Instrumental variable analysis

CMED6020 – Session 8

Eric Lau (ehylau@hku.hk)

School of Public Health
The University of Hong Kong

15 Mar 2021

Session 8 learning objectives

After this session, students should be able to

- Estimate treatment effect using instrumental variable analysis for non-controlled experiment
- Understand the assumptions instrumental variable analysis
- Interpret results from instrumental variable analysis

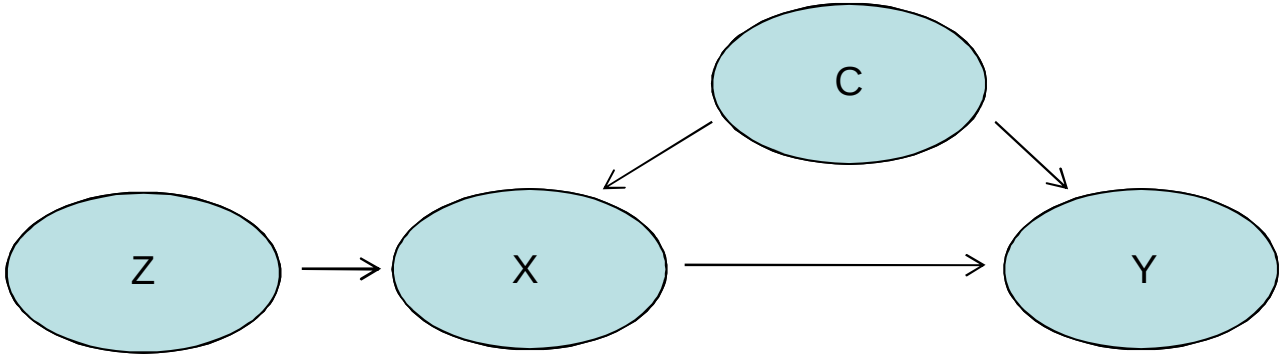
Instrumental variable analysis

Instrumental variable analysis

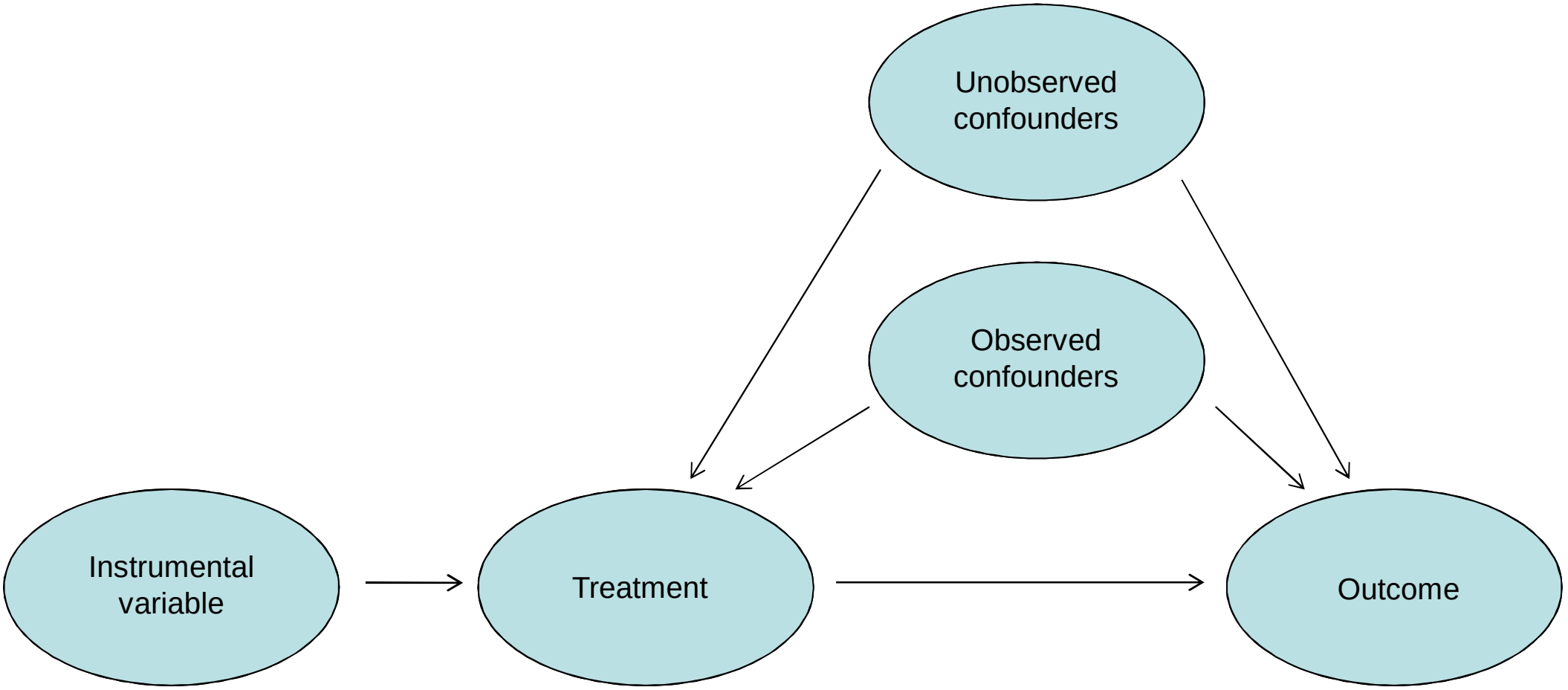
- to estimate treatment effect under selection bias
  - able to deal with hidden selection bias / confounding
  - ordinary regression and propensity score methods: only deal with observed confounder
- selection bias
  - endogeneity: treatment may tend to select patients with certain characteristics, i.e. treatment determined by other variables in the model
  - unobserved individual heterogeneity: patients with certain characteristics may select treatment
- consider outcome $Y_i = \alpha C_i + \beta X_i + e_i$
  - with selection bias due to unobserved characteristics
  - cov(X, e) ≠ 0 (endogeneity)
