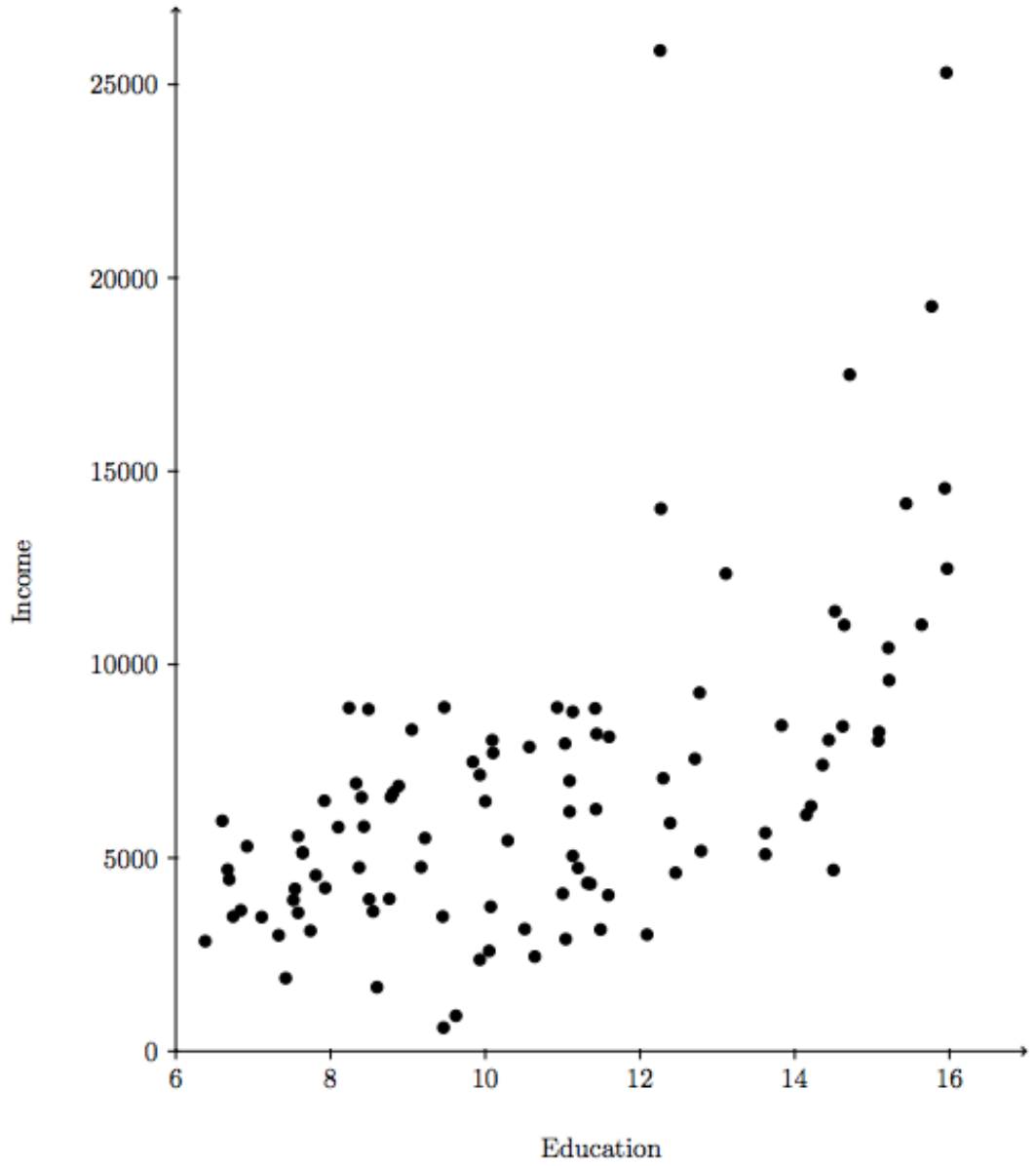
Mendelian randomisation

Nikolai Czajkowski

Education on income

- Does higher levels of education lead to higher income?



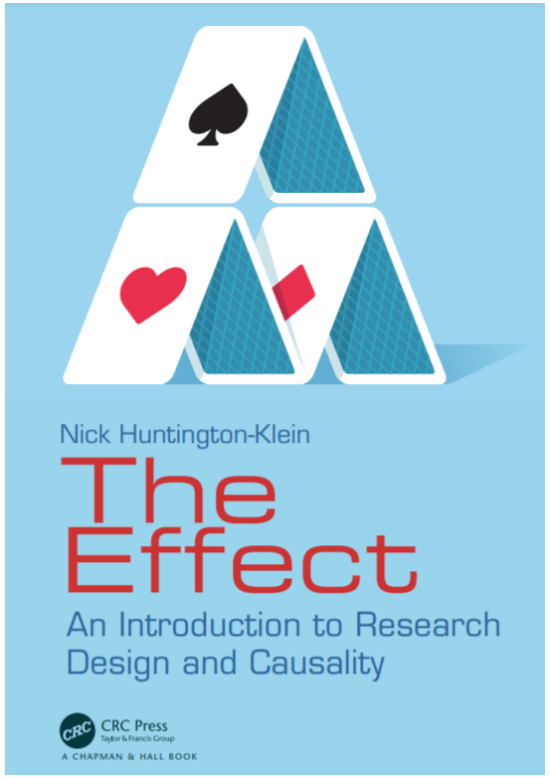
Today

- Focus to causal environmental influences
- Directed Acyclic Graphs (DAG’s)
- Instrumental variable approach to causal inference
- Mendelian randomisation - Genes as "instruments" in natural experiments

Causality

Causal inference

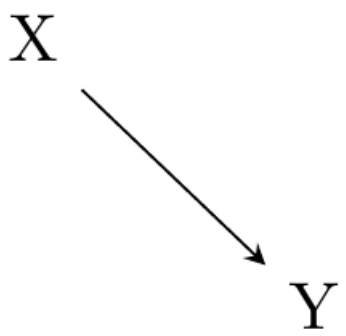
- Is our ultimate aim in science to establish *associations*?
  - "Aspirin is linked to headaches"
- For non-causal relationships, the reverse also holds
  - "Headaches are linked to aspirin."



https://theeffectbook.net/

We can say that X ***causes*** Y if, were we to intervene and *change* the value of X, then the distribution of Y would *also* change as a result.

- *or at least the probability of Y changes*



Endogeneity

- **Endogenous influences** come from within the system, **exogenous influences** come from outside.

- **Endogeneity problem** when a non-causal influence give rise to a correlation between a treatment and an outcome
  - More formally; when the explanatory variable is correlated with the error in the regression model.

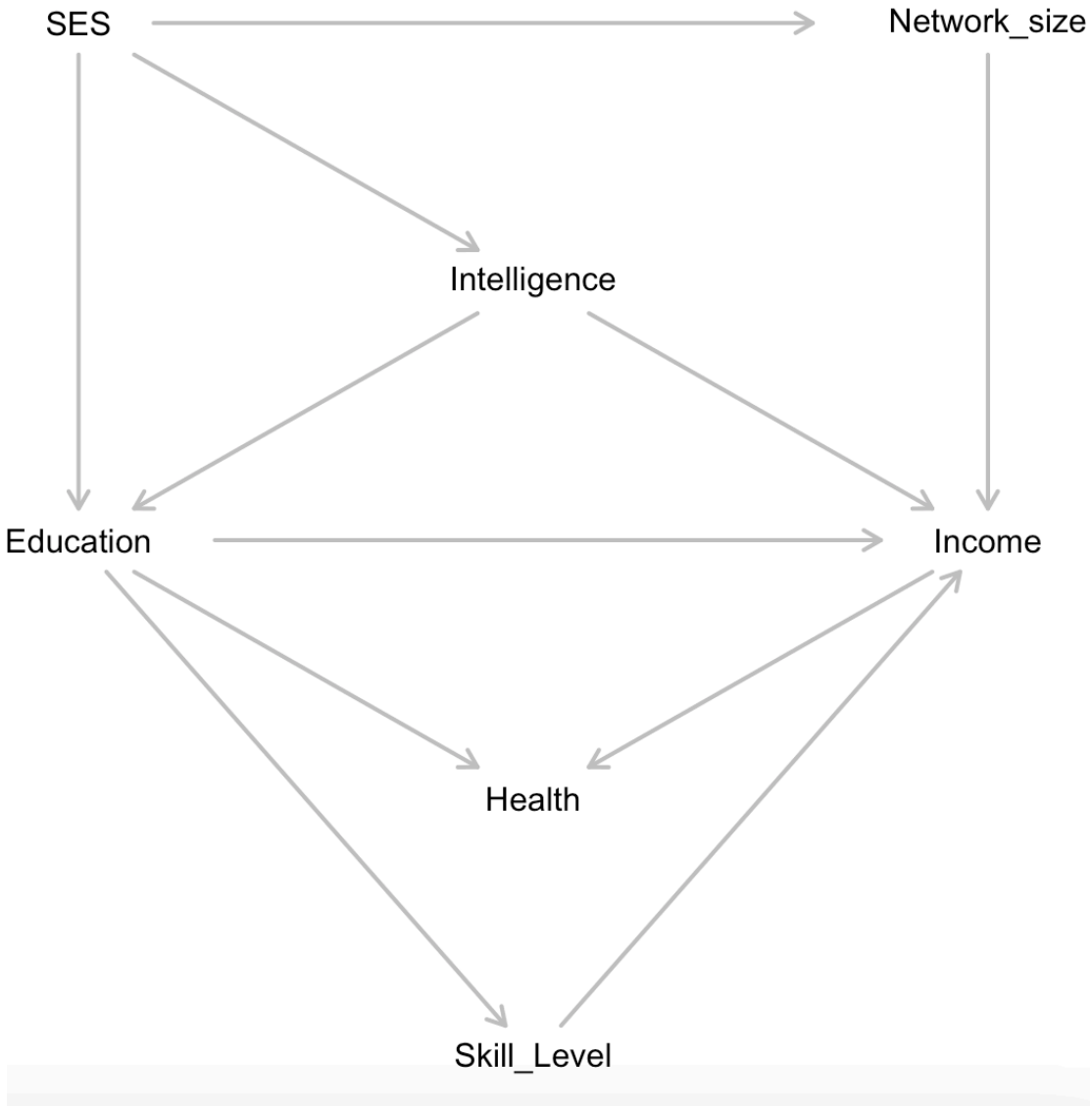
- Common sources of endogeneity
  - Reverse Causation
  - Omitted Variable Bias

- Endogenous influences come from within the system, exogenous influences come from outside.
- Endogeneity problem when a non-causal influence give rise to a correlation between a treatment and an outcome
  - Common sources of endogeneity
  - Reverse Causation
  - Omitted Variable Bias

Example: Vitamin Use and Birth Defects

- Assume you are interest in whether higher education causally leads to higher income.

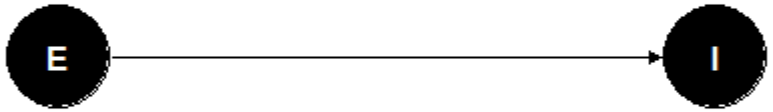
- What should we control for?
  - SES, self-reported health, skill-level, intelligence, network size?



DAG’s – a tool for causal inference

Causal diagrams

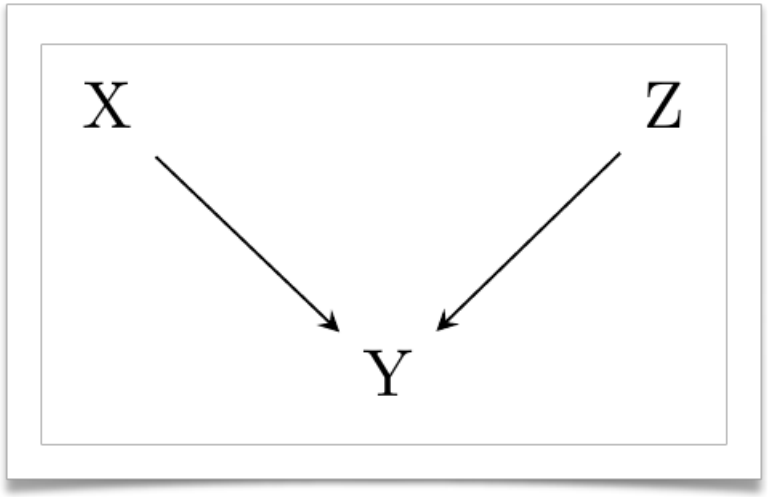
- A **causal diagram** is a graphical representation of a data generating process (DGP).
- Developed in 90's by Judea Pearl for causal inference by computer.



