Dr. Christian Czymara

FORSCHUNGSPRAKTIKUM I UND II: LÄNGSSCHNITTDATENANALYSE IN R

Linear regression with cross-sectional and longitudinal data
session iii

AGENDA

- A run through the OLS estimator
- … and its assumptions
- … and while panel data may violate some
- Specification of linear model

THE ORDINARY LEAST SQUARES ESTIMATOR

LINEAR REGRESSION

- We model y as a function of other variable(s) x
- With real-world data, y is never a perfect function of x: ε_i
- $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$
- $i = 1, \ldots, n$ units; k variables
- (Simple cross-sectional model)

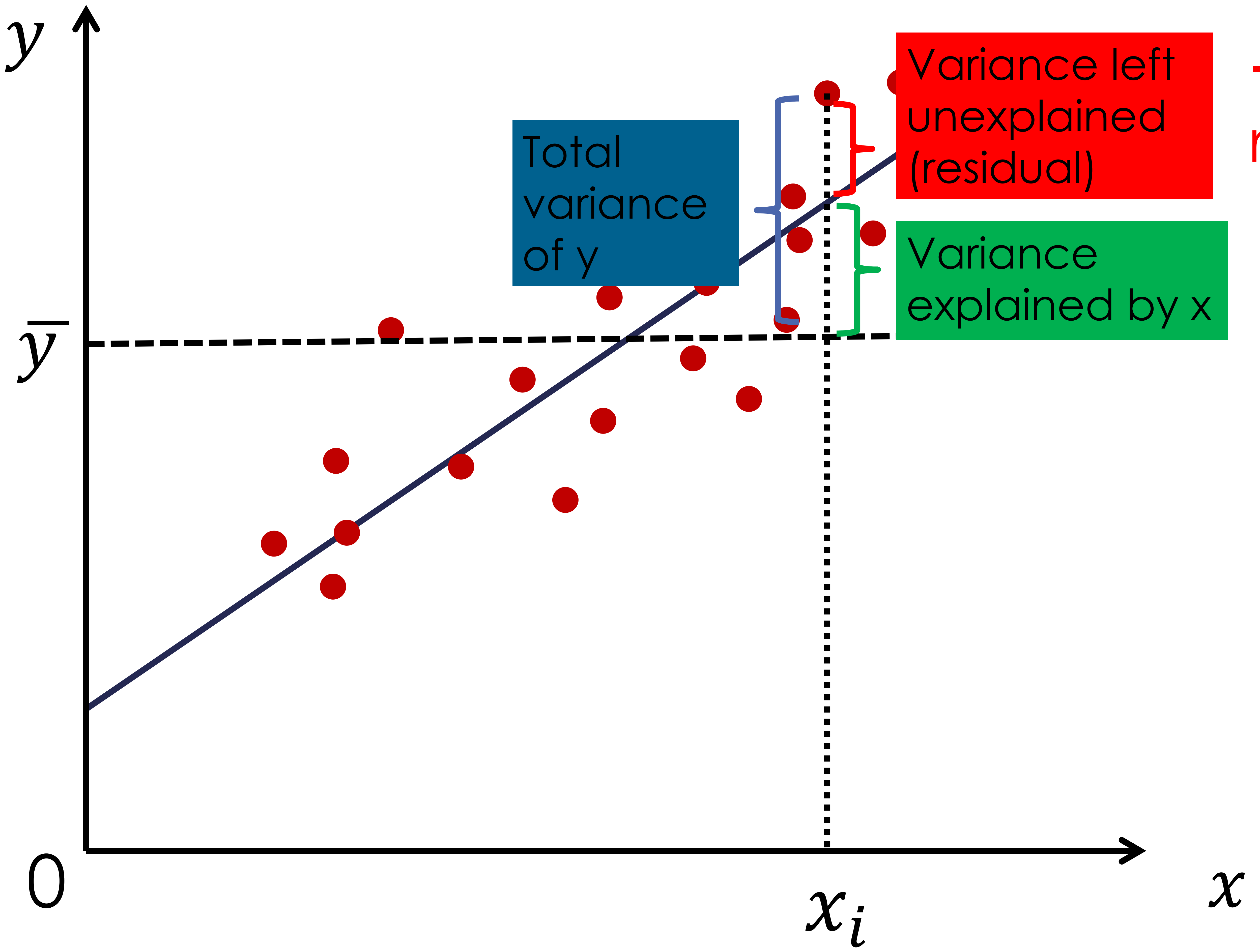
LINEAR REGRESSION

- $y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}_{\text{Systematic part}} + \underbrace{\varepsilon_i}_{\text{Stochastic part}}$
- $\beta_0, \beta_1, \beta_k$ are the parameters which have to be estimated

