LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

CSM Forum

Statistical methods for causal inference

Rhian Daniel

Friday 14th October, 2011 · 12.45pm

Summary of 'Confounded about confounding?' session

- We introduced the formal mathematical language needed for causal discourse
  - Y^x is the value outcome Y would have taken had we intervened on exposure X and set it to x.
- The key assumption that allows us to make causal statements from observational data is **conditional exchangeability**
  - Association is causation within strata of Z if exposed and unexposed are conditionally exchangeable given Z, i.e. if $\{Y^0, Y^1\} \perp\!\!\!\perp X \,|\, Z$.
- Causal diagrams help us to decide, based on our assumptions regarding the causal structure of our variables, on a set of variables $Z = \{Z_1, Z_2, \ldots Z_n\}$ given which conditional exchangeability holds.

What next?

- Suppose our diagram tells us that $Z = \{Z_1, Z_2, \ldots Z_n\}$ is sufficient to control for the confounding.
- HOW do we do it?
- A whistle-stop tour in 14 minutes ...

Outline

Outline

Outline

Analyses following the back-door criterion

- If the number of confounders is small and categorical/binary, we could stratify on them.
- We would then calculate our effect of interest (e.g. an odds ratio) in each stratum and then combine these in the usual way (Mantel–Haenszel), or report them separately if there are effect modifiers of interest, etc.
- If Z is continuous and/or high-dimensional, the natural extension is to adjust for Z in a regression model.
  1. Draw a DAG
  2. Identify a sufficient set of confounders Z
  3. Include them (appropriately) in a regression model
- Our regression model is now a *causal* model.

More formally...

- Consider, for example, Y continuous.
- Associational regression model:

$$E(Y|X = x, Z) = \alpha + \beta x + \gamma^T Z$$

- Causal regression model:

$$E(Y^x|Z) = \alpha' + \beta' x + \gamma'^T Z$$

- If
  1. the set Z has been correctly selected from a correctly-specified DAG
  2. the regression model has been correctly specified

  then $(\alpha, \beta, \gamma) = (\alpha', \beta', \gamma')$ and β can be given a causal interpretation.

The two conditions

1. the set Z has been correctly selected from a correctly-specified DAG

—We dealt with this last week/three weeks ago.

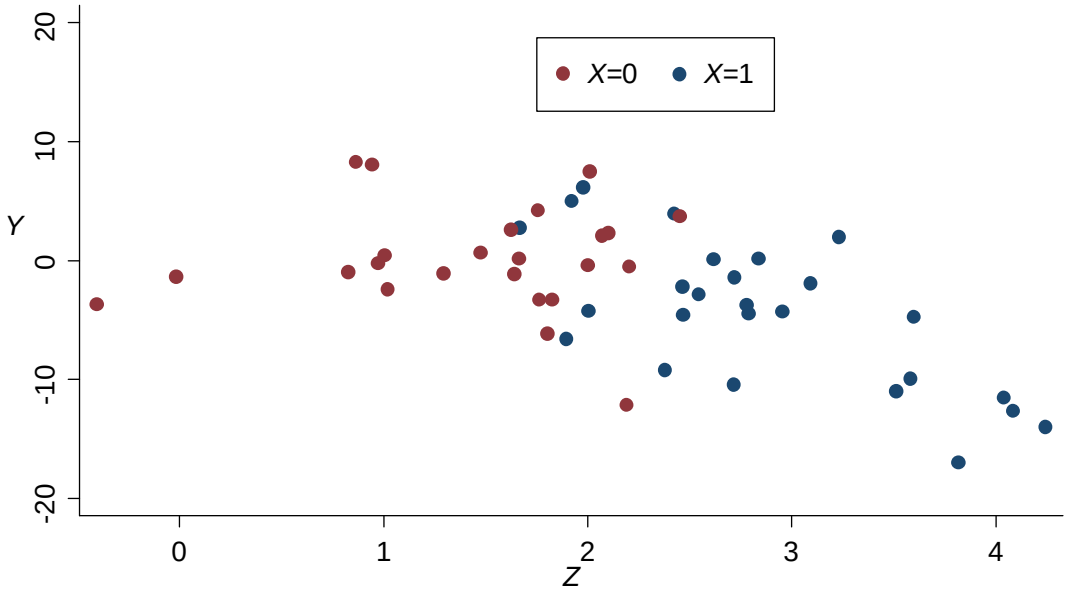
2. the regression model has been correctly specified

—This must be assessed using the usual model-checking techniques.
—But this can be difficult, particularly if there is low overlap across exposure groups.