  - violate key assumptions of OLS to estimate β

Instrumental variable

- to isolate variation in X not correlated to unobserved confounders
- an instrumental variable (IV) Z satisfies the following conditions
  - associated with exposure X: cov(Z, X) ≠ 0
  - uncorrelated with e: cov(Z, e) = 0
  - all effects of the IV on outcome mediated through the treatment only
- we consider the case with homogeneity of treatment effect only



Instrumental variable



- effects of the IV on outcome are only mediated through the treatment
- IV uncorrelated with unobserved confounders

Examples of IV

Outcome	**‘Risk factor’**	**IV**
Health	Smoking	Tobacco tax
Complications	Drug	Drug prescribing preference
Child nutrition	Mother’s schooling and education	Mother’s frequency of surfing the internet

Estimation of treatment effect

1. Wald estimator

- $\beta_{XY} = \beta_{ZY} / \beta_{ZX}$
- Standard error approximated by Fieller's theorem:
- For two normally distributed random variable a, b

$$Var\left(\frac{a}{b}\right) = \left(\frac{a}{b}\right)^2 \left(\frac{Var(a)}{a^2} + \frac{Var(b)}{b^2}\right)$$

Estimation of treatment effect

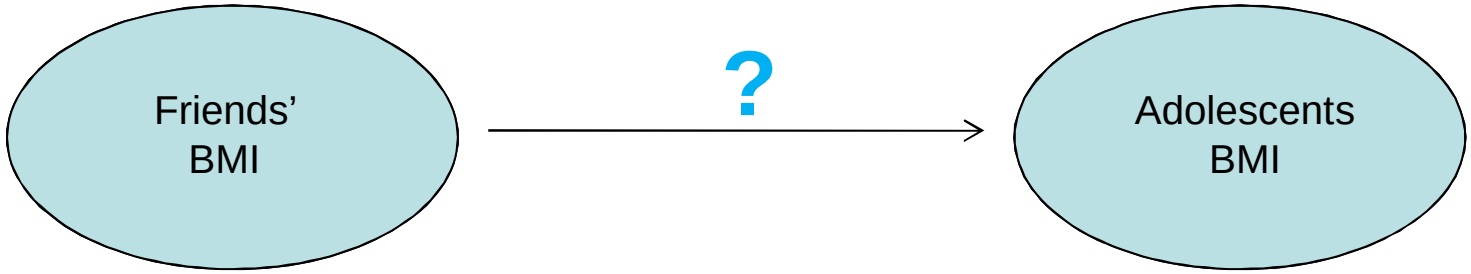
2. two stage least square (2SLS)