ORDINARY LEAST SQUARES (OLS) ESTIMATOR

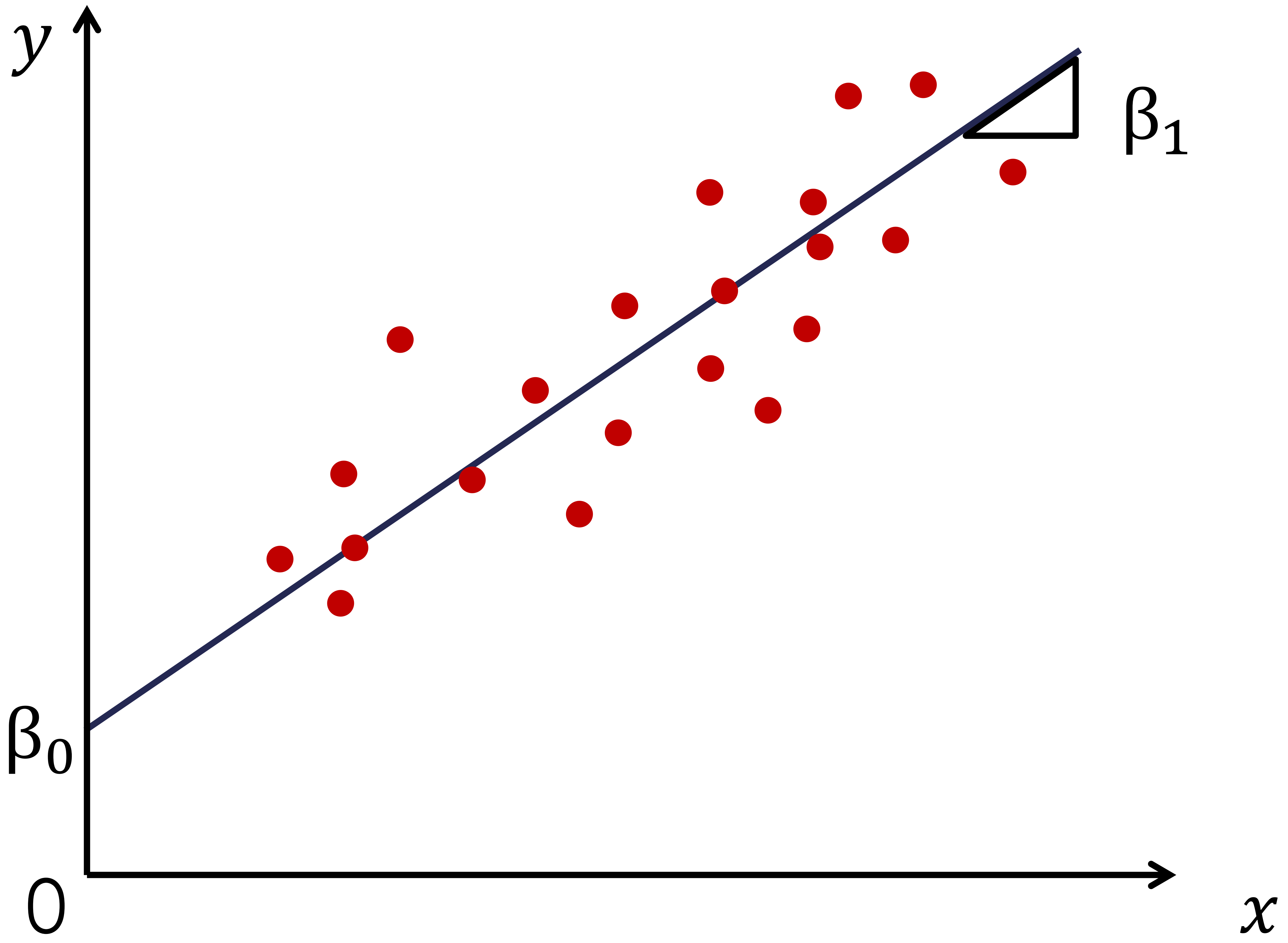
- Which values for unknown parameters $\beta_0$and β_1?
- OLS minimizes (squared) differences between the observed and the predicted values
- Gauss–Markov theorem: OLS is *BLUE*, given certain assumptions
  - **B**est: most efficient (lowest standard errors)
  - **L**inear
  - **U**nbiased: estimated parameters identical with “true” parameters
  - **E**stimator

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$



→(squared and) minimized by OLS

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$



OLS ASSUMPTIONS

ASSUMPTIONS

1. (For inference statistics) random sample
2. Model linear in its parameters $\beta_0, \beta_1, \dots, \beta_k$
3. x neither constant nor linear combinations of other x

ASSUMPTIONS

5. Error not correlated with x (strict exogeneity): $E(\varepsilon_i|x_{1i} \dots x_{ki}) = 0$
6. Error has constant variance across all x (homoscedasticity): $var(\varepsilon_i|x_{1i} \dots x_{ki}) = \sigma^2$
7. Error uncorrelated: $corr(\varepsilon_i, \varepsilon_j|x_{1i} \dots x_{ki}) = 0$
8. Error normally distributed with mean 0 and variance σ^2: $\varepsilon_i \sim Normal(0, \sigma^2)$

EXOGENEITY ASSUMPTION

EXOGENEITY ASSUMPTION

- Assumption 5 means that the error term is independent from x

→Model includes all relevant variables and has correct functional form (*correctly specified*)

→Measurement error is random (does not depend on x)

- Ensures unbiased estimates
- Crucial assumption for estimating "true" (i.e. unbiased) parameters