Poor overlap

A simulated example (1)

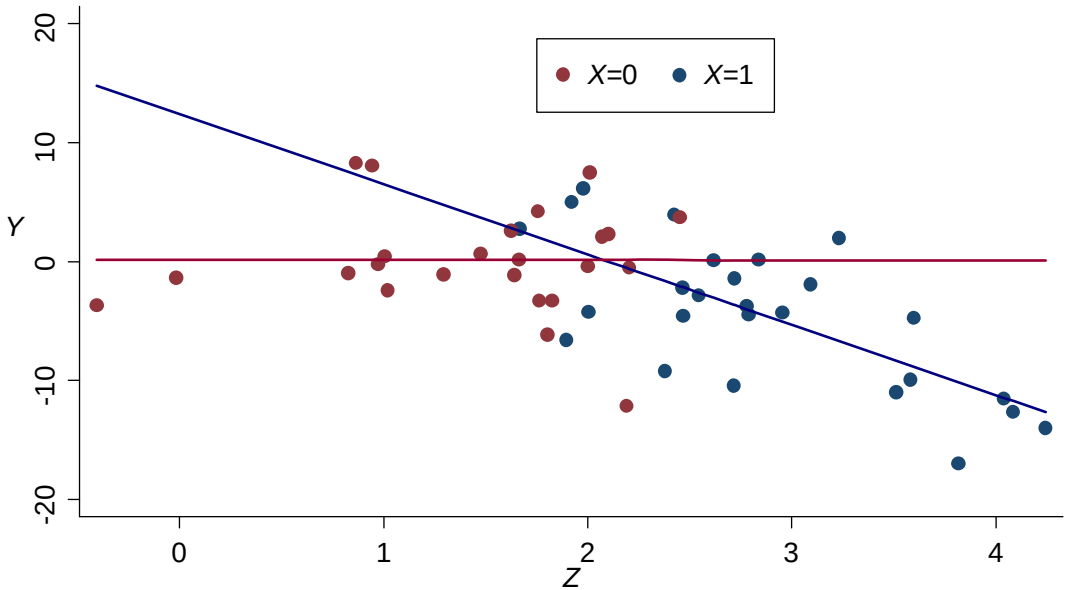
Consider the following example, in which there is little overlap between the Z-values of the $X = 0$ and $X = 1$ groups:



Poor overlap

A simulated example (2)

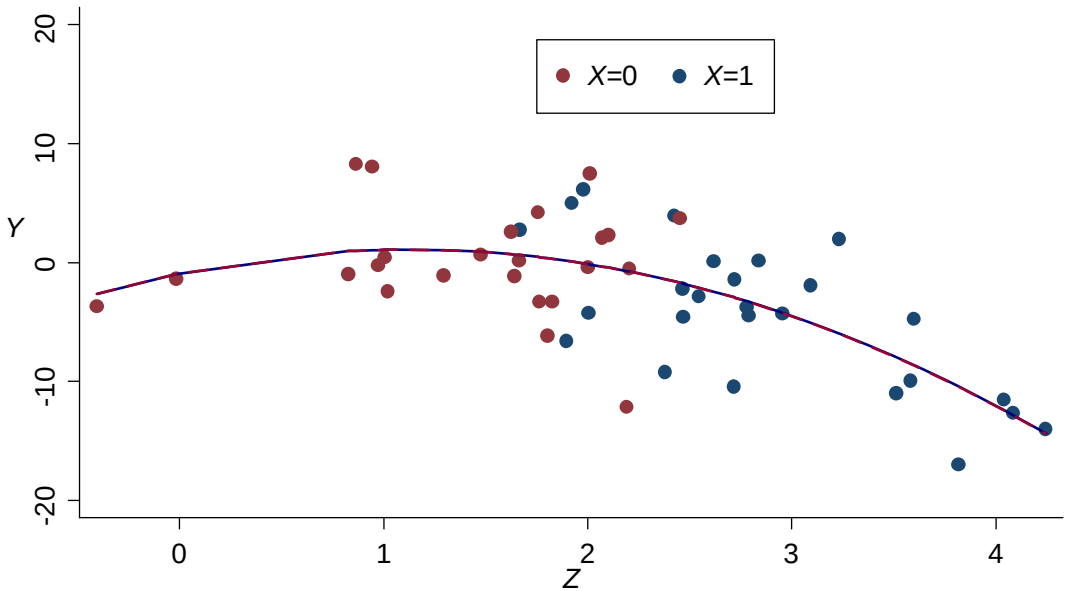
How do we choose between a linear model with an interaction, showing a large causal effect. . .



Poor overlap

A simulated example (3)

. . . and a quadratic model (also with interactions) showing no causal effect?



Too many covariates

Logistic and Cox regression

- The parameters of regression models are typically estimated by maximum likelihood.
- ML estimators are *asymptotically* unbiased.
- For logistic and Cox regression, ML estimators can be noticeably biased in small samples.
- In particular, bias increases as the number of events per parameter decreases.
- The more confounders Z we adjust for, the larger the bias.