- first stage regression (exposure on treatment):
  - $X_i = \gamma Z_i + \varphi C_i + u_i$ for continuous or binary treatment X
  - $\text{logit}(p(X_i=1)) = \gamma Z_i + \varphi C_i$ for binary treatment T
  - obtain $\hat{X}, \hat{p}$ (propensity score) from the model

  $\Rightarrow$ effects of the IV on outcome are only mediated through the treatment
- second stage regression (outcome on predicted exposure):
  - $Y_i = \alpha C_i + \beta \hat{X} + e_i$ to estimate risk difference
  - $\text{logit}(p(Y_i=1)) = \alpha C_i + \beta \hat{p}$ to estimate relative risk
- 2SLS estimate of the treatment effect = $\hat{\beta}$

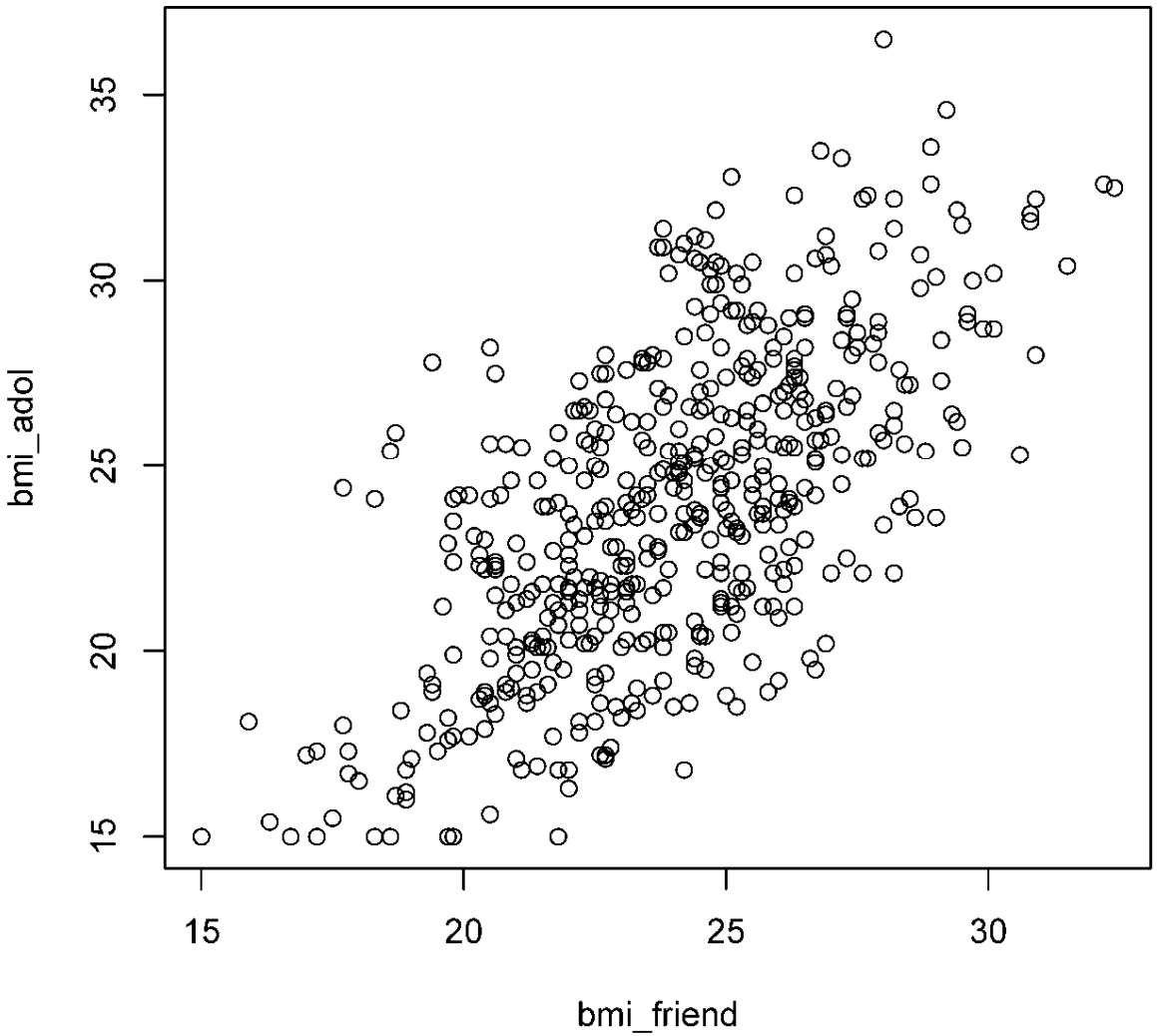
Example - BMI

- are adolescents BMI affected by close friends' BMI?



  - as a preliminary analysis, fit a linear regression to explain adolescents BMI by close friend's BMI