MODEL SPECIFICATION

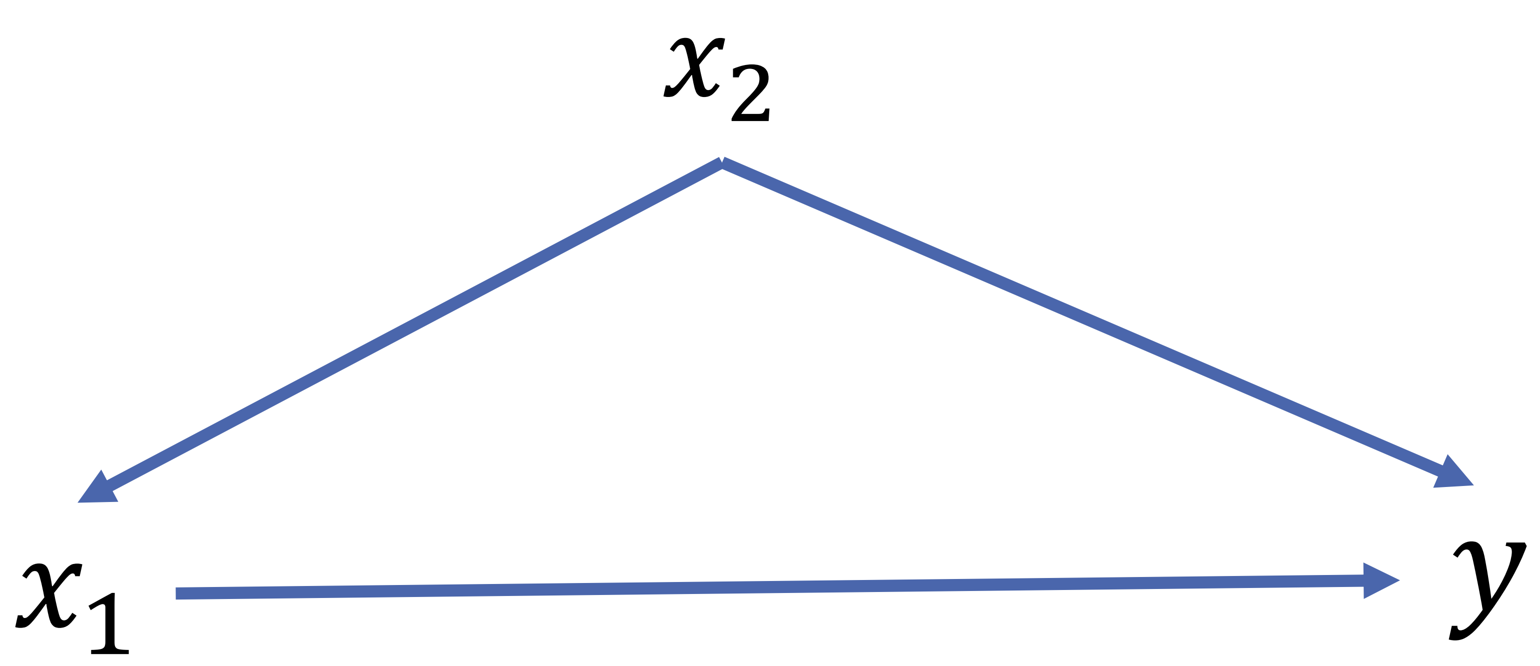
- A correctly specified model includes all relevant x
- Which x are relevant?
- Those that are conceptually or theoretically (!) cause both y and the x of interest
- Not including (omitting) relevant x_2 in a regression model will lead to a biased estimate of β_1
- This is because β_1 in this case carries part of the effect of β_2 on y
- *Avoiding bias is the main point of all statistical analyses!*

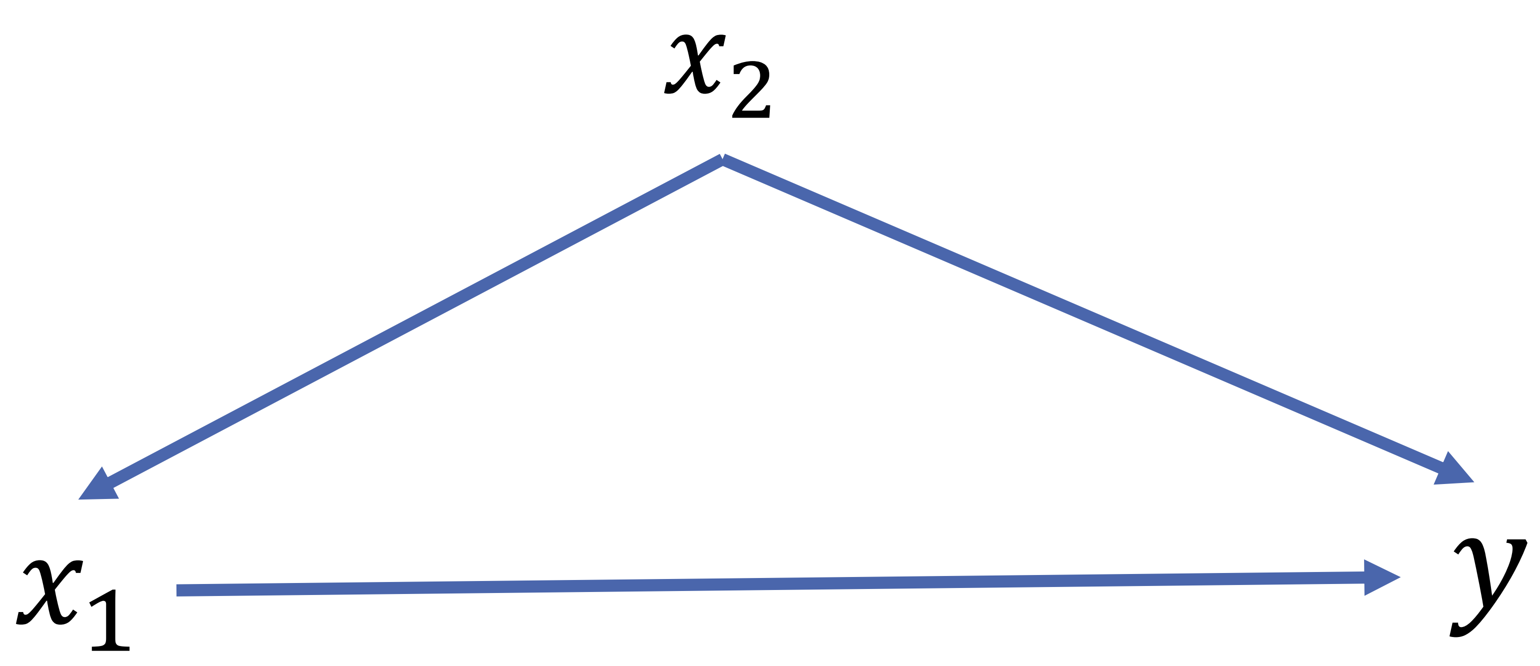
OMITTED VARIABLE BIAS

- True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$
- Unbiased estimation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- New situation: x_2 unobserved
- Biased estimation: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$
- Omitted variable bias: $Bias(\tilde{\beta}_1) = E(\tilde{\beta}_1) - \beta_1 = \beta_2 \frac{\widehat{cov}(x_1,x_2)}{\widehat{Var}(x_1)}$
- Hence no bias if
  - $\beta_2 = 0$
  - or $\frac{\widehat{cov}(x_1,x_2)}{\widehat{Var}(x_1)} = 0$