- Contains only two things
  - Each variable is a *node*
  - Each causal relationship is an *edge (arrow)*

Causal diagrams are heuristic tools

- The exact mathematical relationship is not captured by a the graphs

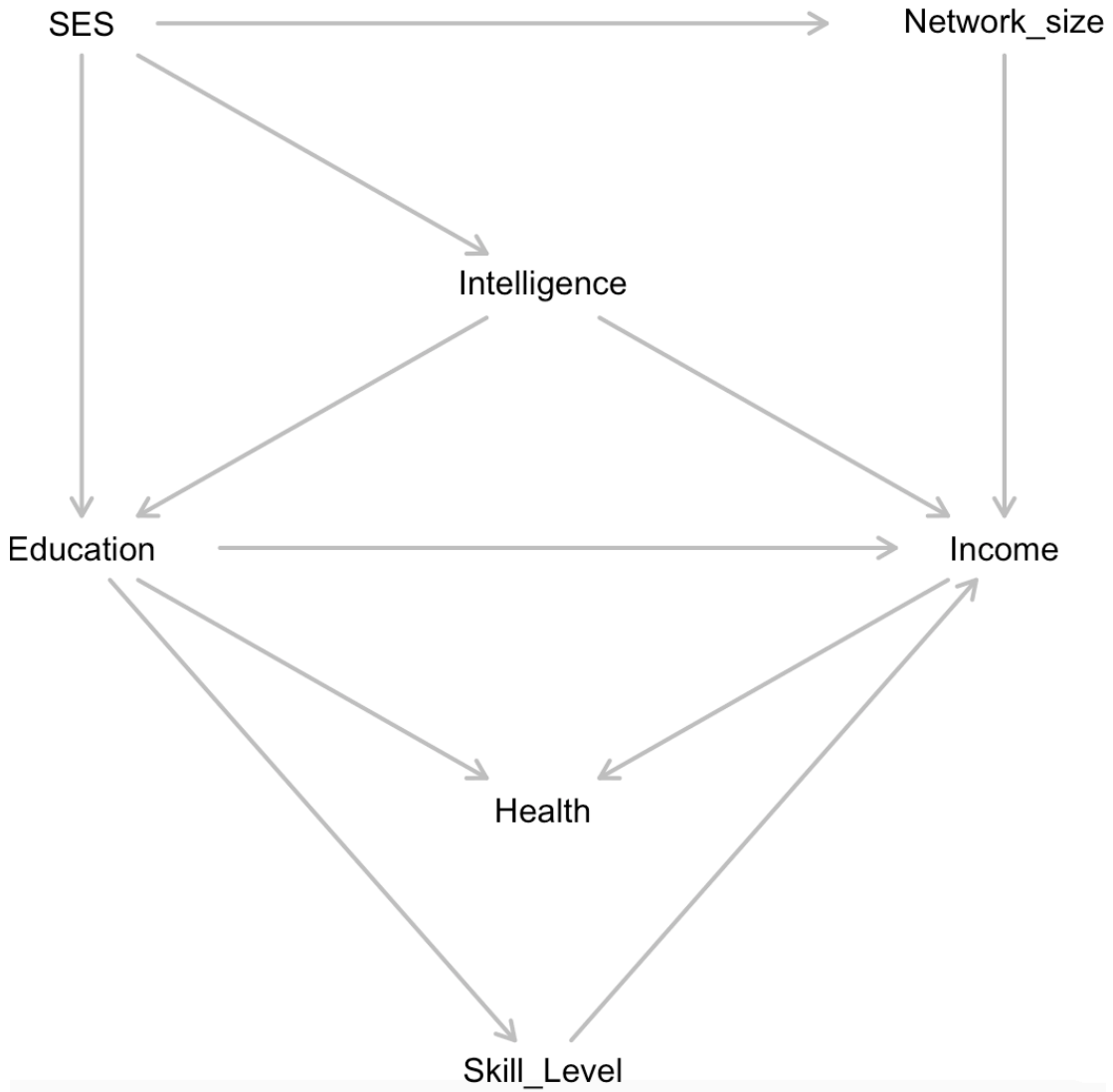


1. $Y = .2X + .3Z$
2. $Y = 4X + 3Z + 2Z^2$
3. $Y = 1.5X + 5Z + 3XZ$
4. $Y = 2X + 3XZ$

Dircted acyclic graphs (DAG's)

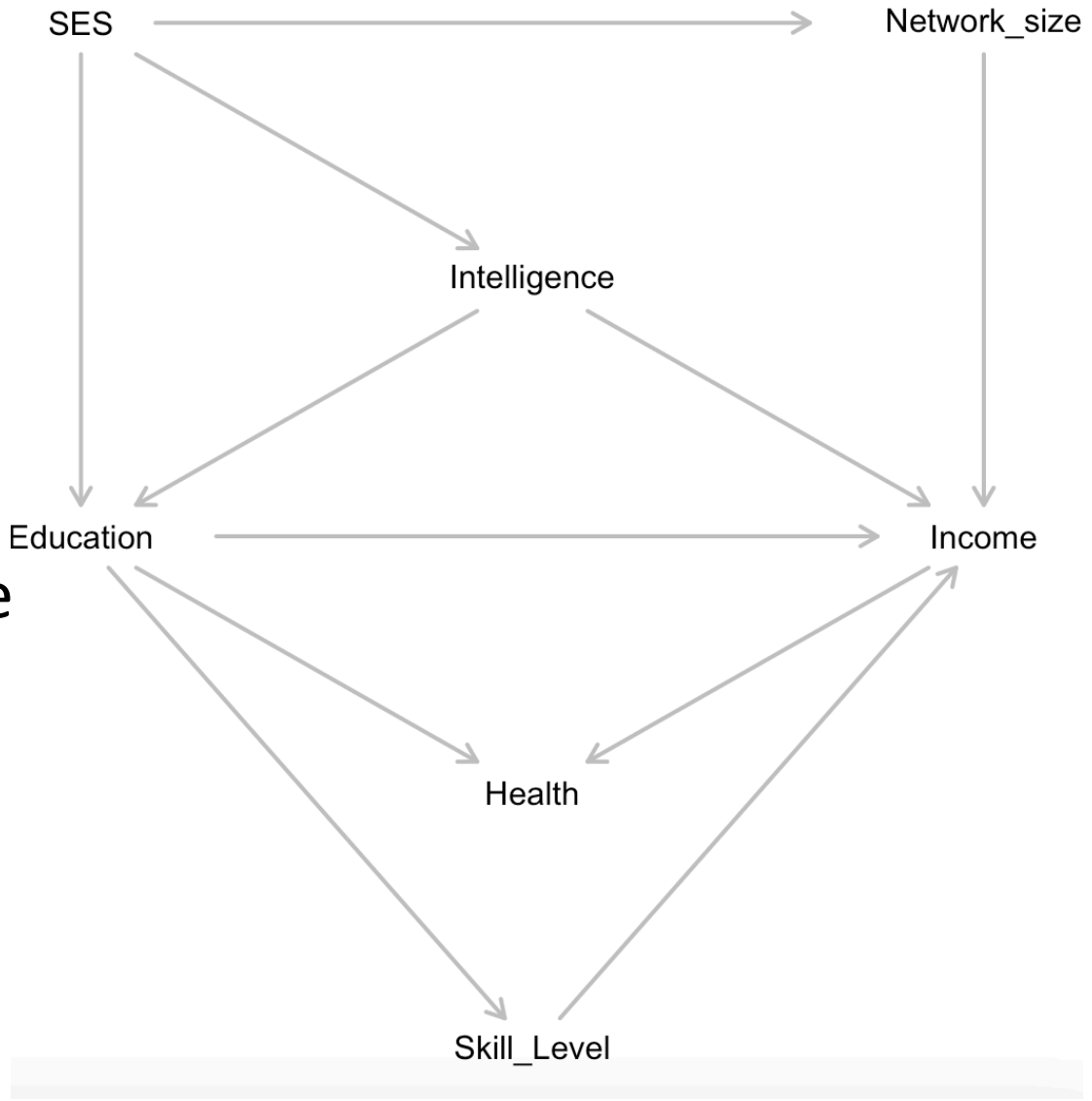
DAG stands for:

- D: directed since each edge is a single-headed arrow
- A: acyclic: it contains no cycles (no variable causes itself)
- G: graph



DAG paths

- **Paths:** sequences of edges connecting two nodes.
- **Direct Path**: This is a path where the nodes are connected directly by a single edge, representing a direct causal effect.
- **Indirect Path**: An indirect path involves one or more intermediate nodes.

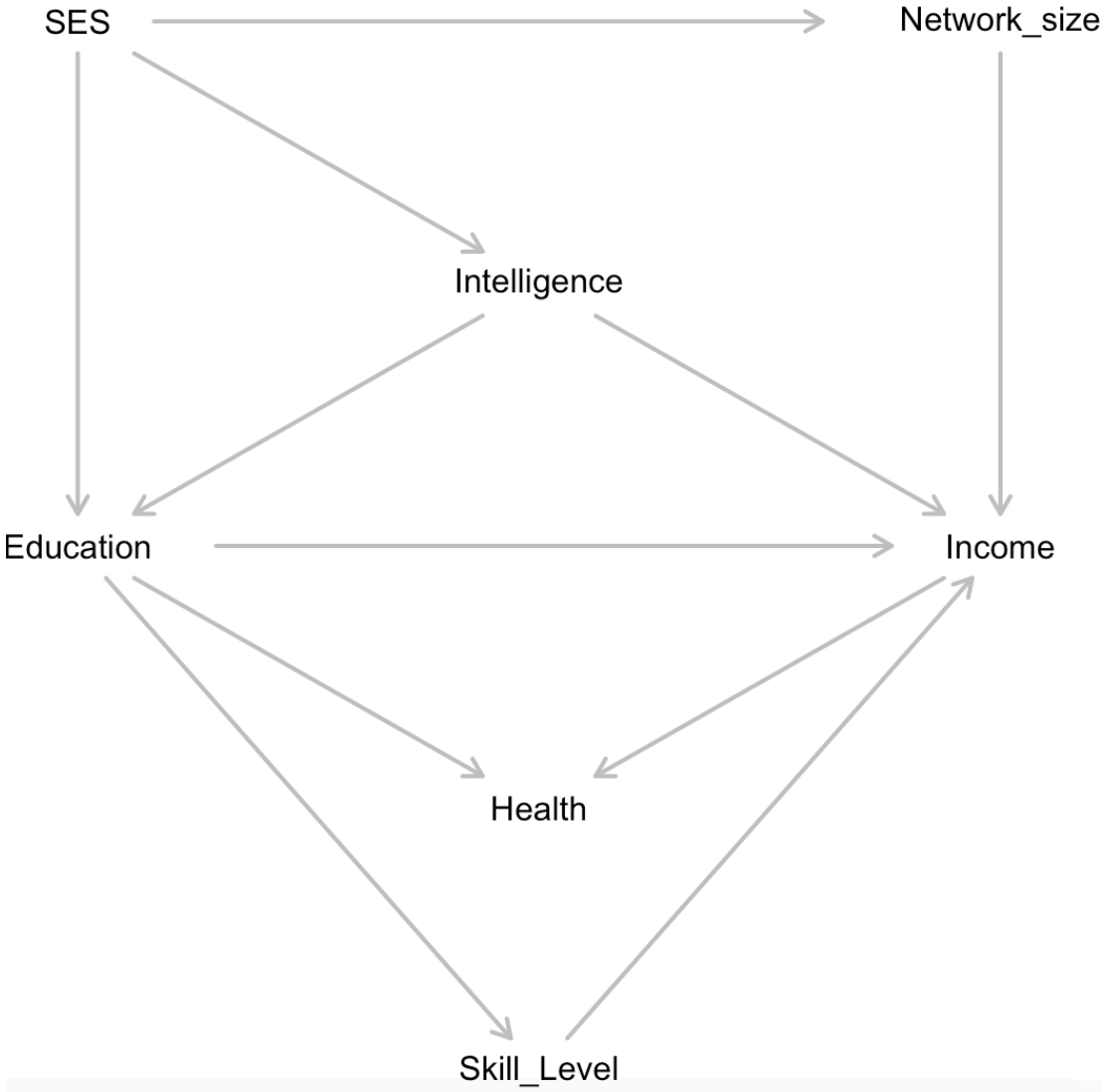


- What are the paths in the DAG?

DAG paths

- What are the paths in the DAG?