Scatter plot - BMI



Ordinary linear regression - BMI

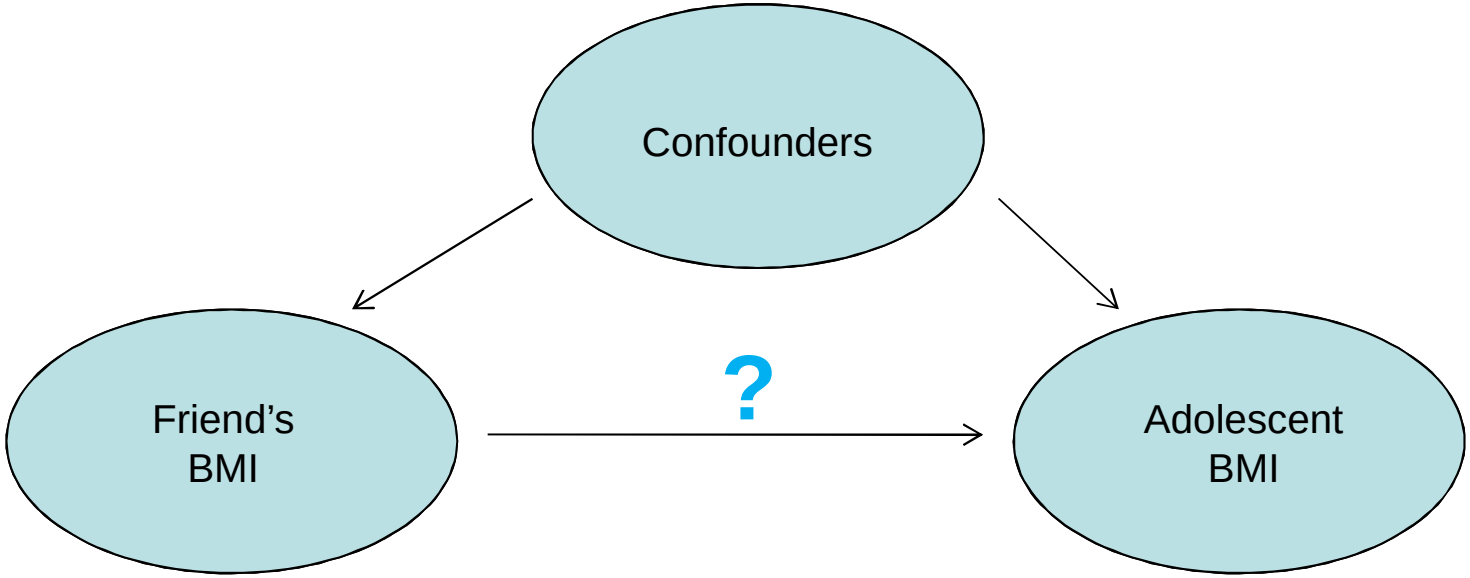
```
> summary(lm(bmi_adol~bmi_friend, data=bmi))


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5919     1.1894   1.338    0.181
bmi_friend    0.9318     0.0492  18.940   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

– an increase of 1 kg/m^2 in close friend’s average BMI is associated with 0.93 kg/m^2 increase of an adolescent’s BMI

– however this result has not accounted for confounding factors

Unmeasured confounders - BMI

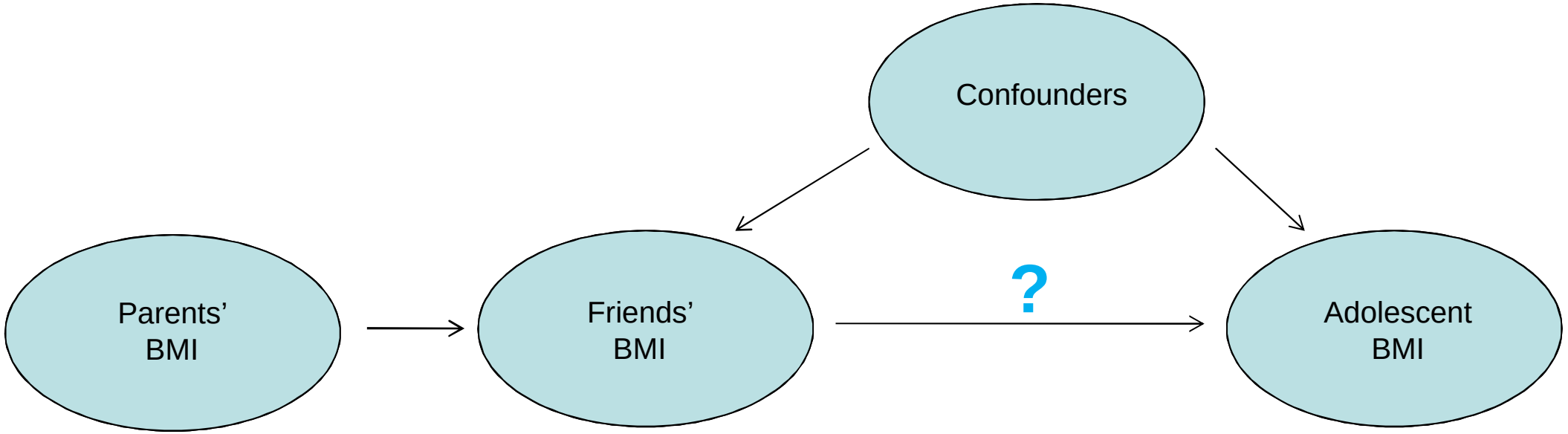
- measured/unmeasured confounders?
  - education, diet, etc…



  - use instrumental variable to correct for unmeasured/unobserved confounders

choice of IV - BMI

- example of appropriate IV
  - average BMI of close friends' parents



  - should be highly correlated with friends' BMI
  - not likely to affect adolescent BMI directly or indirectly except through friend's BMI
- dataset in 'examplelbmiiva.csv'

Wald estimator

- $\beta_{XY} = \beta_{ZY} / \beta_{ZX}$

```
> bmi.zy <- lm(bmi_adol~bmi_parent, data=bmi)
> bmi.zx <- lm(bmi_friend~bmi_parent, data=bmi)
> b.zy <- coef(bmi.zy)[2]
> b.zx <- coef(bmi.zx)[2]
> b.xy <- b.zy/b.zx
bmi_parent
  1.017007
```

Wald estimator (cont'd)

- Fieller's theorem: $Var\left(\frac{a}{b}\right) = \left(\frac{a}{b}\right)^2 \left(\frac{Var(a)}{a^2} + \frac{Var(b)}{b^2}\right)$