OMITTED VARIABLE BIAS

- $\beta_2 = 0$
- $\frac{\widehat{cov}(x_1,x_2)}{\widehat{Var}(x_1)} = 0$





SPECIFYING MODELS IN LINEAR REGRESSION

STATISTICALLY CONTROLLING

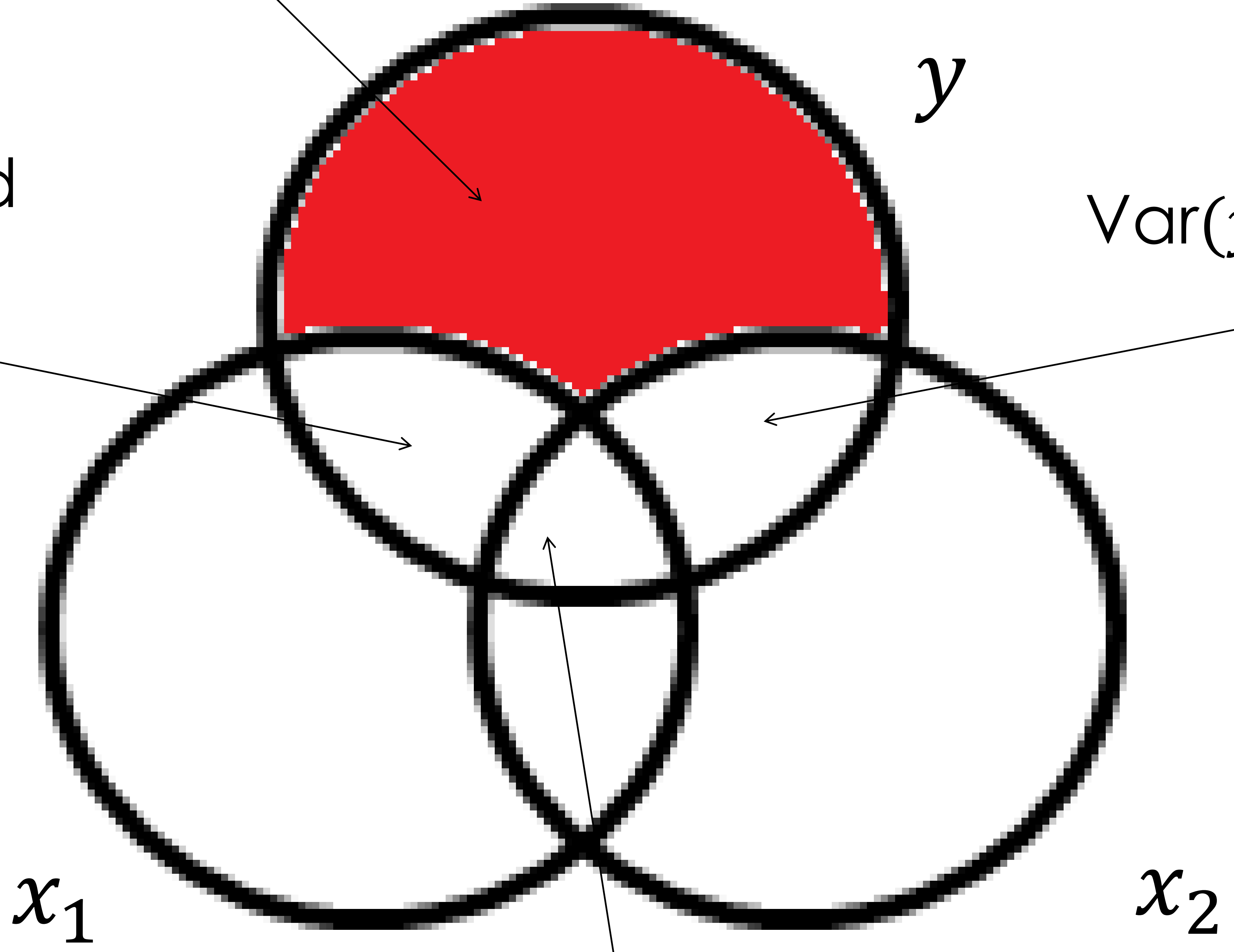
- Confounder x_2 leads to a spurious correlation between x_1 and y
- The most common way to account for this in cross-sectional quantitative studies is statistical controlling
- This means *netting out* the effect of x_2 / *adjusting* for x_2
- The result is the effect of x_1 on y which does not depend on x_2
- The motivation behind this is to remove other "common causes" of x and y