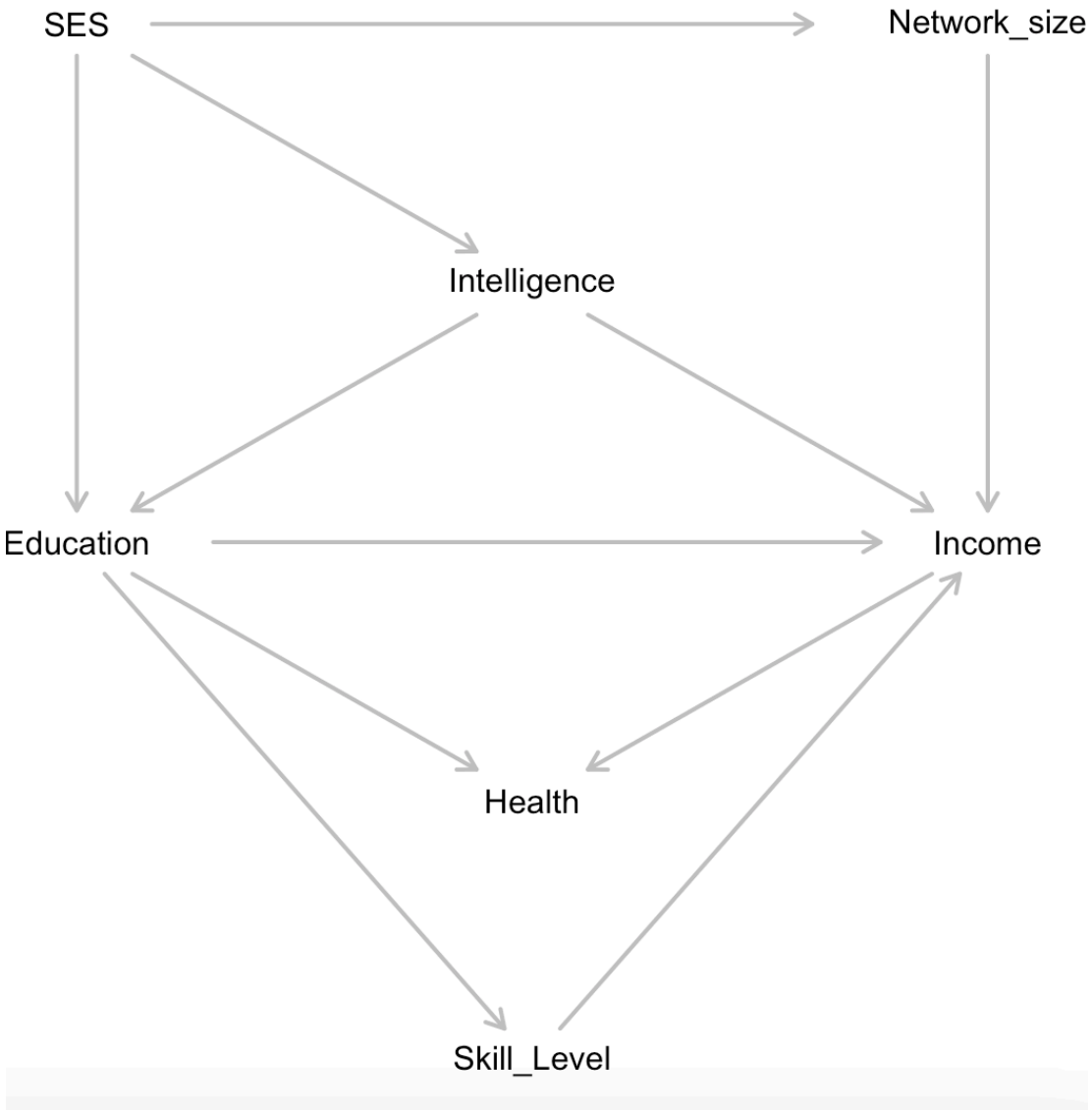
- `Education -> Health <- Income`
- `Education -> Income`
- `Education <- Intelligence -> Income`
- `Education <- Intelligence <- SES -> Network_size -> Income`
- `Education <- SES -> Intelligence -> Income`
- `Education <- SES -> Network_size -> Income`
- `Education -> Skill_Level -> Income`
-



Front and back doors

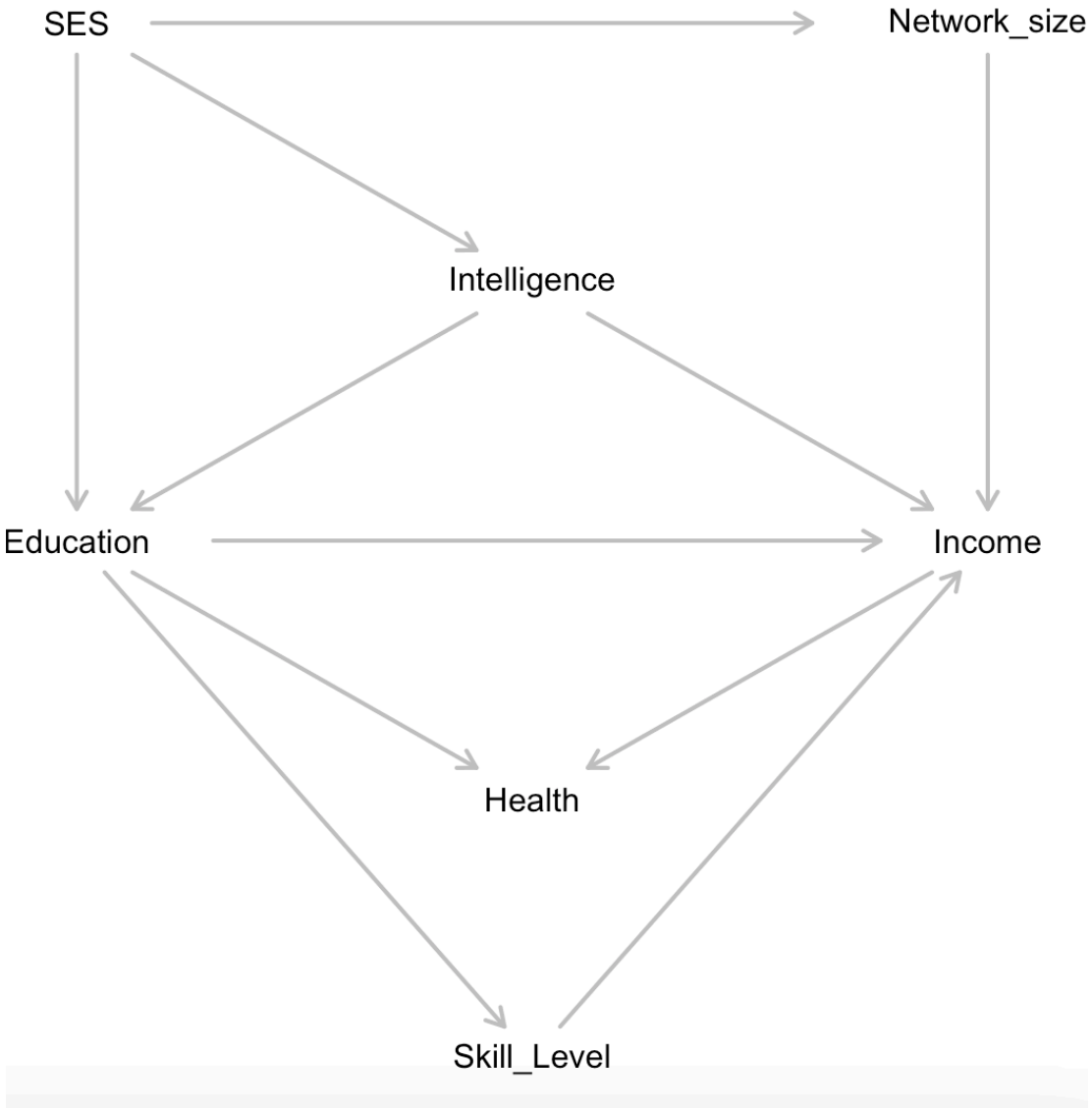
- **Front door paths:** Causal paths where all the arrows point away from exposure.

- **Backdoor paths:** Causal paths where at least one arrow, somewhere along the line, is pointing back towards the exposure (education) variable.



Confounders

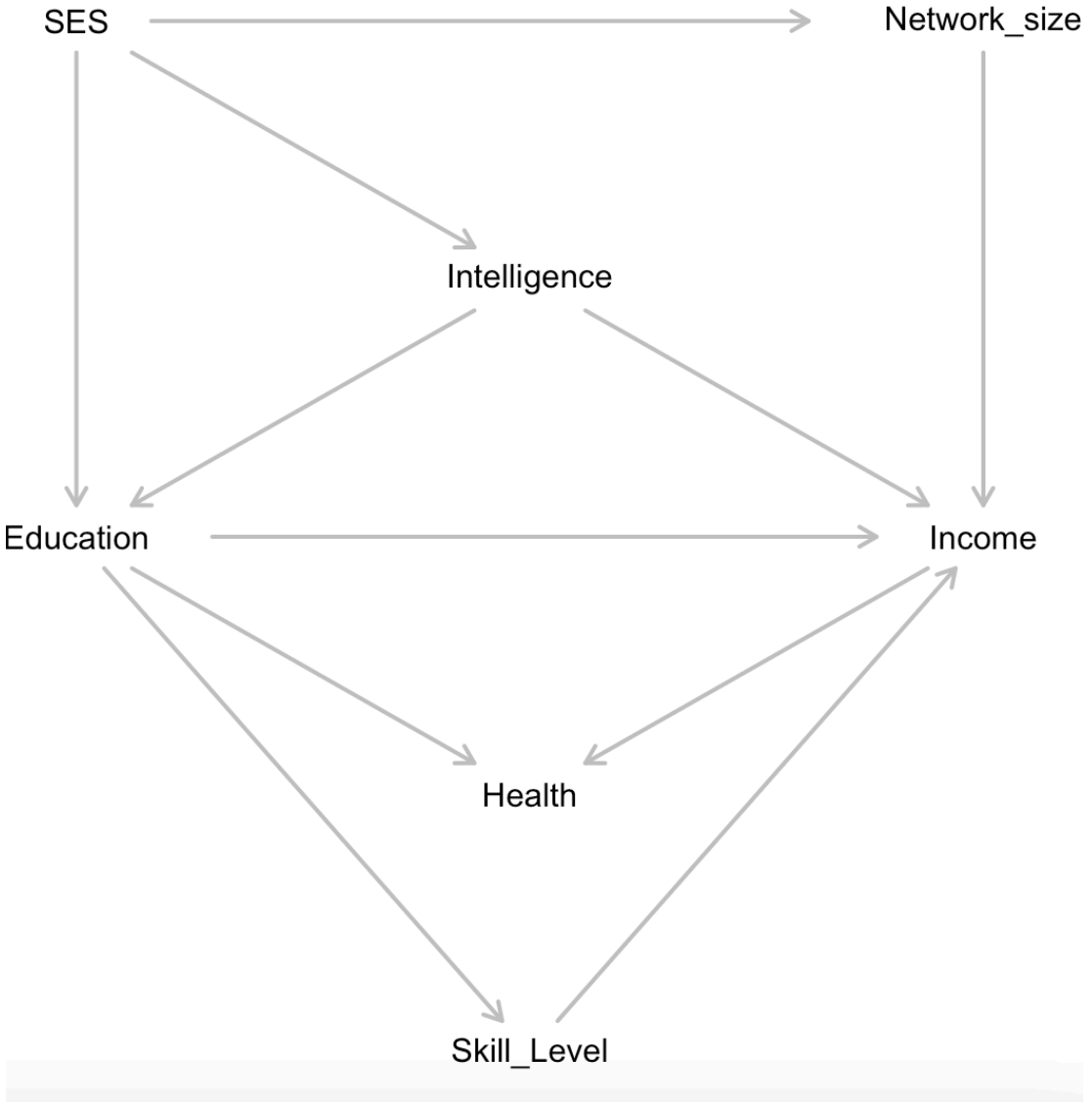
- How can we understand the variable “intelligence”
- A **confounder** is a variable that influences both the independent variable (cause) and the dependent variable (effect), potentially distorting the perceived relationship between them.
- To remove the influence of a confounder we *condition* on it.



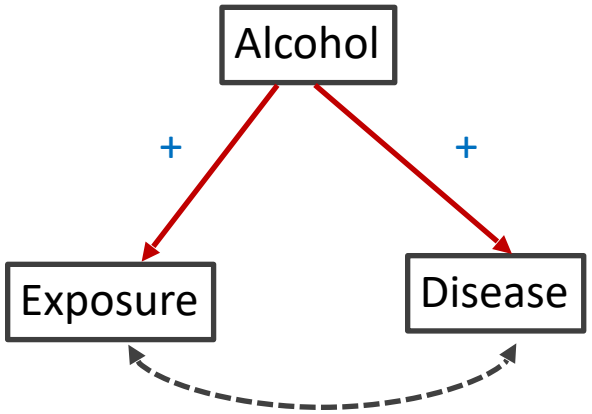
Open and closed paths

- Backdoor paths between exposure and outcome may be closed or open.