```
> var.zy <- vcov(bmi.zy)[2,2]
> var.zx <- vcov(bmi.zx)[2,2]
> var.xy <- (b.zy/b.zx)^2*(var.zy/b.zy^2+var.zx/b.zx^2)
> var.xy
bmi_parent
0.01490696
> sqrt(var.xy)
bmi_parent
  0.122094
```

Two stage least squares (2SLS) in R

ivreg(*formula, data, subset, …*) (in package: AER)

- *formula*: in a form y ~ x | z, x are the endogenous regressors and z is the instrument variable
  - if there are other exogenous regressors, the formula should be y ~ ex + en | ex + z
- *data, subset*: specifies the data frame and subset

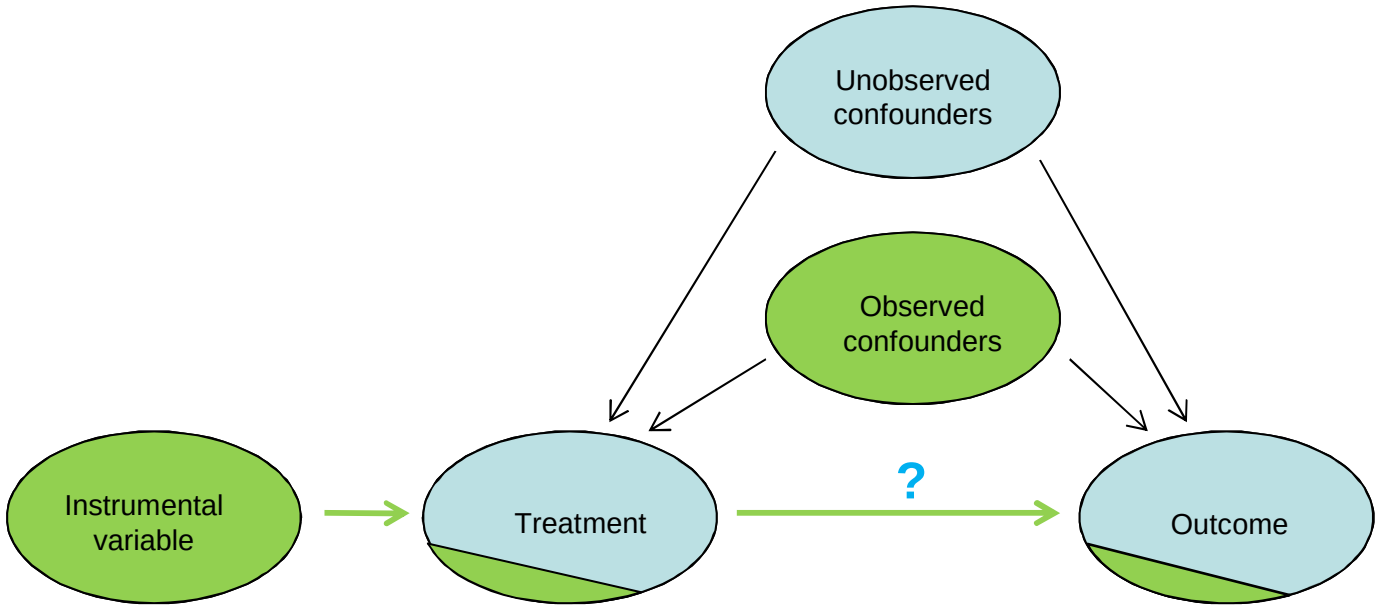
2SLS - BMI

```
> bmi.2sls <- ivreg(bmi_adol~bmi_friend|bmi_parent,
   data=bmi)
> summary(bmi.2sls)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.45315    2.07462  -0.218    0.827
bmi_friend   1.01701    0.08624  11.793   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- conclusion: an increase of 1 kg/m^2 in close friend’s average BMI is associated with 1.02 kg/m^2 increase of an adolescent’s BMI, after controlled for unmeasured confounders