VARIANCE COMPONENTS OF TRIVARIATE REGRESSION

Var($\hat{y}$) explained neither by x_1 nor x_2

Var($\hat{y}$) explained by (only) x_1

Var($\hat{y}$) explained by (only) x_2

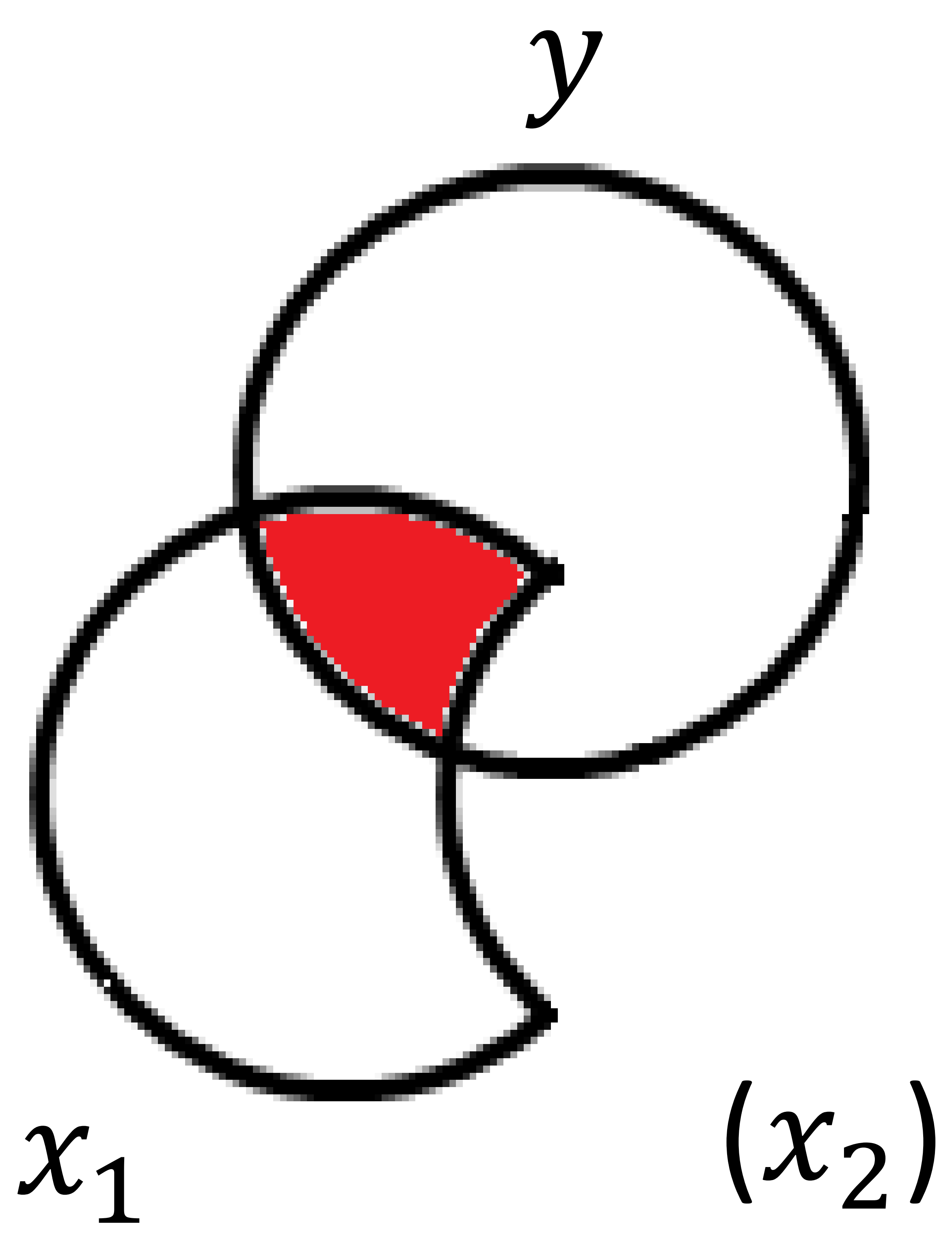


Var($\hat{y}$) explained by x_1 as well as x_2

By definition, residuals will always be uncorrelated with explanatory variables in OLS

STATISTICALLY CONTROLLING

- The effect of x_1 which does not depend on x_2
  - The effect of x_1 on y *controlling* for x_2 (trivariate regression)
  - Interpretation: "a one unit increase in x_1 implies a β_1 increase in y controlling for / net of x_2"

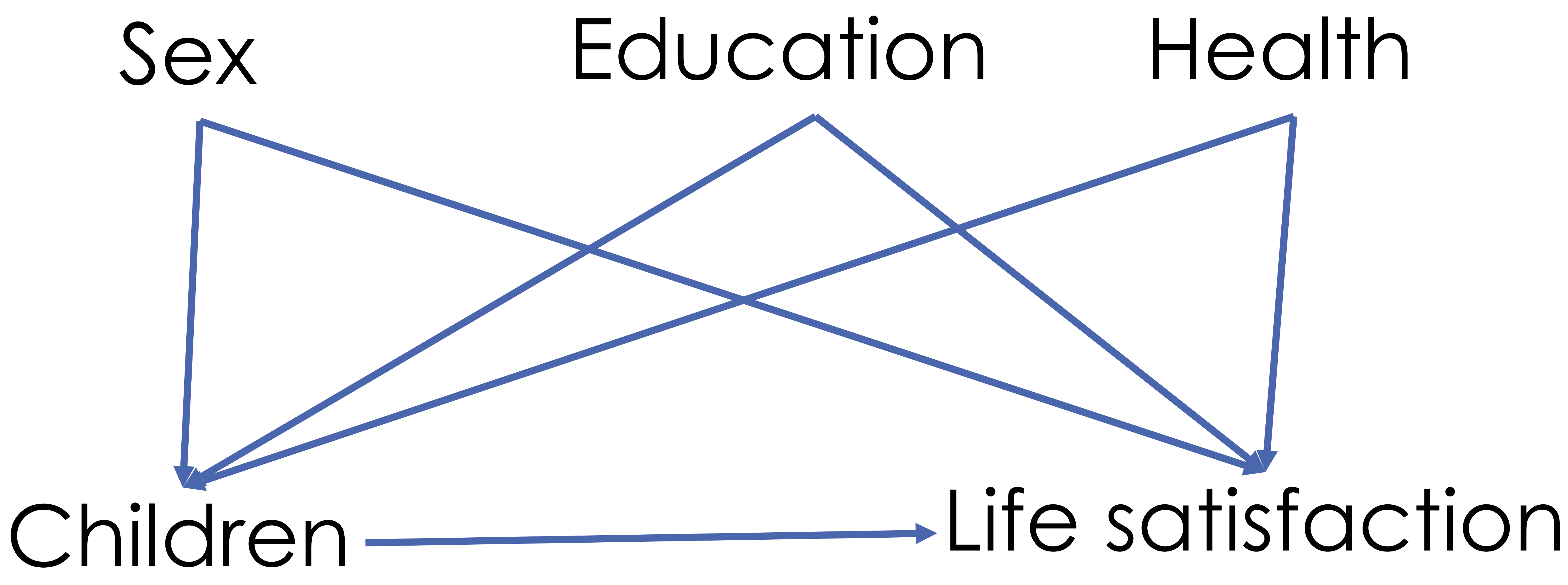


EXAMPLE FROM TUTORIAL

EFFECT OF CHILDREN ON LIFE SATISFACTION

- "Do children make happy? We are interested in the impact of having children (x: no_kids) on life satisfaction (y: satisf_org)."
- Control variables: education, health and sex
- With the proposed model, is the effect causal?

