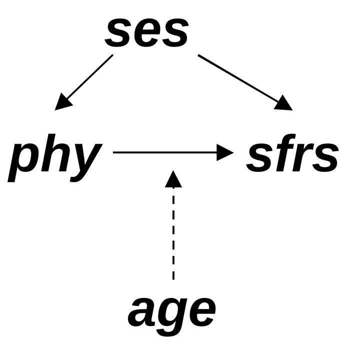
Example:

$$Y = \alpha + \beta_1\text{exposure} + \beta_2\text{confounder} + \beta_3\text{effect modifier} + \beta_4(\text{exposure*effect modifier})$$

```
Y ~ exposure*effect modifier + confounder
```

Example dataset: Cardiovascular risk in the elderly

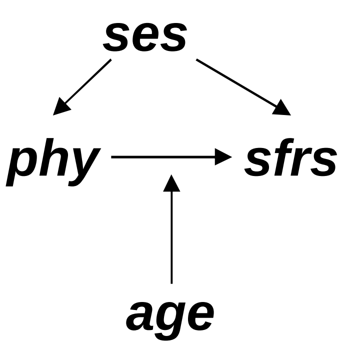
- Dataset: examplecardio.csv
- 1000 subjects aged 60-90y
- Variables
  - *sfrs:* standardized Framingham risk score (10y cardiovascular risk)
  - *phy:* physical activity index, 0–100 (high)
  - *ses*: social economic status, 0–10 (high)
  - *age, bmi, male*
- Suppose age is a potential effect modifier of the *phy* effect on *sfrs*



Cardiovascular risk with age as effect modifier

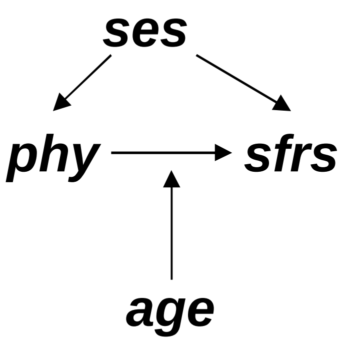
```
> age.int <- lm(sfrs~phy*age+ses, data=cardio)
> summary(age.int)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.051e+01  3.932e-01   26.73   <2e-16 ***
phy         -1.960e-01  7.169e-03  -27.34   <2e-16 ***
age         -1.212e-01  5.187e-03  -23.37   <2e-16 ***
ses          1.504e-01  1.085e-02   13.86   <2e-16 ***
phy:age      2.008e-03  9.415e-05   21.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.541 on 995 degrees of freedom
Multiple R-squared:  0.7084,    Adjusted R-squared:  0.7073
F-statistic: 604.4 on 4 and 995 DF,  p-value: < 2.2e-16
```