Weak instruments

- correlate with unobserved determinants of the outcome
  - cov(Z, e) ≠ 0
  - no significant advantage over linear regression
- weakly correlate with treatment X (weak relevance)
  - large variance in $\hat{\beta}$
  - potential bias in the estimation of β



Test for weak instruments

- test of instrument relevance
  - F-statistics in the first stage regression
  - in case of one treatment / endogenous regressor: weak instrument if F-statistics < 10
- Durbin-Wu-Hausman (DWH) test for endogeneity of regressor
  - H_0: IV and LS both consistent, but LS is efficient
  - H_1: Only IV is consistent
  - DWH statistics = $(\beta_{IV} - \beta_{LS}) / \sqrt{s^2_{\beta_{IV}} - s^2_{\beta_{LS}}} \sim N(0,1)$
  - reject H_0 if $|DWH| > z_\alpha$
    - use β_{IV} even it is less efficient
  - cannot establish instrument exogeneity

Instrument relevance – BMI example

```
> bmi.lm <- lm(bmi_friend~bmi_parent, data=bmi)
> bmi.lm0 <- lm(bmi_friend~1, data=bmi)
> anova(bmi.lm, bmi.lm0)
Analysis of Variance Table

Model 1: bmi_friend ~ bmi_parent
Model 2: bmi_friend ~ 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    498 2858.9
2    499 4250.6 -1   -1391.7 242.42 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- F statistics = 242.4 > 10
  - high correlation between IV and endogenous regressor

DWH test – BMI example

IVA:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.45315    2.07462  -0.218    0.827
bmi_friend   1.01701    0.08624  11.793   <2e-16 ***
```