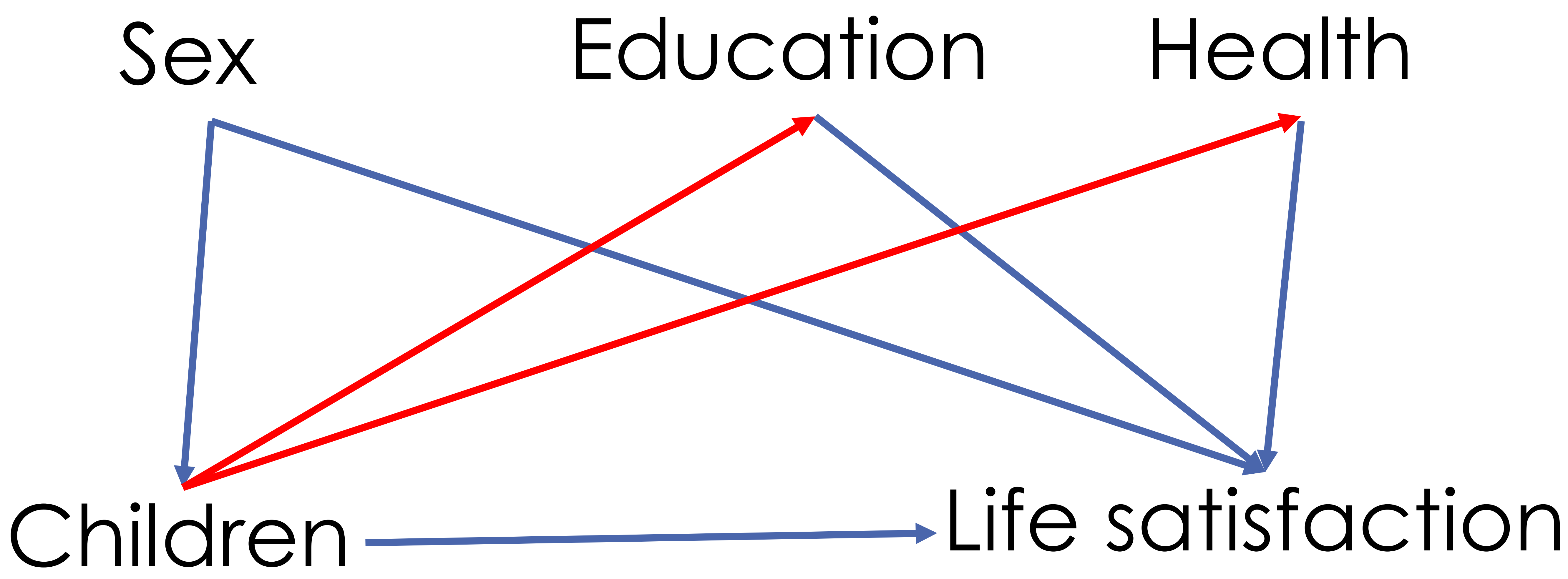
ASSUMED MODEL



- All control variables are confounders ✓
- No conditioning on colliders or mediators ✓
- No unobservables ✓

→Causal effect

ALTERNATIVE SCENARIO 1



- Now education and health are post-treatment, making them mediators

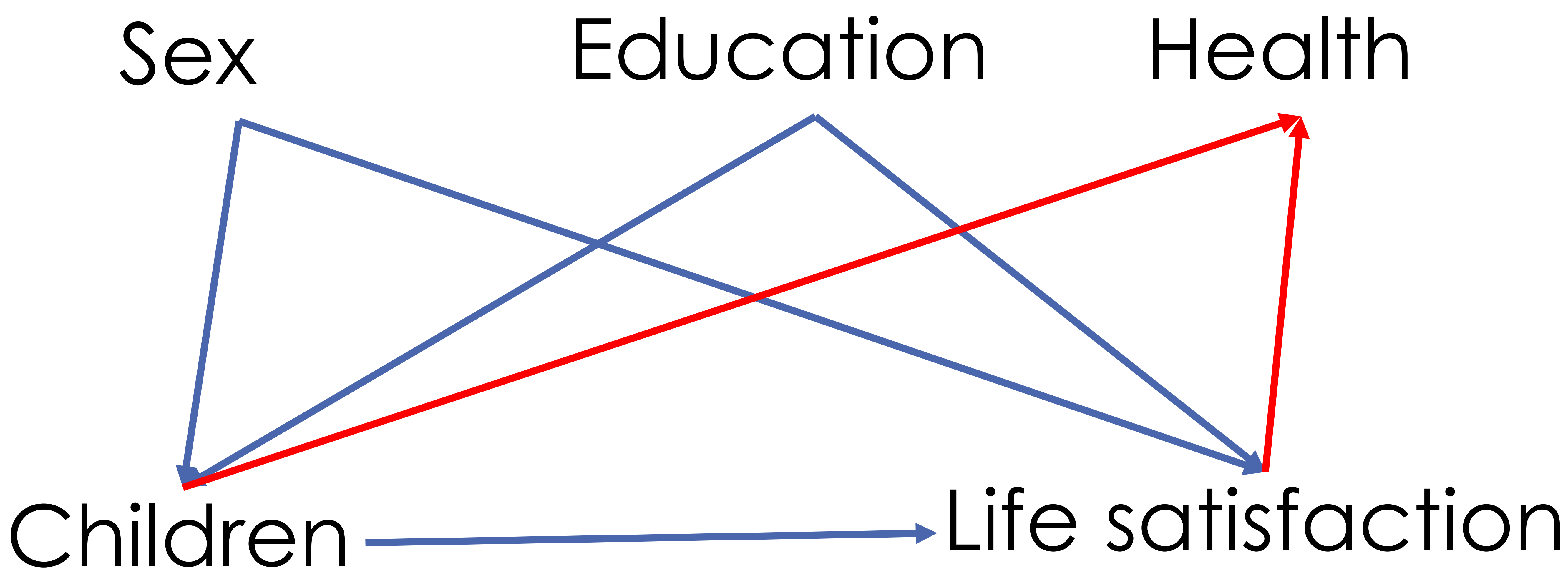
→Direct causal path from children to life satisfaction