Interpreting interaction effect

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.051e+01  3.932e-01   26.73   <2e-16 ***
phy         -1.960e-01  7.169e-03  -27.34   <2e-16 ***
age         -1.212e-01  5.187e-03  -23.37   <2e-16 ***
ses          1.504e-01  1.085e-02   13.86   <2e-16 ***
phy:age      2.008e-03  9.415e-05   21.33   <2e-16 ***
```



ses

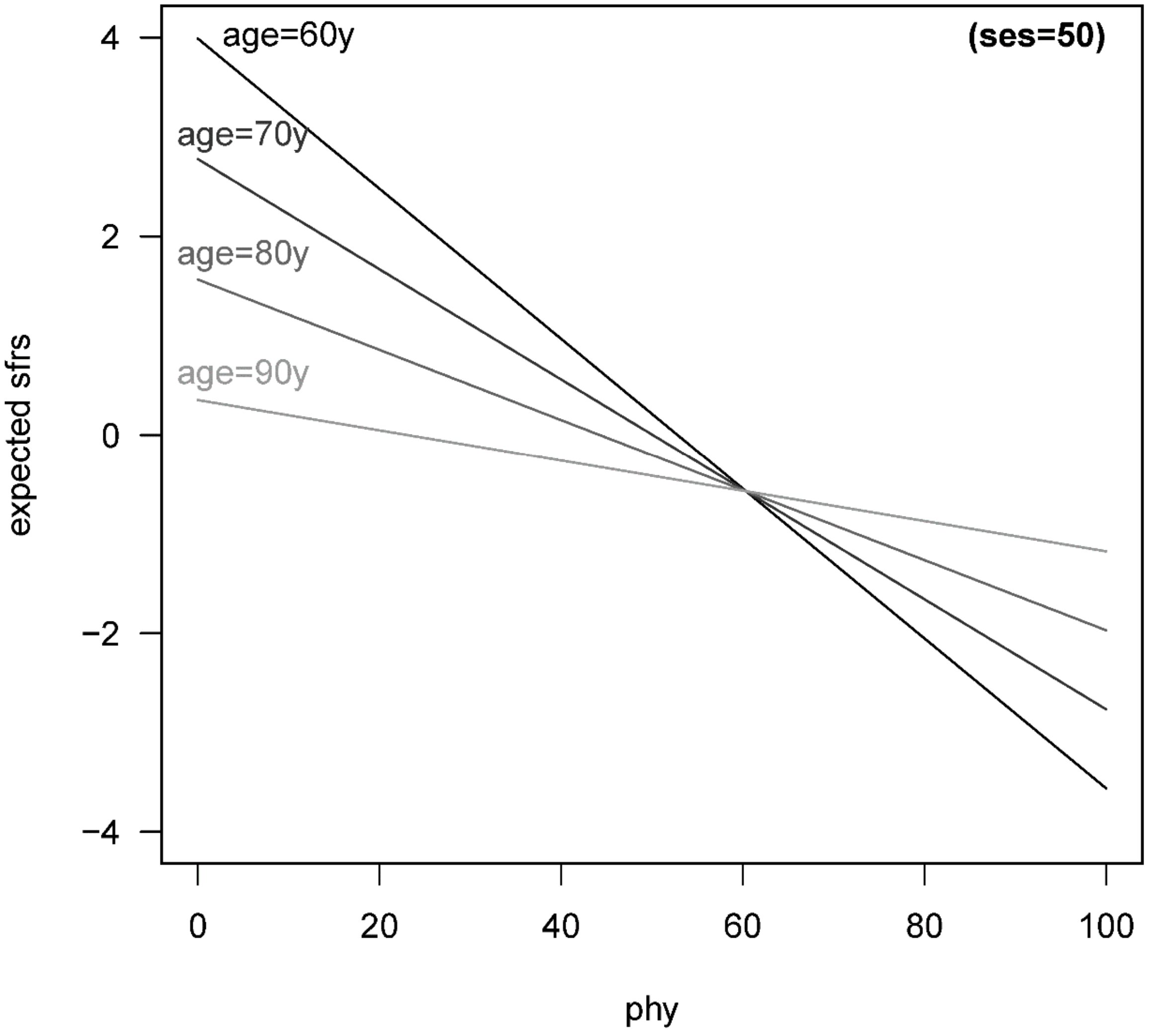
phy

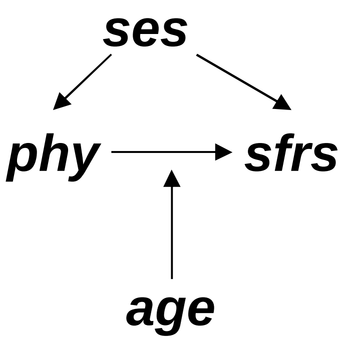
sfrs

age

- The protective effect of physical activity on cardiovascular risk diminished with older age (antagonism)

Visualizing interaction effect

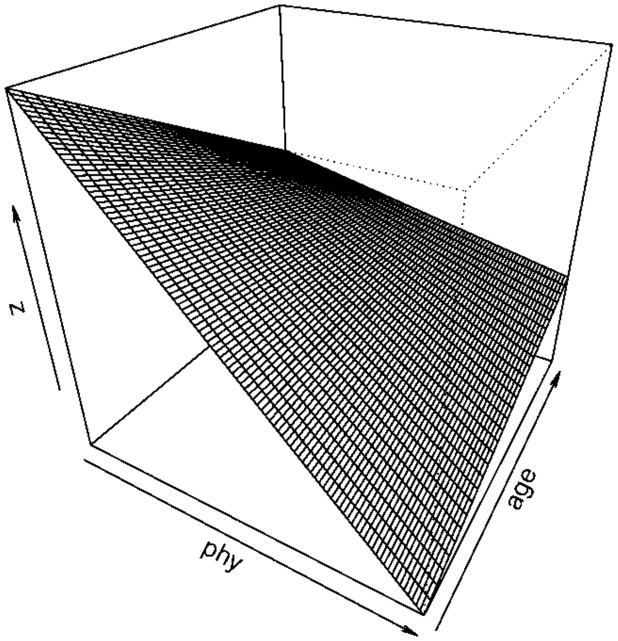



ses

phy

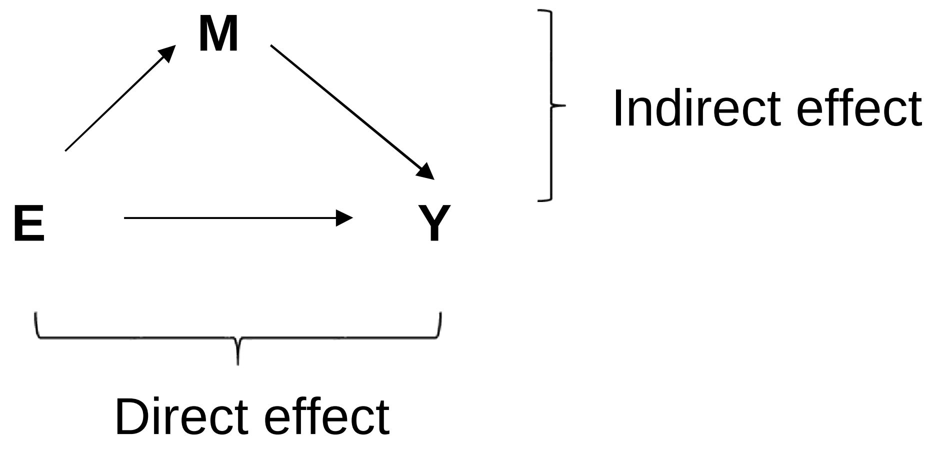
sfrs

age



Mediation

- Mediation effect: when mediator carries the effect of the exposure (E) to outcome (Y)
  - Full mediation: direct effect = 0
  - Partial mediation: direct effect ≠ 0