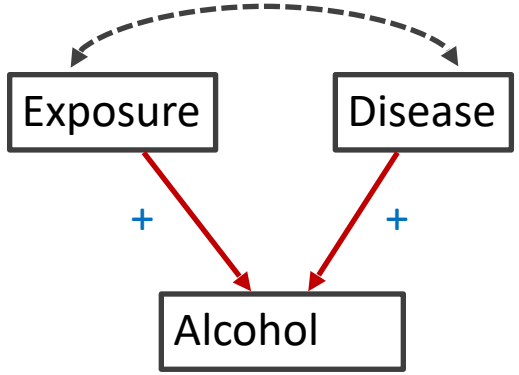
- A path is open if all of the variables along that paths are allowed to vary.
  - An **open** path between to variables leads to a statistical association between them, even if there was no causal relationship.
- A path is **closed** if at least one variable along the path has no variation.



What to control for



Confounder

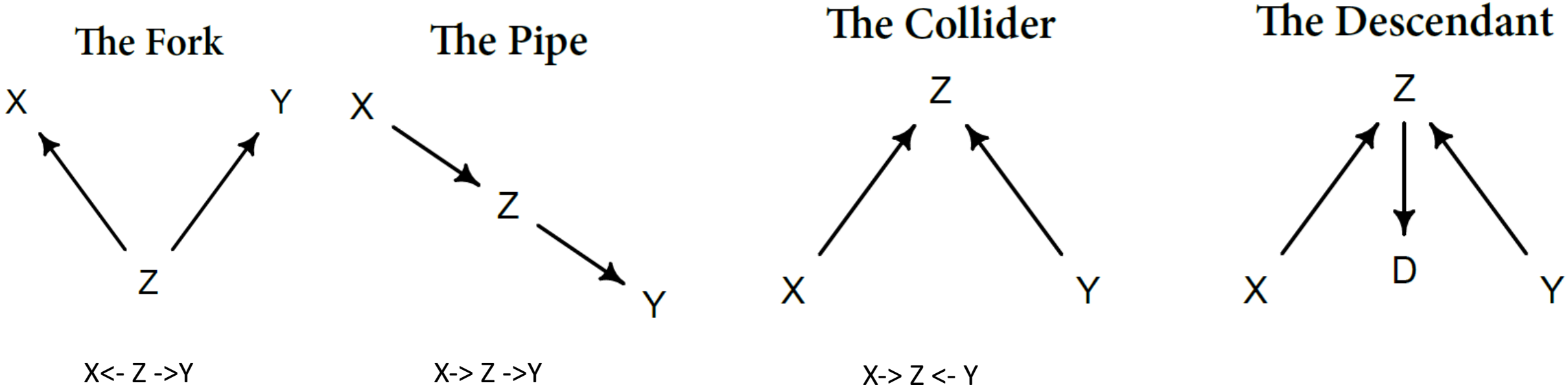


Collider

- What will happen to the correlation between exposure and disease if we condition on alchol consume in the figure on the right?
- Collider: control variables that are a common effect of two other variables (arrowheads collide).
- Controlling for a collider induces a negative correlation between its common causes, opening an additional backdoor path.
  - Are the most newsworthy published studies the least trustworthy?

- A backdoor path is closed, provided we control for at least one non-collider on the path.
- A backdoor path including a collider is closed, provided we do not control for the collider in the statistical model.
- Blocking all confounding paths between some predictor X and some outcome Y is known as shutting the backdoor.

DAG elements



X<- Z ->Y

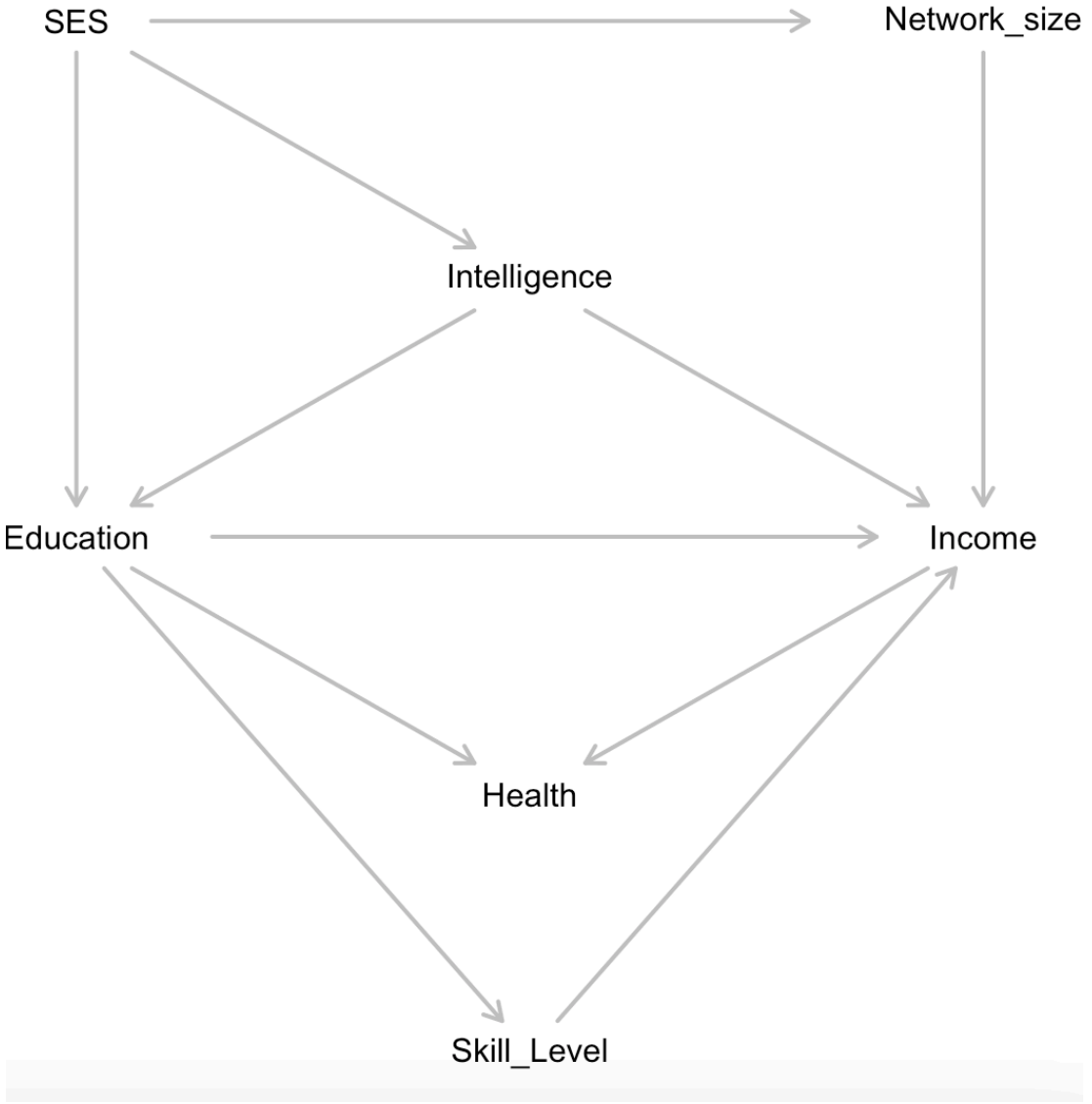
X-> Z ->Y

X-> Z <- Y

- DAG's are always built out of these four types of elements
- What is the effect of controlling for Z?

How do we close paths?

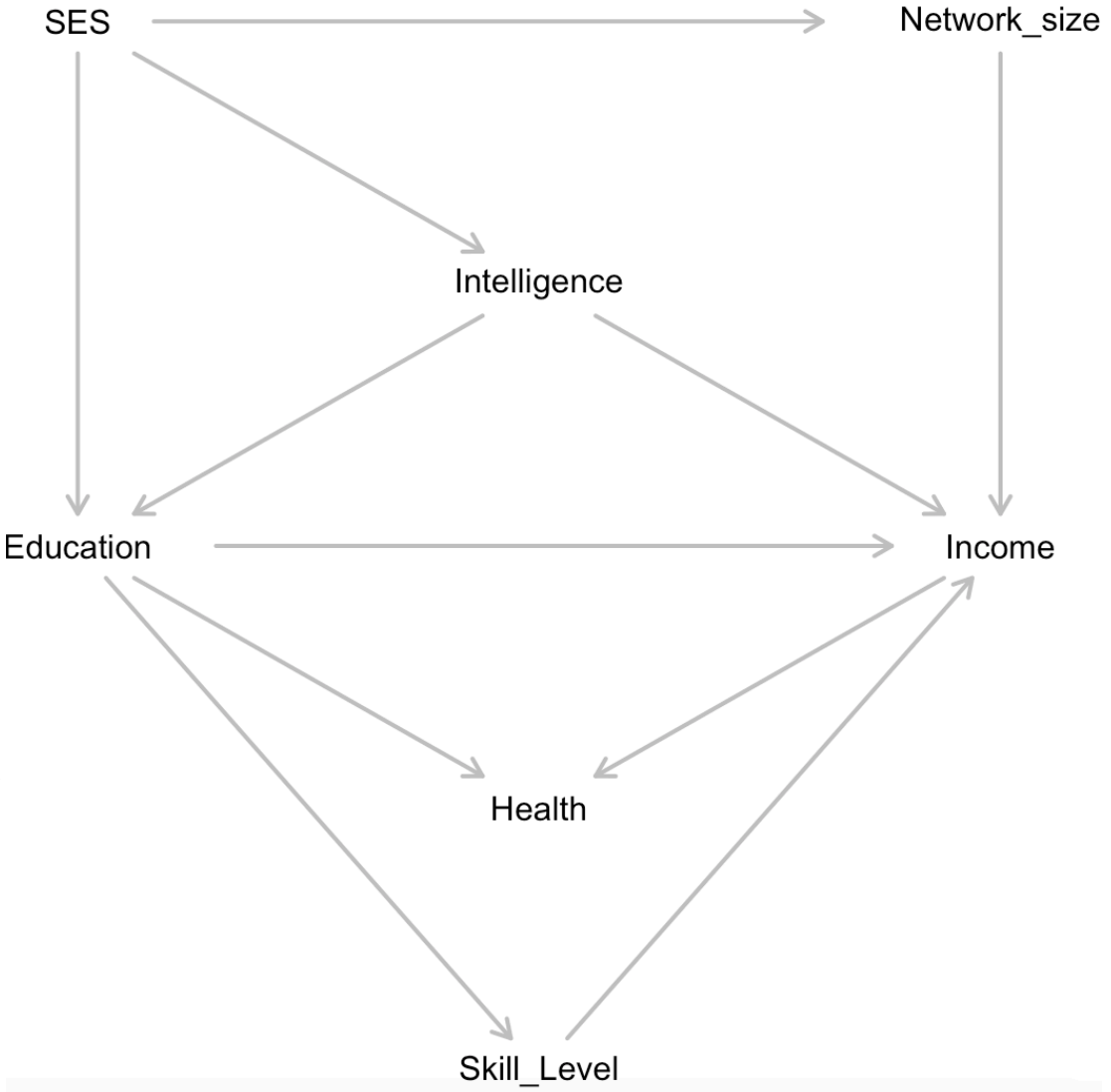
- Which variables should we control for?



How do we close paths?

- Which variables should we control for?