LS:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5919     1.1894   1.338    0.181
bmi_friend    0.9318     0.0492  18.940   <2e-16 ***
```

- $|\text{DWH}| = (1.017 - 0.932) / \sqrt{0.086^2 - 0.0492^2}$

  $= 1.20 < 1.96$

- no evidence that IV estimate is better

may use vcov() to extract the variance directly

Mendelian randomization

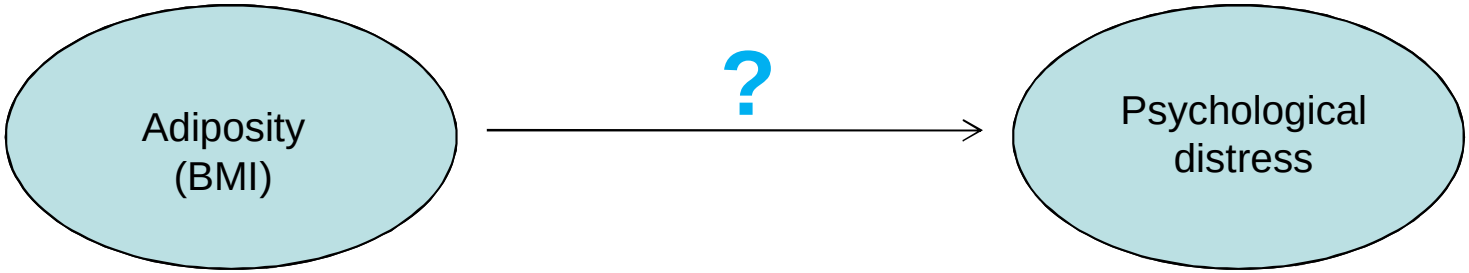
- an application of the instrumental variable analysis
- genetic variant as the IV
  - random allocation of alleles (paternally and maternally inherited)
  - hence not associated with any measured / unmeasured confounders
  - gene is chosen to have known link to a phenotype of interest
- Examples

Outcome	**Gene**	**Variant**	**Risk factor**
BMI	CRP	rs755300	C-reactive protein

(Bochud & Rousson, 2010)

Example: mendelian randomization

- Lawlor et al, J Intern Med, 2011
- interested in the association between adiposity and psychological distress
- n = 53,221



- similar studies are difficult due to multiple source of confounding

Example – adiposity and psychological distress

- instrumental variable
  - Fat mass and obesity associated (FTO) gene (rs9939609)
  - MC4R (rs17782313)
- known gene-trait association:
  - FTO rs9939609 associated with increased BMI
  - MC4R rs17782313 associated with increased waist-hip ratio (WHR)

Example – adiposity and psychological distress

	Multivariable association	IV analysis 1	p-value 1	IV analysis 2	p-value 2
Estimates of odds ratio of outcome per 1 age and sex standardised z-score increase in BMI					
Stress/nervous	0.97 (0.95, 0.99)	0.65 (0.46, 0.91)	0.032	0.36 (0.19, 0.69)	0.005
Not accomplishing	1.11 (1.09, 1.13)	0.69 (0.47, 1.01)	0.016	0.48 (0.24, 0.96)	0.022
Wanting to give up	1.10 (1.06, 1.13)	0.63 (0.34, 1.15)	0.067	0.67 (0.21, 2.09)	0.32
Using antidepressants	1.04 (1.01, 1.08)	0.57 (0.29, 1.10)	0.093	0.49 (0.14, 1.70)	0.24

- inverse association between adiposity and psychological distress was found
- the positive association from conventional analysis may be due to residual confounding

Remarks on IVA / Mendelian randomization

- with a valid IV, a powerful tool to control for unmeasured / unobserved confounder
- assumption of exogeneity of IV not testable from data
  - need to establish validity of the IV
  - need reasoning or knowledge of the empirical context
- using an invalid IV worse than not using any IV

Overview of the course

Exam format

- Date: Mar 29, 2021 (Monday)
- Time: 6:30 – 9:30pm
- Venue: online
- Open book (printed or electronic materials)
- Use your own laptop / computer
  - allow access to R help
  - suggest to pre-install the packages
  - any communication is not allowed
- Choose 2 out of 3 questions:
  - structured questions on analyzing dataset using R
  - correct choice of method and interpretation of the results, clear presentation

R programming

- Open-source, free
- Provide a wide range of packages for different statistical analysis
- More likely to provide most updated techniques
- Need practice
  - Learning process includes committing and correcting errors
- Potential to generate nice graphics
- Relatively easy to reproduce analysis with updated data
- Potential to generate dynamic documents (e.g. ‘knitr’)

Statistical methods covered

- Poisson / negative binomial regression
- Conditional logistic regression
- Propensity score method
- Instrumental variable analysis
- Inverse probability weighting
- Meta analysis
- Model diagnostics
- Assessing confounder, effect modifier and mediator
- Random number generation

Generalized linear regression (GLM)

- Largely determined by the dependent variable
- Linear, logistic, poisson, negative binomial regression models all belong to GLM
- Interpretation of the estimates depending on the link function
  - e.g. mean difference (identity link), relative risk (log link), odds ratio (logit link)

Main components of GLM:

- Systematic component:
  - Link function *g* which links the outcome variable to the linear predictor
  - $g(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$
- Random component:
  - Specifying variance to mean relation

General strategy for analysis

- Exploratory analysis to inform best model
  - Identify potential outliers
- Specify the method before the analysis
  - To reduce experimenter bias
- Assess fitness / assumption of the models
  - Model diagnostics
  - Identification of outliers or influential observations
- Interpret results and draw conclusion
  - Sensitivity analysis for an assessment

General strategy for addressing confounding, effect modifier and mediator

- Based on a priori subject-matter knowledge of the causal structure, include appropriate confounder(s) and effect modifiers in the model
- Empirical model selection from the data
  - Confounding:
    - Model selection by AIC (backward selection is preferred to reduce the number of multiple testings)
    - Change in estimate of the exposure / intervention
    - ‘Standard rules’ for confounding
  - Effect modifier:
    - Test of interactions
  - Mediator:
    - Baron and Kenny criteria
    - Sobel test

Further reading

- Khandker SR, Koolwal GB, Samad HA. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington: The World Bank. 2010.

- Kane RL, Radosevich DM. *Conducting Health Outcomes Research.* Jones & Bartlett Learning. 2011.