→Indirect causal path flowing from children through education and through health

→Total causal effect of children: direct + indirect effects

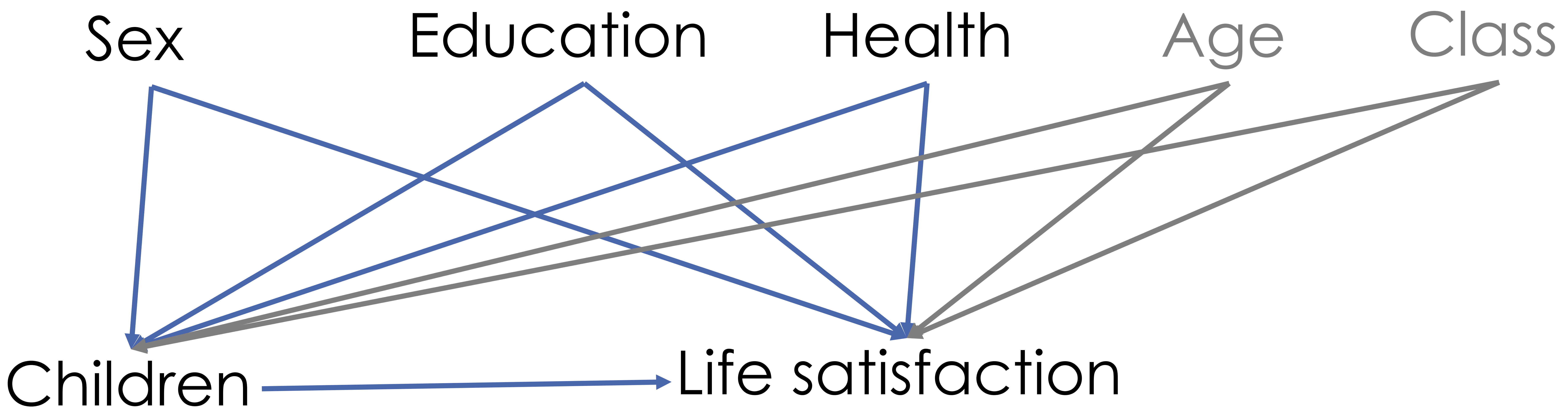
→Controlling education and health would lead to *overcontrol bias*

ALTERNATIVE SCENARIO 2



- Now health is post-treatment and post-outcome, making it a collider
- The causal effect of children can be estimated by controlling sex and education – but *not* health
- Controlling health would open the non-causal path C→H←LS

ALTERNATIVE SCENARIO 3



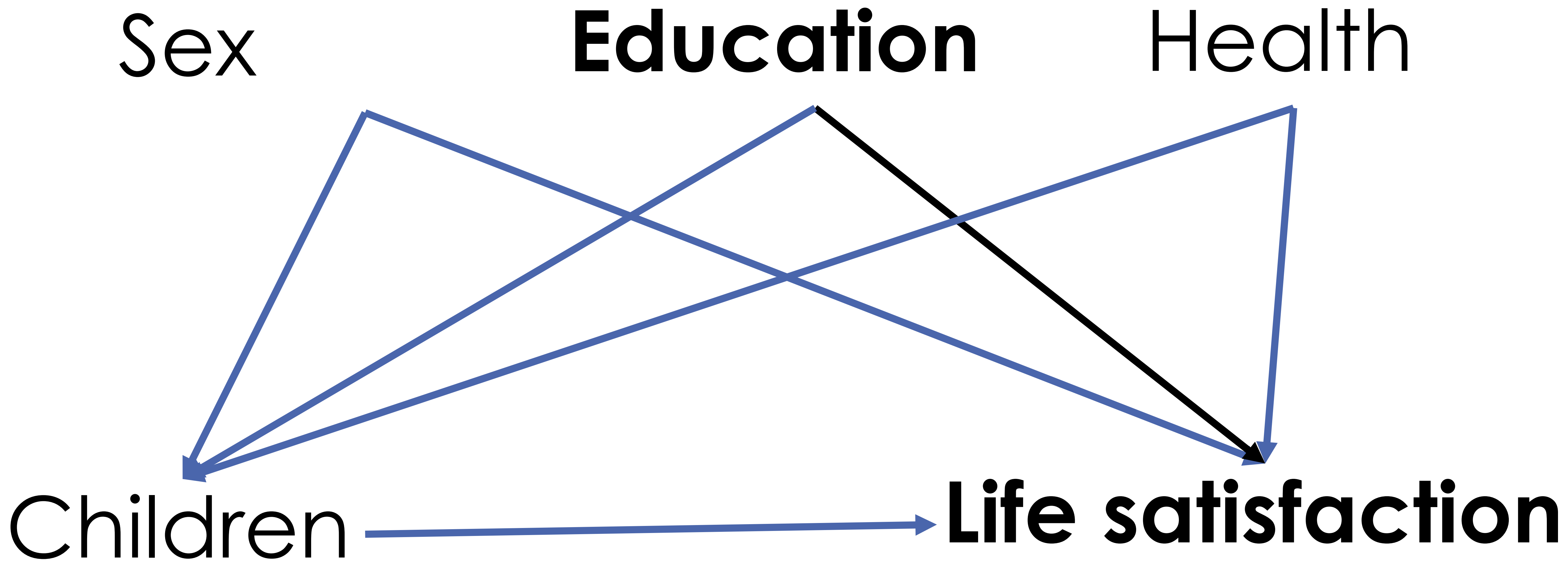
- Age and class are confounders, but are not observed
- Causal effect of children on life satisfaction not estimable
- No easy solution (with common cross-sectional models)
- One of the most common critiques of empirical studies (*"You should control for …"*)

WHICH MODEL IS CORRECT?