Reference

Peduzzi *et al* (J Clin Epidemiol, 1995 & 1996) gave a rule of thumb of "10 or more events per variable".

- What if we have 100 events and more than 10 confounders in Z?

Outline

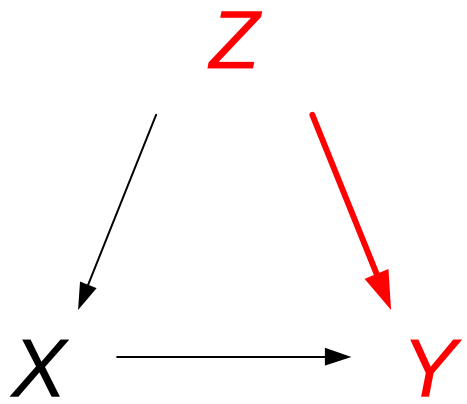
Methods based on the propensity score

Basic idea (1)

Key reference

Rosenbaum and Rubin (Biometrika, 1983) "The central role of the propensity score in observational studies for causal effects".

Instead of modelling this:



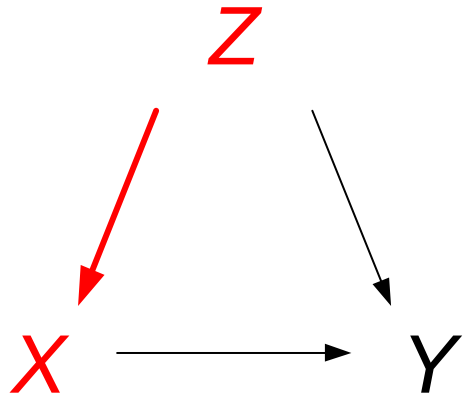
Methods based on the propensity score

Basic idea (2)

Key reference

Rosenbaum and Rubin (Biometrika, 1983) "The central role of the propensity score in observational studies for causal effects".

We model this:



Definition of the propensity score

- The propensity score $p(Z)$ is the conditional probability that $X = 1$ given Z.

$$p(Z) = Pr(X = 1 | Z)$$

- A scalar, irrespective of the dimension of Z.
- Can be estimated from a logistic regression of X on Z.
- Validity of methods based on propensity score relies on correctly modelling $X | Z$.

Some intuition

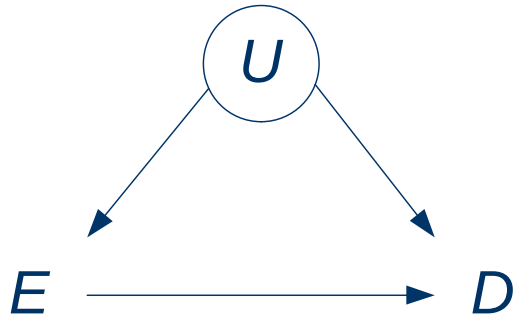
1. If an exposed and unexposed person have the same value of the propensity score, say 0.25, they were equally likely to have received the exposure.
2. As far as we can tell on the basis of their confounders, a coin was tossed to decide which one was exposed.

How it is used

- There are several ways of incorporating our model for $X|Z$ into the analysis: stratification, matching, adjustment, weighting.
- These alternative methods are valid only if correct confounders Z are included and we model $X|Z$ correctly.
- No free lunch!

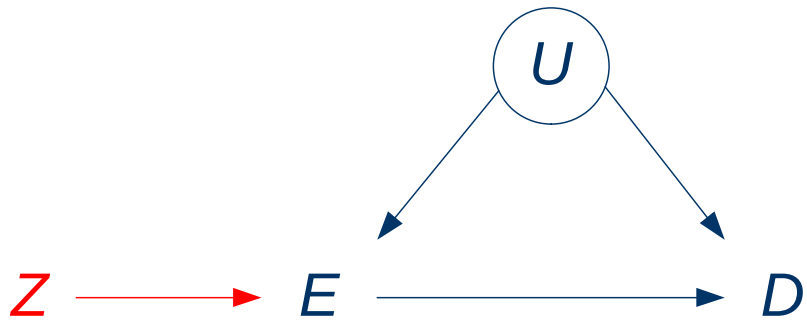
Outline

The problem of unmeasured confounding



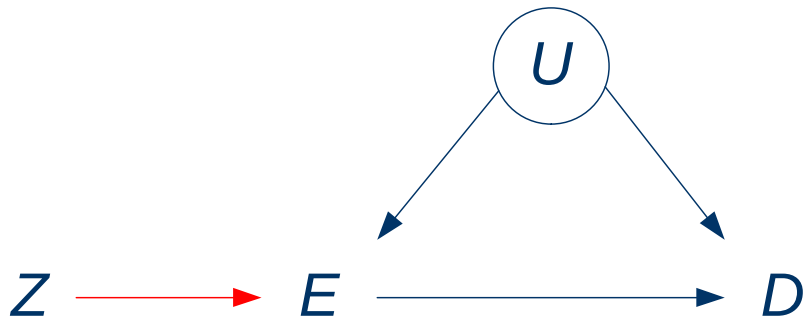
- So far, we have concentrated on methods that aim to measure a sufficient set of covariates to control the confounding of the association between E and D.
- What if this is impossible, and our relationship will certainly suffer from unmeasured confounding?