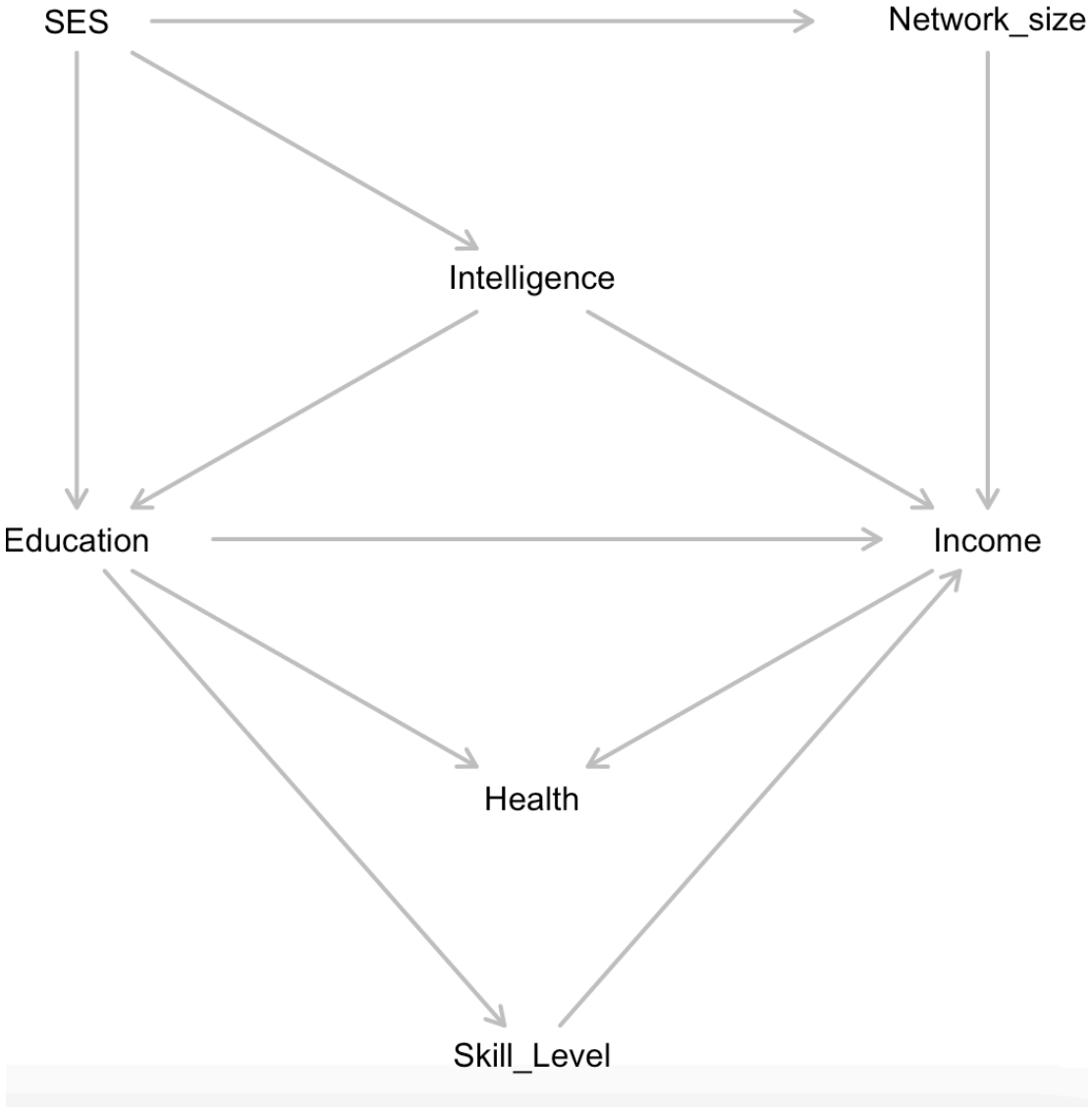
```
> paths(dag_2 , from="Education" , to="Income" )
$paths
[1] "Education -> Health <- Income"
[2] "Education -> Income"
[3] "Education -> Skill_Level -> Income"
[4] "Education <- Intelligence -> Income"
[5] "Education <- Intelligence <- SES -> Network_size ->
Income"
[6] "Education <- SES -> Intelligence -> Income"
[7] "Education <- SES -> Network_size -> Income"

$open
[1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[7]  TRUE
```



How do we close paths?

- Which variables should we control for?

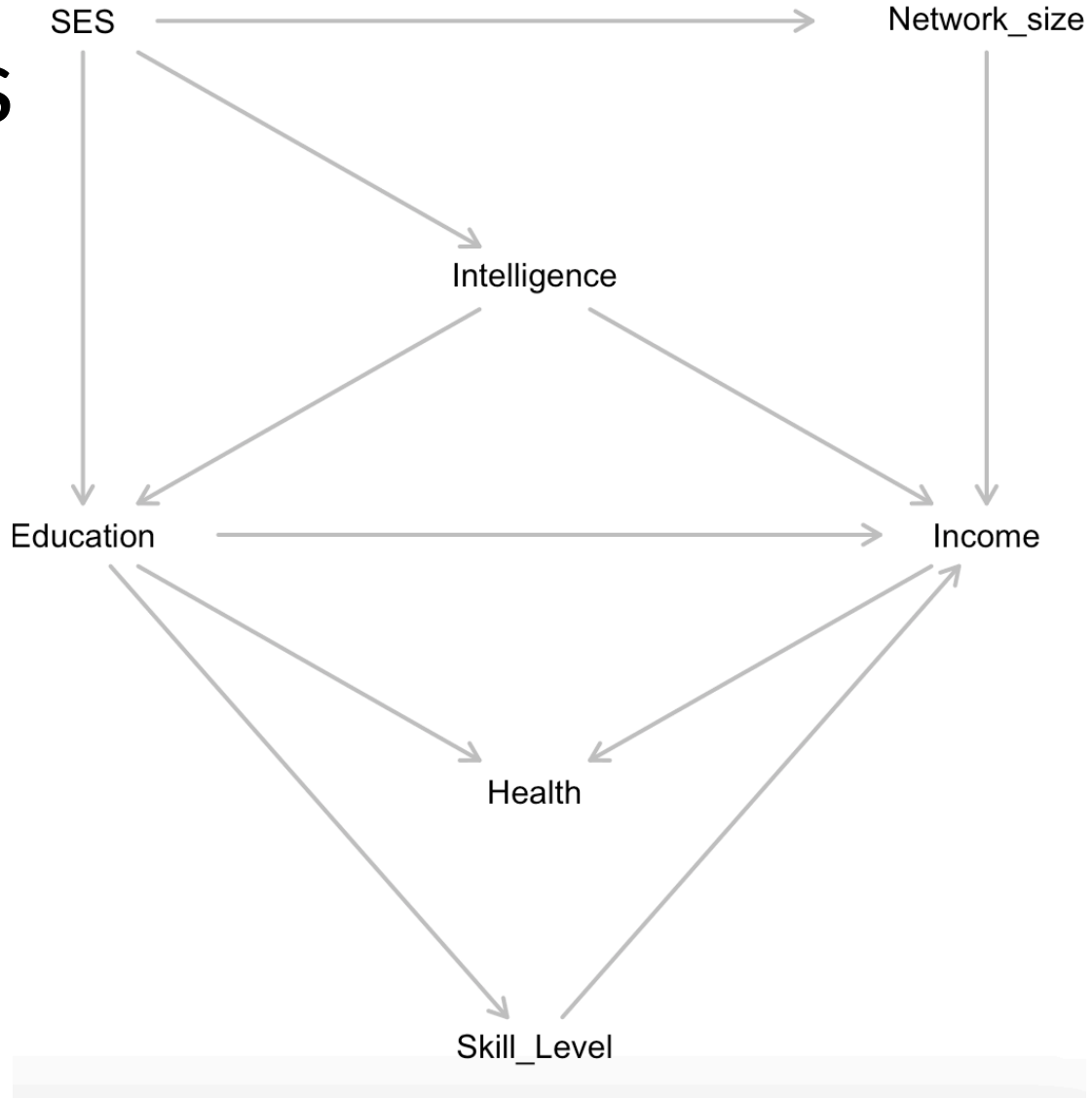
```
> adjustmentSets( dag_2 , exposure="Education" ,
outcome="Income" )

{ Intelligence, Network_size }
{ Intelligence, SES }
```



Who controls for colliders

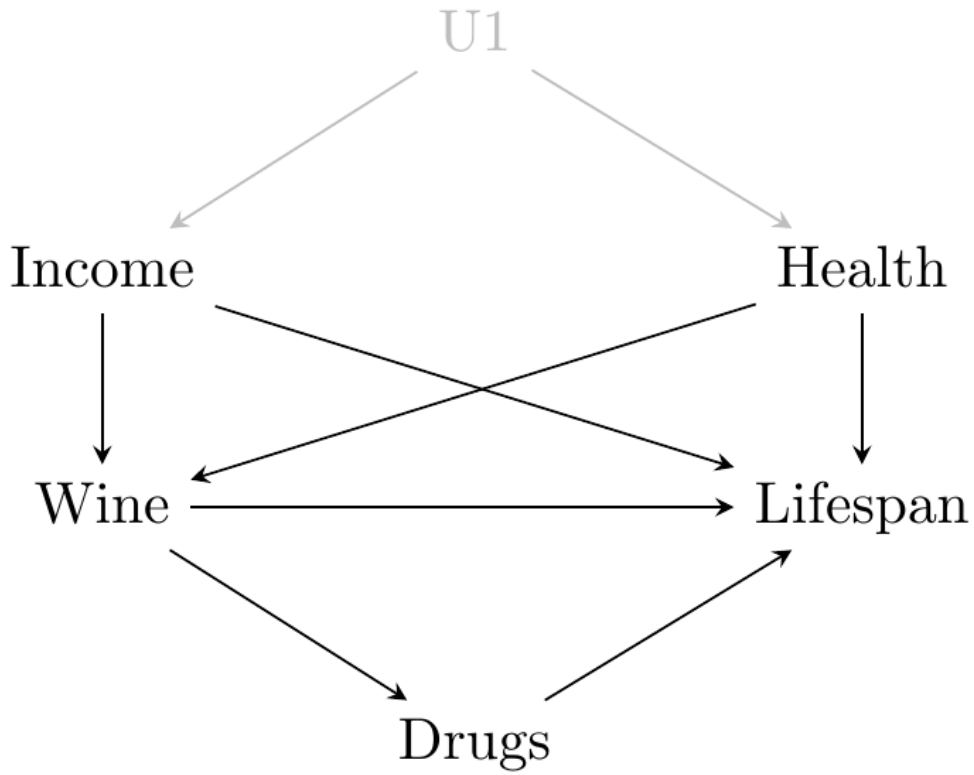
- Colliders are only a problem if you control for them, so just don't control for them?

1. Colliders are often disguised as variables it feels like you *should* control for.
2. One common way we control for colliders is by *selecting a sample.*
   -



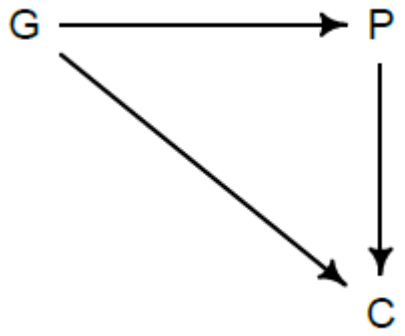
Problem 1: Unmeasured confounders

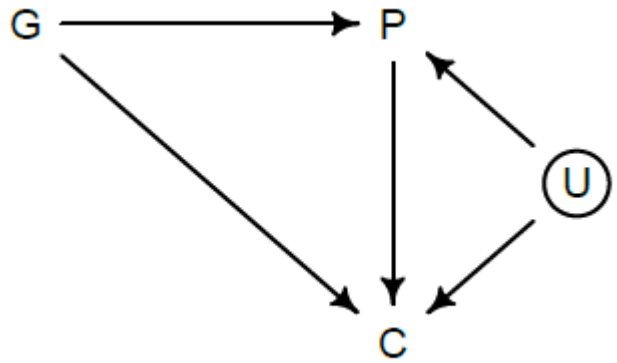
- If we can control for at least one variable on each of our bad paths without controlling for anything on one of our good paths, we have identified the answer to our research question.
- Which variables would you control for here?



Problem 2:

- Suppose for example that we are interested in inferring the direct effect of grandparents (G) on the educational achievement of children (C).





- Now P is a common consequence of G and U, so if we condition on P, it will bias inference about G → C, even if we never get to measure U.

Statistical methods for causal inference

An alternative to closing back doors

- Closing back doors is a widely used approach to strengthen claims of causal effects, but has a number of problems.
  - Do you kniow what the confounder is?
  - Is the confounder measured?
  - Can the confounder be measured?