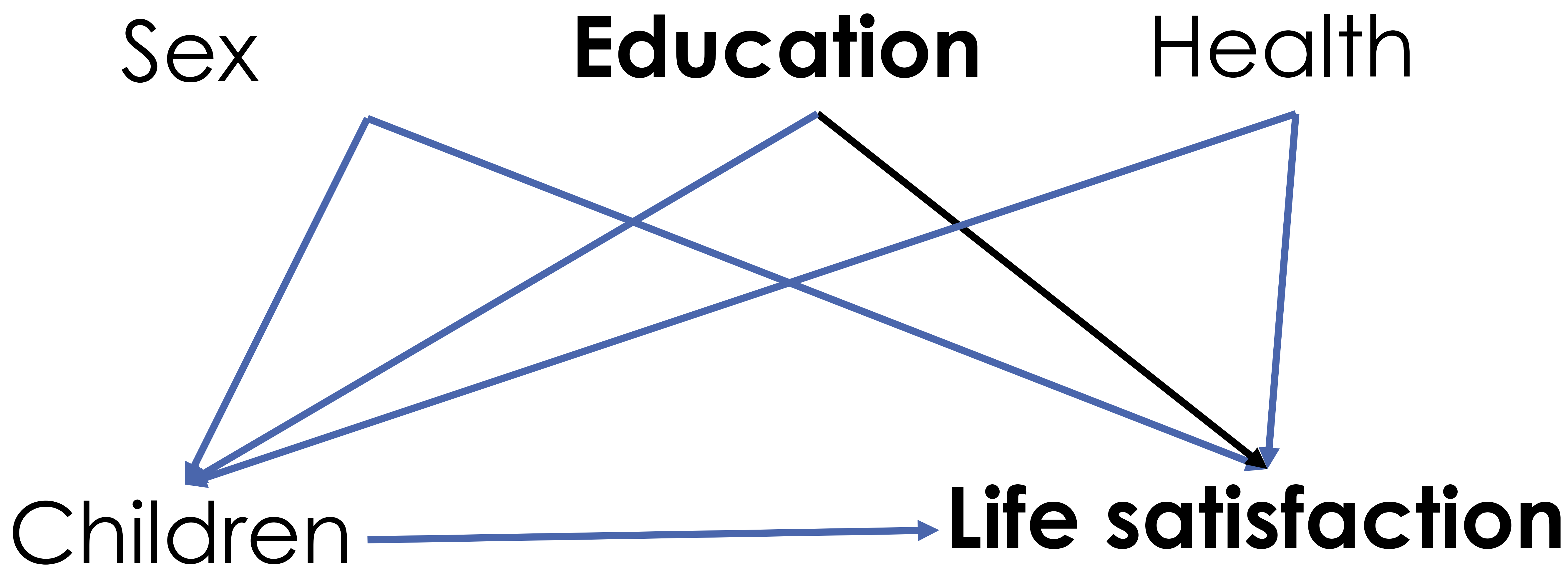
- You tell me
- With the initial model, we assume that neither age nor class (or anything else) affect children and life satisfaction
- If one buys this assumption, we have estimated a causal effect
- Given existing research, however, this is a strong assumption that is hard to defend
- Scenario 3 more likely to be convincing

SCENARIO 4

- Same data, same DAG but we are actually not interested in the effect of children but in the effect of education



ASSUMED MODEL



- If the DAG we assumed before is correct…
- *Children* is a mediator on the causal path from education to life satisfaction and should *not* be controlled
- Even worse, *Children* is a collider blocking the non-causal paths *Education*→*Children*←*Sex*→*Life satisfaction* and *Education*→*Children*←*Health*→*Life* satisfaction and, thus, *must not* be controlled (unless *Sex* and *Health* are controlled as well)
- Since we assume no association between either *Sex* or *Health* and *Education*, controlling both would neither induce nor remove bias (as long as *Children* is not controlled)
- *Different research question require different modelling strategies*

LIMITS OF STATISTICAL CONTROLLING

- Within the standard linear regression framework, one can only control variables that are in the data
- Many things, however, are not observed
- Especially when working with secondary data
- Some techniques for longitudinal data analysis can tackle this problem
- Tbc.

ASSUMPTION OF UNCORRELATED ERRORS

PANEL DATA

- Panel data means the same individuals are observed over time (interviewed repeatedly)
- Person A is interviewed in time point 1 and in time point 2

→For each variable x, there are two data points for person A (x_{A1} and x_{A2})

→Same for person B (x_{B1} and x_{B2})

- In contrast to cross-sectional data analysis, the units of analysis are *not individuals*, but individual interviews!
- … because each individual is in the data multiple times (as often as she was interviewed)

OLS WITH PANEL DATA

- It is reasonable to assume that data points are not independent
- x_{A1} is likely to have more in common with x_{A2} than with x_{B1} (or x_{B2})
- For example, income of person A in 2015 is not independent from her income in 2014 (chances are high it's actually the same)
- Put differently, observations (interviews) cluster within individuals
- … which separates them from interviews of other individuals

→Likely a violation of the assumption of independent errors

ASSUMPTION OF INDEPENDENT ERRORS

- Violation of the assumption of independent errors means observations are not statistically independent
- Sample size is inflated
- There is less information in the data than it seems (because it is partly correlated)

→More data leads to lower standard errors (erroneously, in this case)

- Underestimated standard errors lead to wrong p-values and confidence intervals
- Results look "too significant"
- Should be modelled

LITERATURE

- Cinelli, Forney & Pearl (forthcoming). A crash course in good and bad controls. Sociological Methods and Research.