- To understand the mechanism of the exposure



M

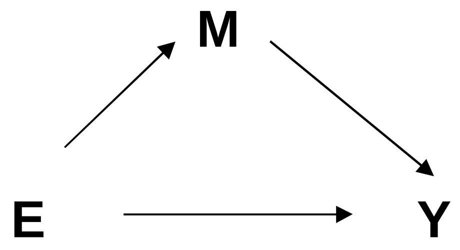
E

Y

Indirect effect

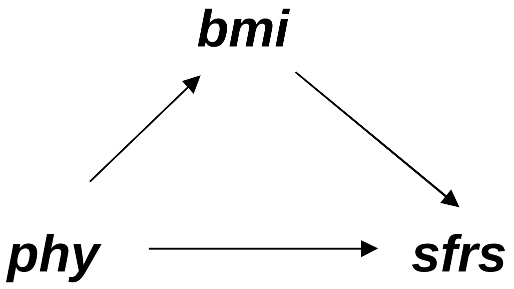
Direct effect

Baron and Kenny criteria



- Four steps to assess mediation:
  - Is exposure associated with the mediator? (M~E)
  - Is mediator associated with the outcome? (Y~M)
  - Is exposure associated with the outcome? (Y~E)
  - Does adjusting for the mediator reduce the association between the exposure and the outcome (attenuation) (compare with Y~E+M)
- The above association can be assessed by regression
- Note that the Baron and Kenny criteria are an approximation which is only valid in some simple situations but is not always valid (discussed in CMED6030)

Example dataset: Cardiovascular risk in the elderly



bmi

phy

sfrs

- Dataset: examplecardio.csv
- Suppose *bmi* is a potential mediator
- Assessing Baron and Kenny criteria:

```
> summary(lm(bmi~phy, data=cardio))
```

β = -0.096, $p < 0.001$

```
> summary(lm(sfrs~bmi, data=cardio))
```

β = -0.009, $p = 0.130$

```
> summary(lm(sfrs~phy, data=cardio))
```

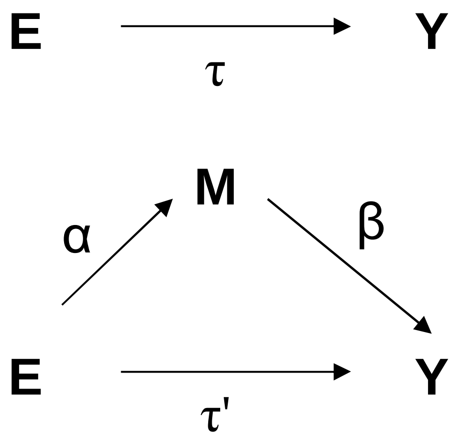
β_{phy} = -0.035, $p < 0.001$

```
> summary(lm(sfrs~phy+bmi, data=cardio))
```

$\beta_{phy'}$ = -0.041, $p < 0.001$ (no attenuation)

- *bmi* does not satisfy the 4th (and 2nd) B & K criterion

Testing for Mediation: Sobel test



E
τ
Y
M
α
β
E
τ'
Y

- The indirect effect is τ – τ' = αβ
- Sobel test, H_0: αβ = 0, under H_0

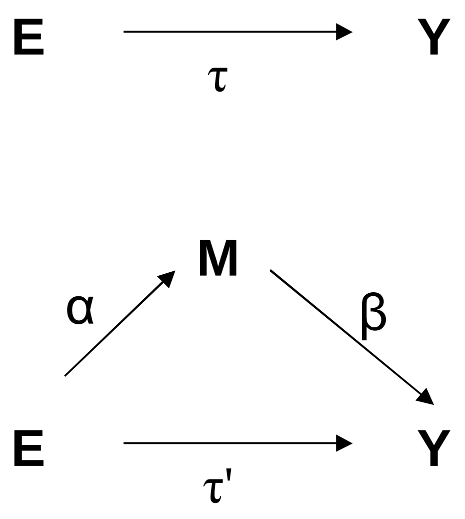
$$z = \alpha\beta / \sqrt{\alpha^2\sigma_\beta^2 + \beta^2\sigma_\alpha^2}$$

- Critical value at 5% significance level = 1.96
- α is obtained by fitting M~E
- β is obtained by fitting Y~E+M
- Other versions of Sobel-Goodman test have different forms of the denominator
- A more complicated version is needed for binary outcomes

Sobel test in R

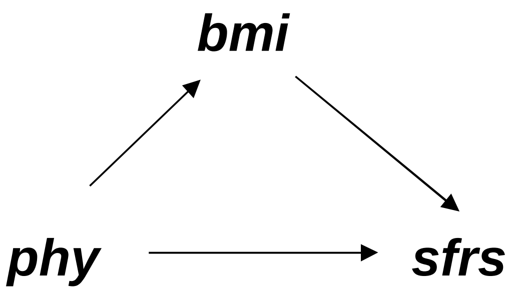
sobel(*iv*, *mv*, *dv*) (in package: multilevel)

- ***iv*** is the independent variable (E)
- ***mv*** is the mediator variable (M)
- ***dv*** is the dependent variable (Y)



Example dataset: Cardiovascular risk in the elderly

- Dataset: examplecardio.csv
- Suppose *bmi* is a potential mediator



bmi

phy

sfrs

```
> require(multilevel)
> out <- with(cardio, sobel(phy, bmi, sfrs))
```

(“with” function: use cardio dataset)

```
> 2*pnorm(out$z.value, lower.tail=F)
[1] 5.466551e-23
```

- Reject H_0 and conclude that there is significant mediation effect.
- ‘out’ stored the results for the relevant models for the mediation analysis

Sobel test (practice)

- The sobel test statistics is given by

$$z = \alpha\beta / \sqrt{\alpha^2\sigma_\beta^2 + \beta^2\sigma_\alpha^2}$$

- Reproduce the result in the previous slide by fitting appropriate models

Sobel test (practice)

- The sobel test statistics is given by

$$z = \alpha\beta / \sqrt{\alpha^2\sigma_\beta^2 + \beta^2\sigma_\alpha^2}$$

```
> m1 <- lm(bmi~phy, data=cardio)
> m2 <- lm(sfrs~bmi+phy, data=cardio)
> alpha <- coef(m1)[2]
> beta <- coef(m2)[2]
> s.alpha <- summary(m1)$coef[2,2]
> s.beta <- summary(m2)$coef[2,2]
> z <- alpha*beta/sqrt(alpha^2*s.beta^2+beta^2*s.alpha^2)
> p.value <- 2*pnorm(z, lower.tail=F)
```

Some comments on Sobel test

- Usually regarded as more accurate than Baron and Kenny criteria
- Need large sample size
  - for smaller sample size the normal assumption for the test statistics may not be valid
  - could be handled by bootstrapping (will be covered in CMED6040)
- May also need to adjust for confounding (depending on the causal structure)
- Functions like mediation.test do not consider other confounders

Review

- We have discussed multicollinearity
- We have discussed how to handle confounding, effect modification and mediation in regression analysis

Further reading

- Gelman, A., Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007

- Velentgas P., et al. *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Agency for Healthcare Research and Quality, 2013

- Vittinghoff, E, et al. *Regression methods in biostatistics.* Springer, 2005