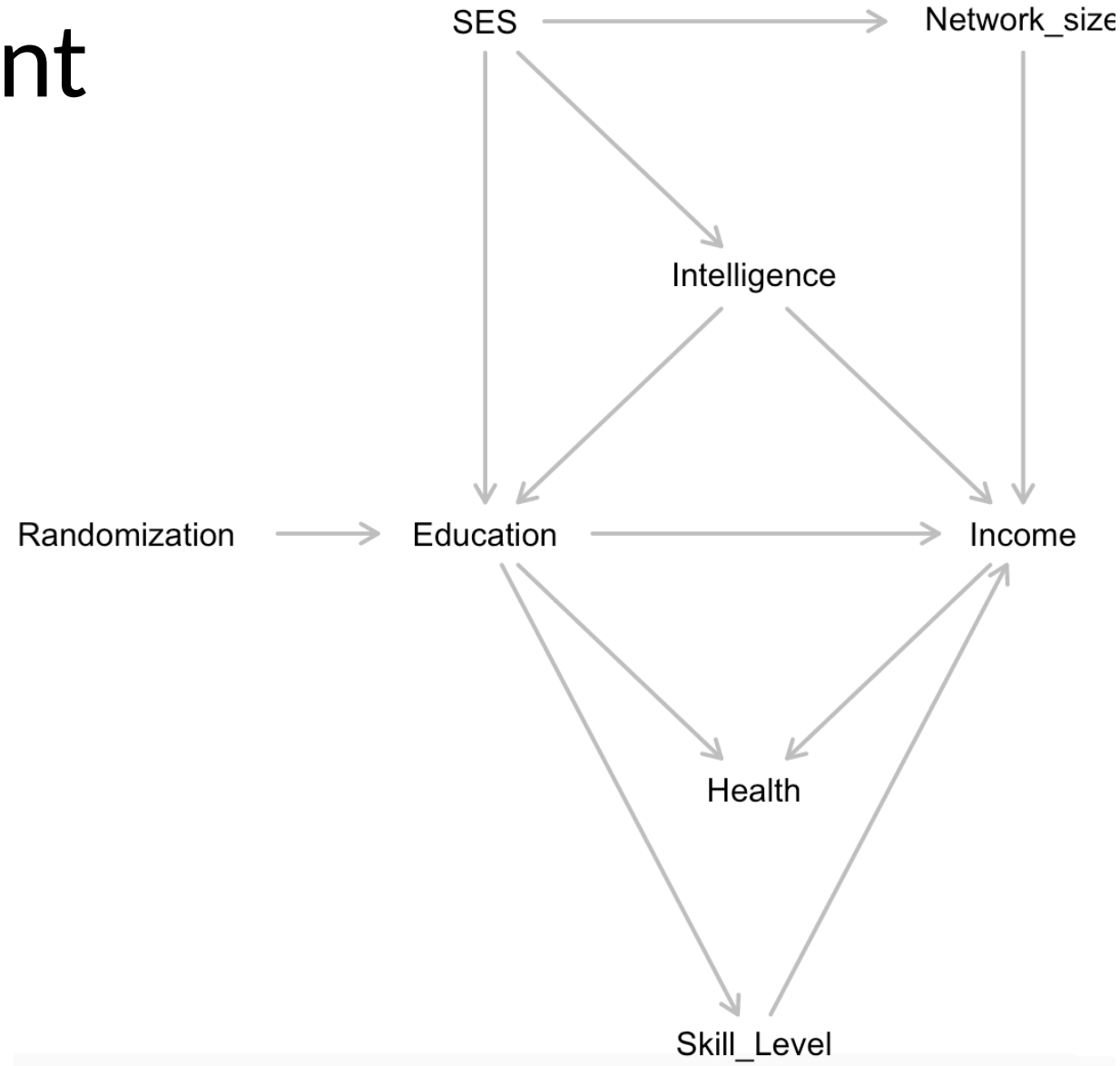
- An alternative approach to identifying the answer to a research question, instead of actively closing back doors, is to find ways of *isolating just Front Doors*.

Two approaches to causal inference

- Removing back doors vs. opening front doors.
- A host of methods are used to «close all back doors» by design.
- Randomized controlled trials
  - Natural experiment

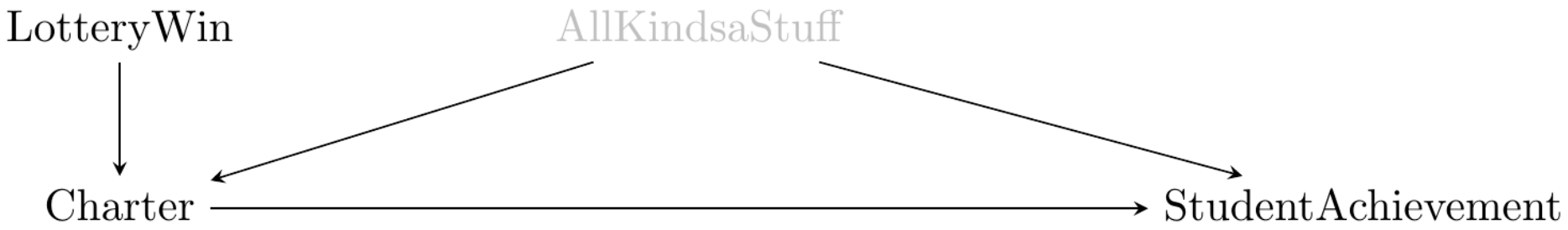
Randomized experiment

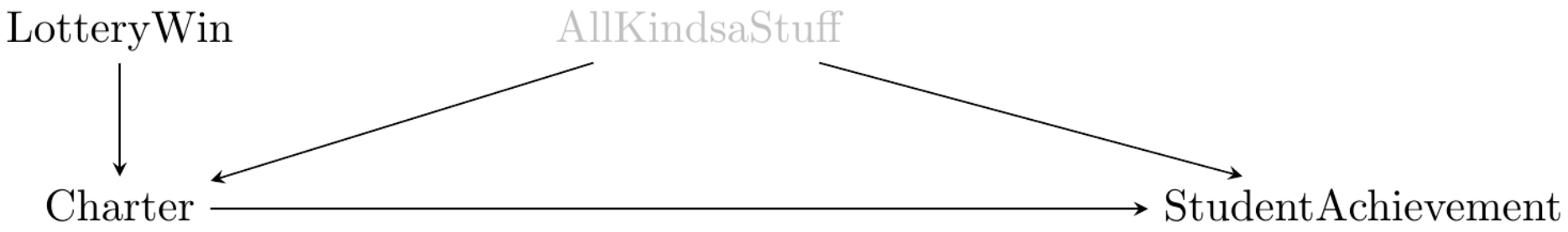
- How can we isolate the causal path?
- The most famous solution is to run an experiment. If we could assign education levels at random, it changes the graph



Natural experiments

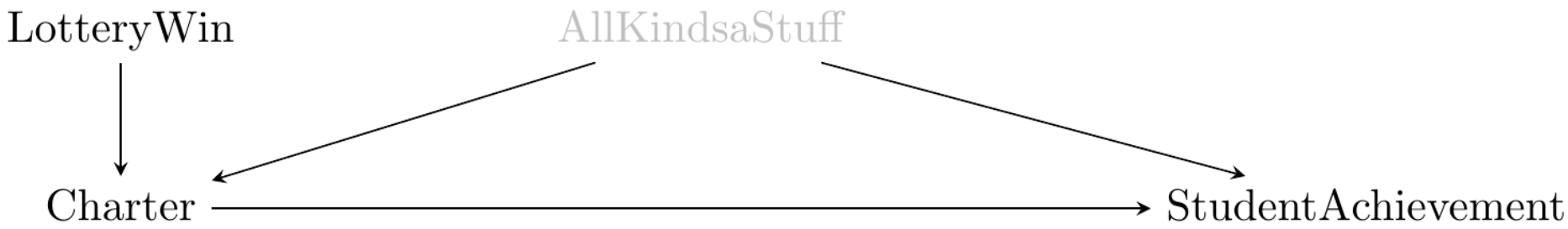
- **Natural experiments**: When randomization of a treatment occurs without a researcher controlling the randomization.





- Does not remove the requirement of closing back doors, but hopefully makes it easier.
  - We need to close the back doors between randomisation and outcome, as well as any front doors that does not pass through treatment.

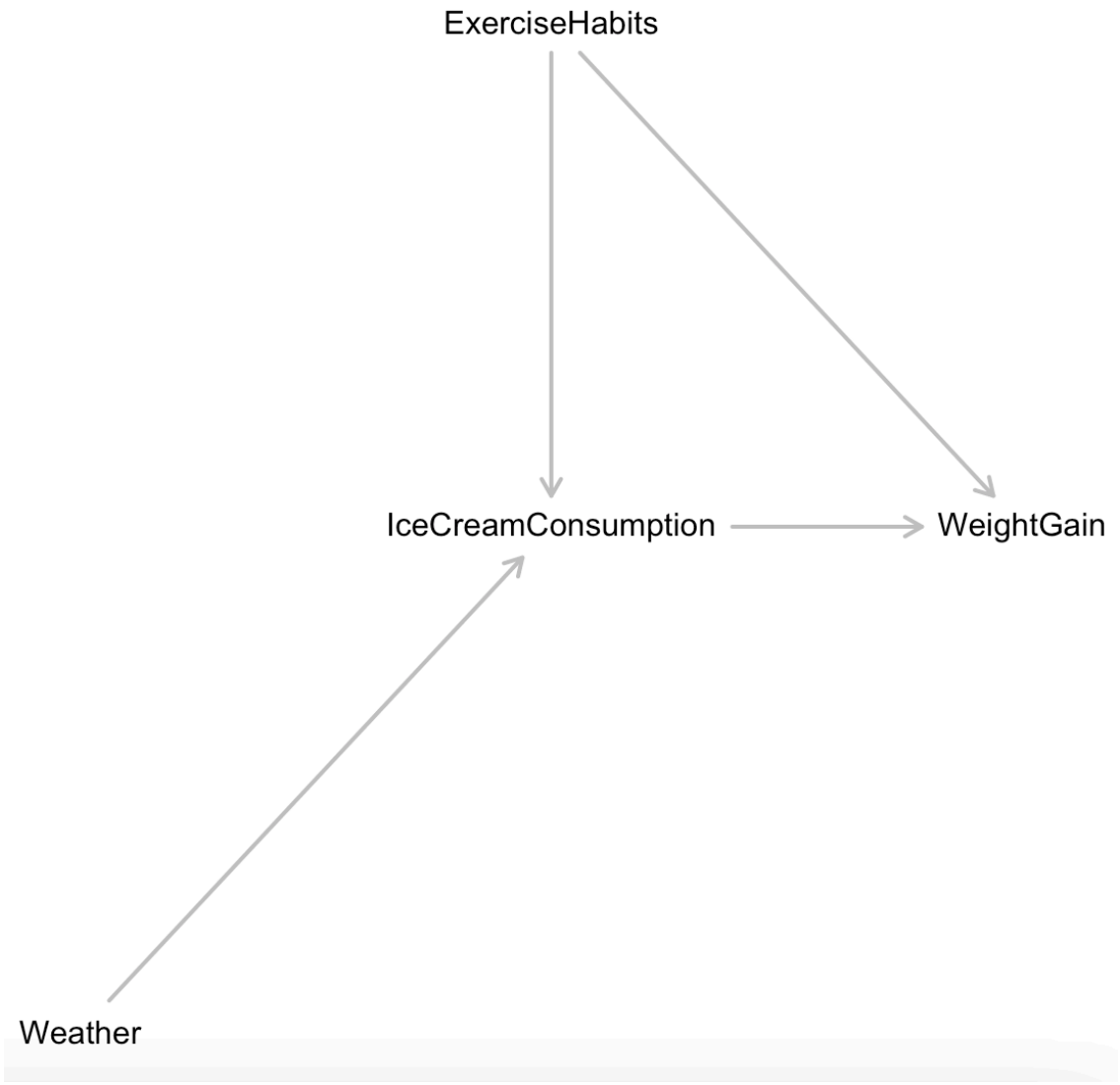
Instrumental variable overview



1. Use the instrument to explain the treatment
2. Remove any part of the treatment that is *not* explained by the instrument
3. Use the instrument to explain the outcome
4. Remove any part of the outcome that is *not* explained by the instrument.
5. Look at the relationship between the remaining, instrument-explained part of the outcome and the remaining, instrument-explained part of the treatment

Ice cream example

- **Relevance**: Weather is related to ice cream consumption because people are more likely to buy and eat ice cream when it’s nice outside.
- **Exogeneity**: Weather is independent of personal characteristics that could affect weight gain, such as metabolism or exercise habits. It’s random in the sense that it doesn’t change based on a person's behavior or characteristics.
- **No Direct Effect**: We assume weather doesn’t directly affect a person's weight (people don't gain or lose weight simply because it’s sunny or rainy). So, any effect we see of weather on weight is likely through its impact on ice cream consumption.



Experiments vs. natural experiments

1. Sometimes there *will* be back doors from instrument to the outcome, which doesn't happen with pure randomization.
2. Natural experiments are more natural, people may not even realize they're a part of an experiment
   - Sample sizes are often larger
3. Because we are isolating just the variation in treatment that is driven by the Natural Randomness, we are tossing out any treatment that occurs for other reasons.
4. Convincing others that the variance is exogenous can be difficult.

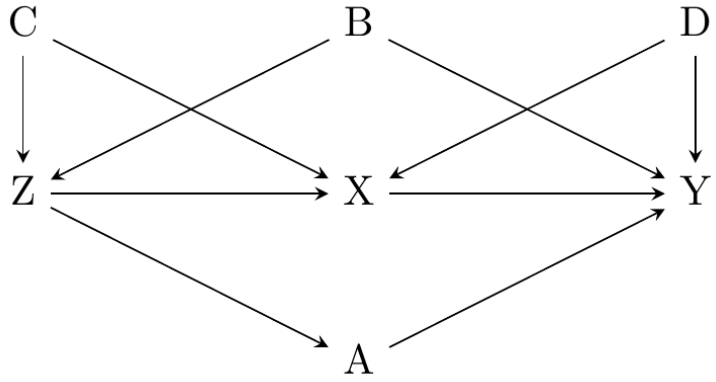
Not all instrumental variables are good

- If you go back to the 1970s or 1980s you can find people using things like parental education as an instrument for your own

- A really good instrument usually takes one of two forms:
  1. it represents real randomization
  2. is one that you would never think to include in a model of the outcome variable, and may be surprised to find that it ever had anything to do with assigning treatment
     - Eg. balanced sex ratio as IV for maternal employment.

Assumptions for Instrumental Variables

- **Relevance**: The idea of instrumental variables is that we use the part of X, the treatment, that is explained by Z, the instrument.
  - If Cov(Z,X) is small, we'd call Z a *weak instrument* for X, if zero, the approach breaks down.
- **Exogeneity:** No confounding influences between Z and X.
- **Exclusion Restriction**: Any paths between the instrument Z and the outcome Y must either pass through the treatment X or be closed.
  - in effect, the assumption that the instrument Z *is* a variable that has no open back doors of its own
- **No Direct Effect:** Of Z onto Y .

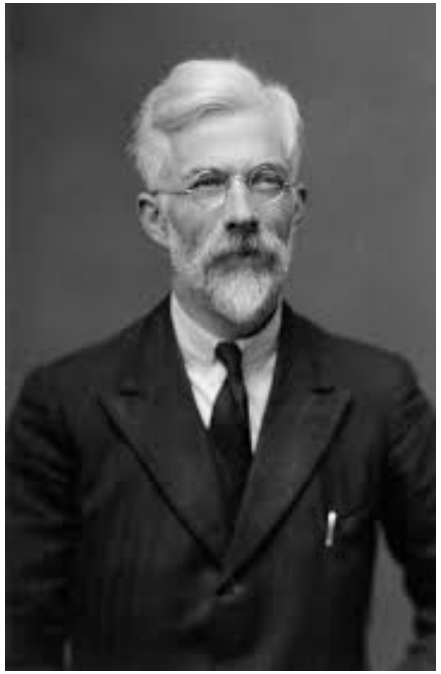


Mendelian randomization

Genetic recombination is a natural experiment

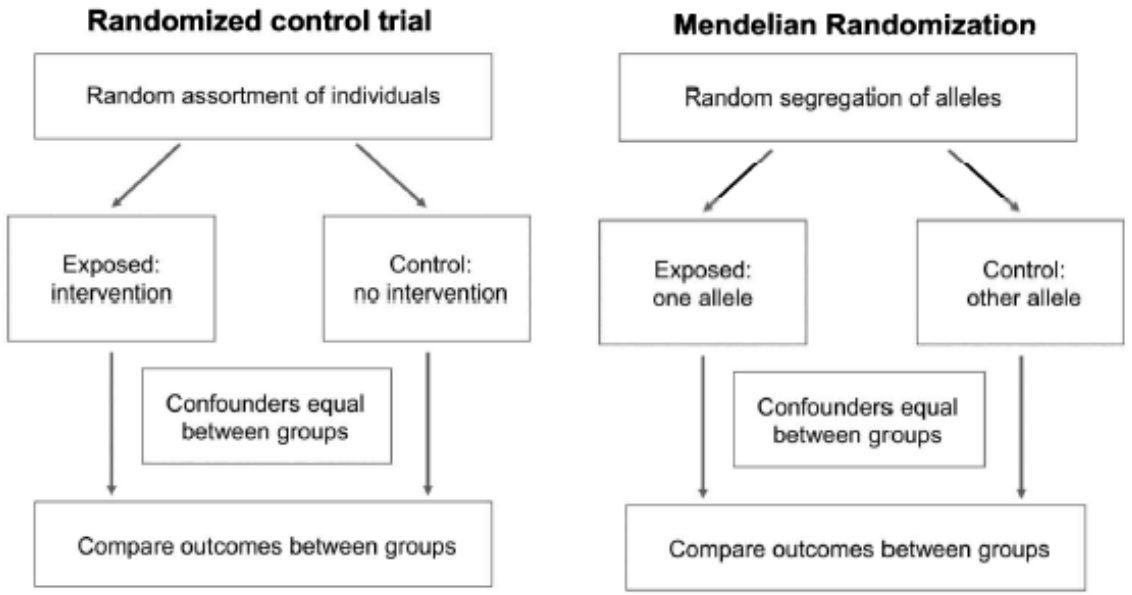
- *Genetics is indeed in a peculiarly favoured condition in that Providence has shielded the geneticist from many of the difficulties of a reliably controlled comparison. The different genotypes possible from the same mating have been beautifully randomized by the meiotic process. A more perfect control condition is scarcely possible, than that of different genotypes appearing in the same litter.*

  - Ronald Fisher (1951)



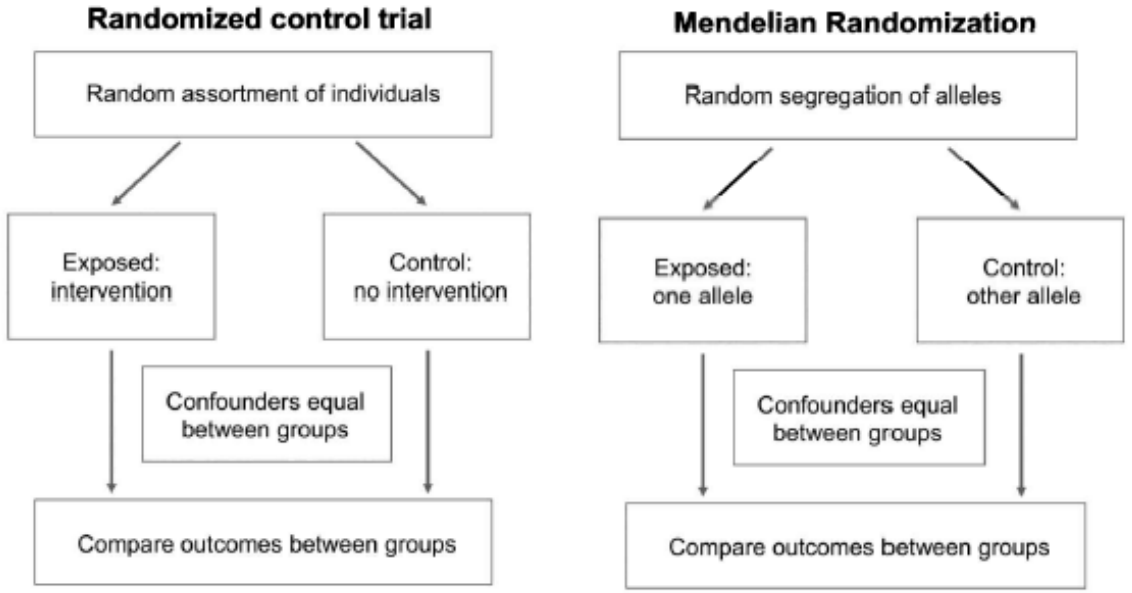
Mendelian *randomisation*

- **Mendelian Randomization**: A genetics-based method for inferring causality between risk factors and health outcomes.
- **Mendelian Inheritance**: Genes are transmitted randomly from parents to offspring, with random allocation of genes at conception.
- **Mendel's second law of inheritance**, the law of independent assortment states that a pair of traits segregates independently of another pair during gamete formation.

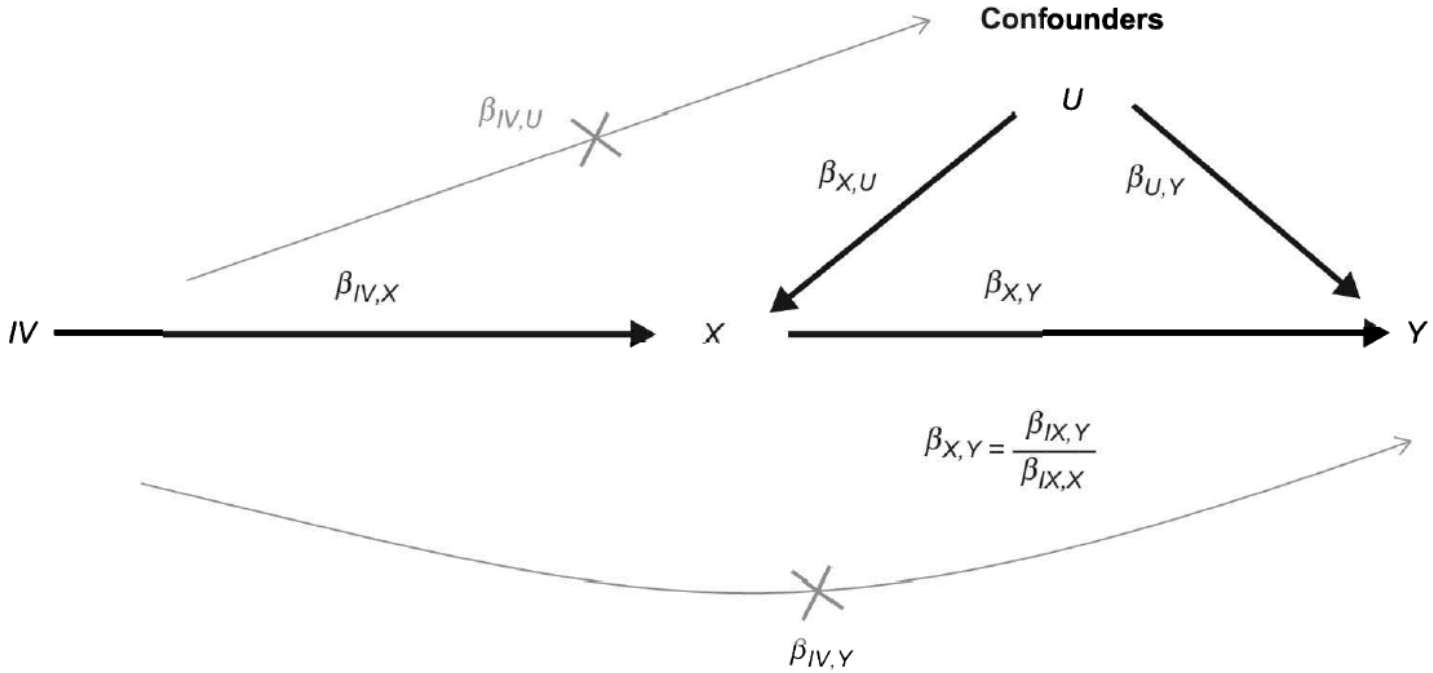


Mendelian *randomisation (MR)*

- **Random Allocation of Genes**: MR leverages the Genetic Variants as Proxies: Utilizes genetic variants (like SNPs) associated with risk factors (such as high cholesterol) as instrumental variables.
- Similar to a blinded randomized controlled trial.

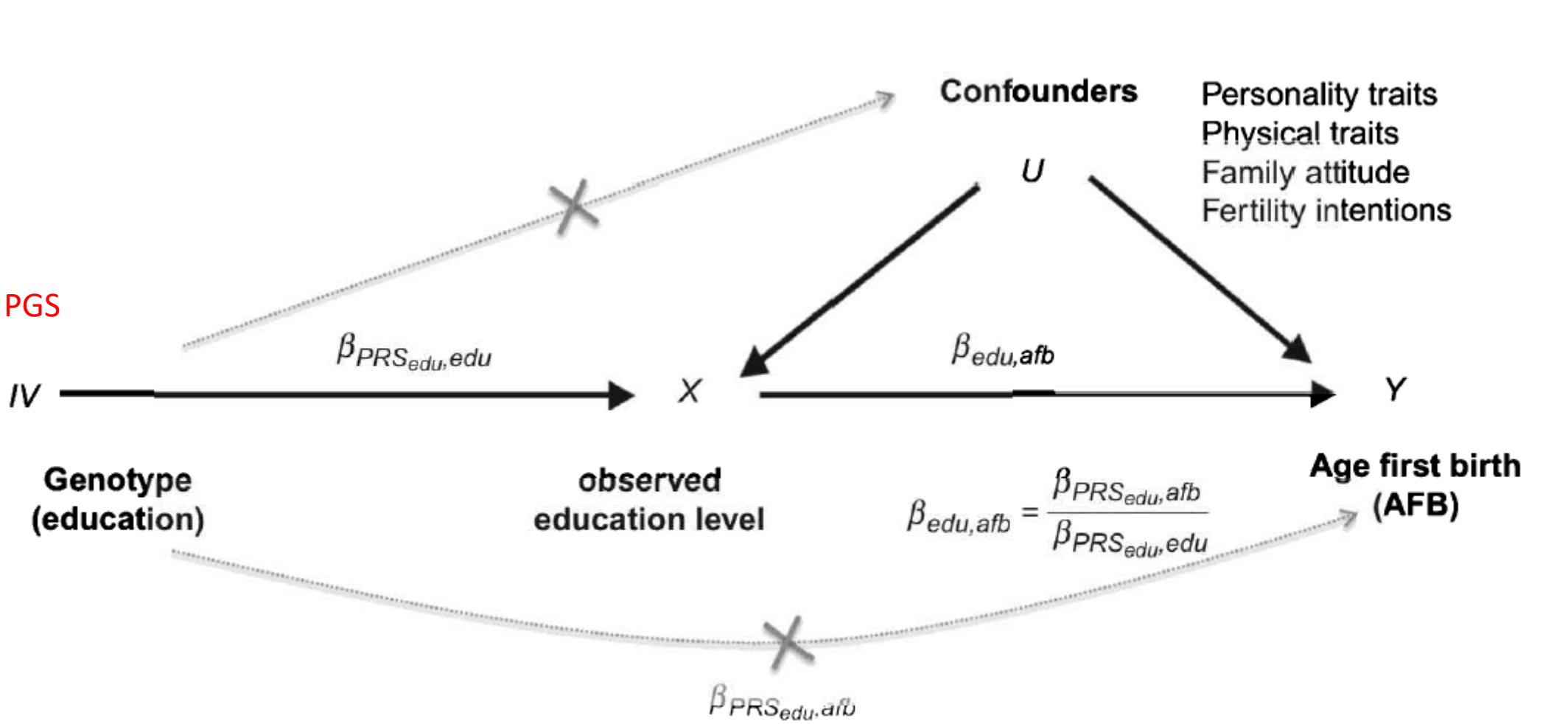


MR – the central idea



- If we find a genetic variant that is deemed causal for the exposure variable, and we can plausibly argue that this variant does not have a direct causal effect on the outcome of interest, an association between the genetic variant and the outcome variable can only be observed via the causal effect of exposure on outcome variable.

Age of first birth to educational attainment



Estimating the parameters

- Two stage least-squares

1. In the first stage we regress X on the IV

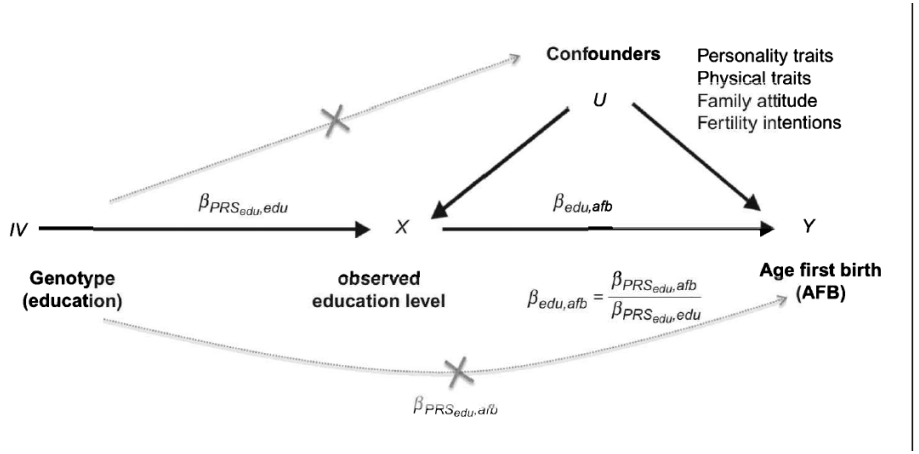
$$X = \mu_1 + \gamma IV + \varepsilon_1$$

2. In the second stage we regression Y on the predicted value of X, based on the first stage

$$Y = \mu_2 + \beta_{2SLS}\hat{X} + \varepsilon_2$$

This is equivalent to

$$Y = \mu_3 + \rho IV + \varepsilon_3 \qquad \frac{Cov_{y,z}}{Cov_{x,z}}.$$



The Causal Effects of Education on Adult Health, Mortality and Income: Evidence from Mendelian Randomization and the Raising of the School Leaving Age

Neil M. Davies
University of Bristol

Matt Dickson
University of Bath and IZA

George Davey Smith
University of Bristol

Frank Windmeijer
University of Bristol

Gerard J. van den Berg
University of Bristol and IZA

MARCH 2019

- 74 SNPs that associated with educational attainment at genome-wide significance levels ($p<5x10-08$) in the discovery sample of the educational attainment GWAS (see Okbay et al., 2016

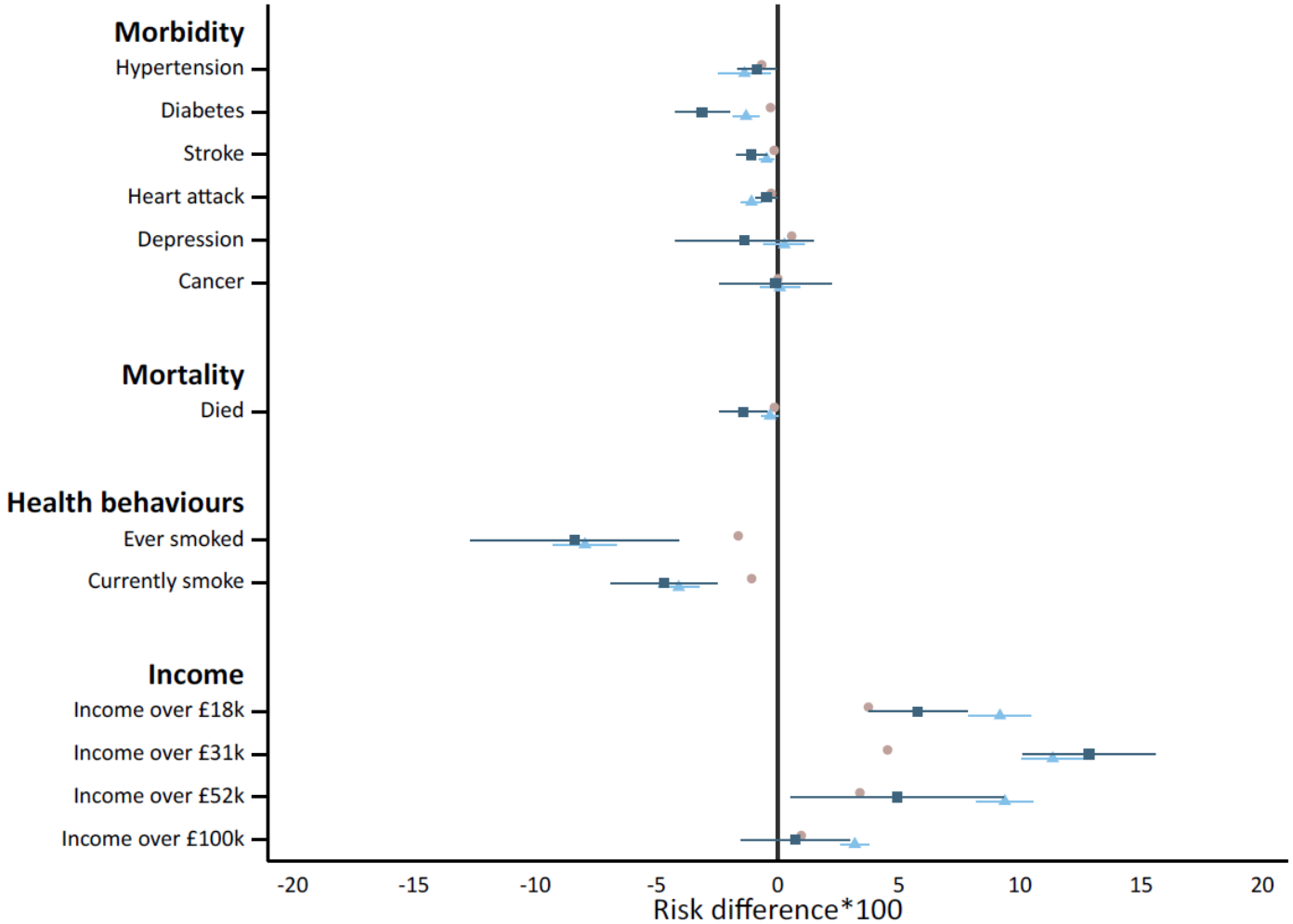
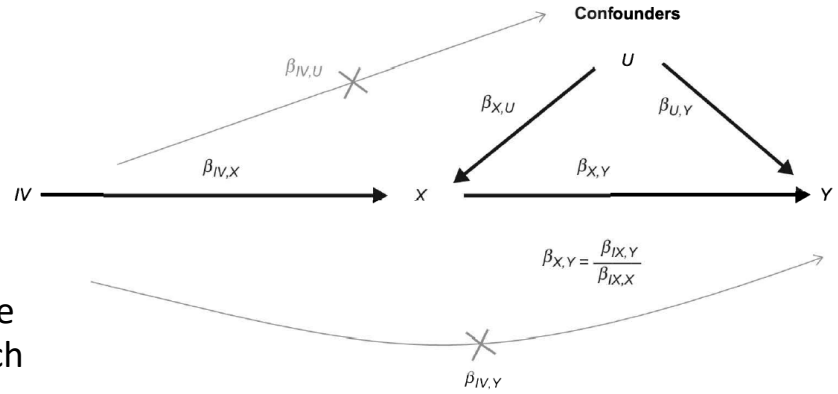


Figure 6: The effect of one additional year of schooling on morbidity, mortality and socioeconomic outcomes, estimated via OLS ●, and instrumenting education using the 1972 Raising of the School Leaving Age ■, and the polygenic educational attainment genetic risk score ▲.

Assume interest in the relationship between a genetic predisposition to high cholesterol and the risk of heart disease.

- **Weak Instruments Assumption:**
  - A gene variant that's weakly associated with cholesterol levels, and won't be a good predictor of cholesterol levels and thus can't be a reliable instrument.
- **Exclusion Restriction:**
  - The gene affecting cholesterol levels should not affect heart disease by any path other than through cholesterol levels.
  - If a gene variant not only influences cholesterol but also blood pressure independently we can't be sure if it's the cholesterol or blood pressure (or both) causing heart disease, which violates the exclusion restriction.
- **Independence Assumption:**
  - The gene variants used as instruments must be independent of any confounders that affect both cholesterol levels and heart disease.
  - This is like saying the genetic lottery for high cholesterol is random and not linked to other factors, such as diet or exercise habits that could independently affect heart disease risk.
- **No Genetic Assortative Mating:**
  - Genetic assortative mating occurs when people choose partners based on traits that are genetically influenced, which could bias MR studies.
  - If people with a genetic predisposition to high cholesterol are more likely to mate with others who also have heart disease risk factors, the offspring's genes are not randomly assorted in relation to the disease.

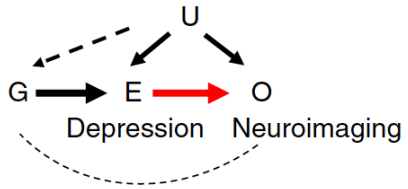


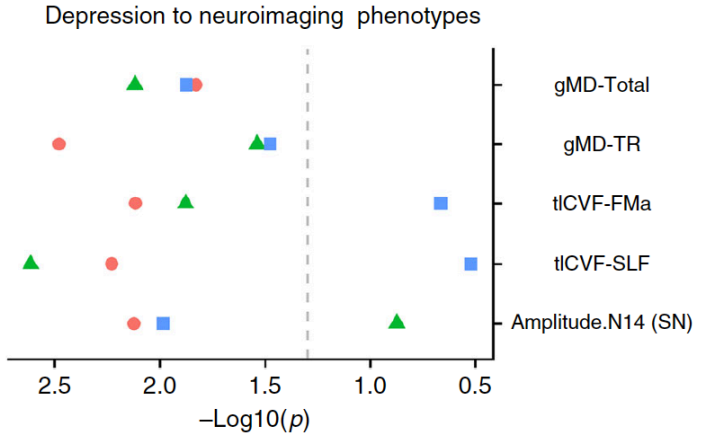
Ex: causal consequences of depression

A phenome-wide association and Mendelian Randomisation study of polygenic risk for depression in UK Biobank

Xueyi Shen[1], David M. Howard[1,2], Mark J. Adams[1], W. David Hill[3,4], Toni-Kim Clarke[1], Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium*, Ian J. Deary[3,4], Heather C. Whalley[1,153] & Andrew M. McIntosh[1,3,4,153]

https://doi.org/10.1038/s41467-020-16022-0





- GWAS have identified 44 and 102 risk-associated genetic variants for depression.
- Depression is phenotypically correlated with many behaviours, brain structure and function measures, cognition and physical conditions.
- Findings suggest that variation in white matter microstructure is a causal consequence of depression.

Further examples

- **Consequences of alcohol consumption**
  - Association between ALDH2 and hypertension supports a causal effect of drinking alcohol.

- **Body mass index and mortality**
  - PGS of BMI in sample of N=335,000 in UK biobank suggests 1 unit increase in BMI increases risk of death by 3%.

- **Educational attainment and Alzheimers disease**
  - IV analyses using both school laws and PGS of educational attainment finds evidence for a causal effect.

Conclusion?

- A shift to environmental influences, specifically "causal" ones
- Directed Acyclic Graphs – DAG’s
- Instrumental variable approach to causal inference
- Mendelian randomisation
- Genes as «instruments» in natural experiments