What is an instrumental variable? (2)



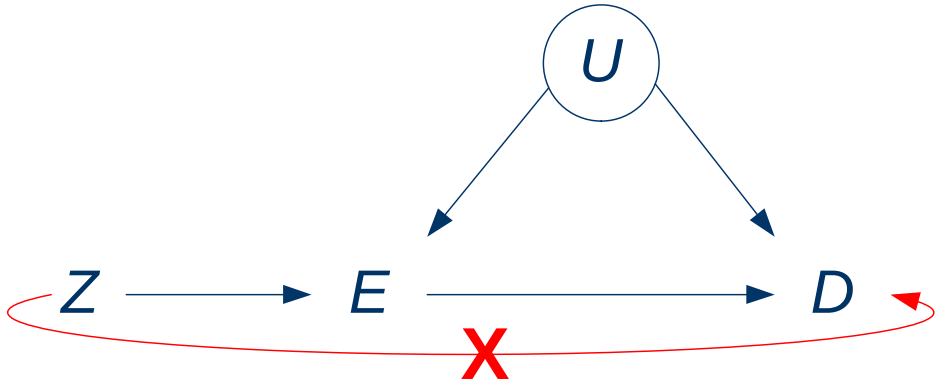
- It turns out that we can still proceed, IF we have measured an instrumental variable Z.

Assumptions (1)



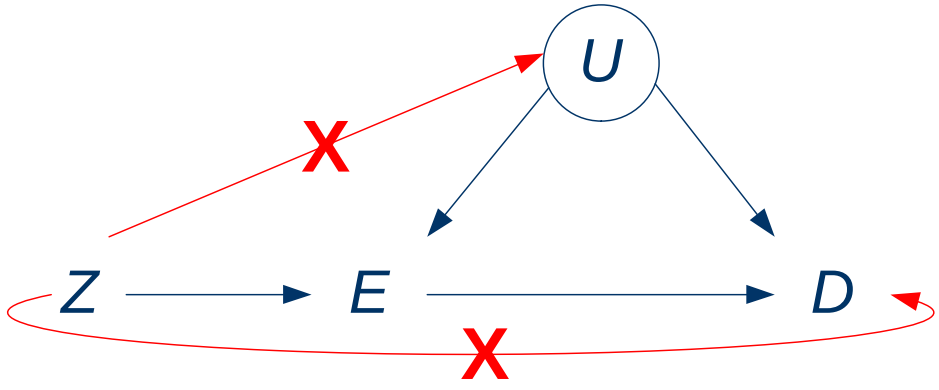
- An instrument Z must be a cause of the exposure.

Assumptions (2)



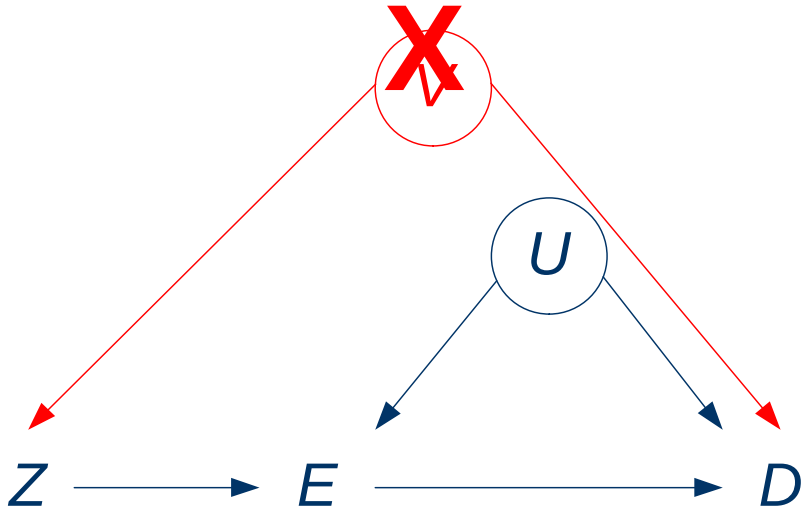
- It must NOT affect the outcome except through its effect on the exposure.

Assumptions (3)



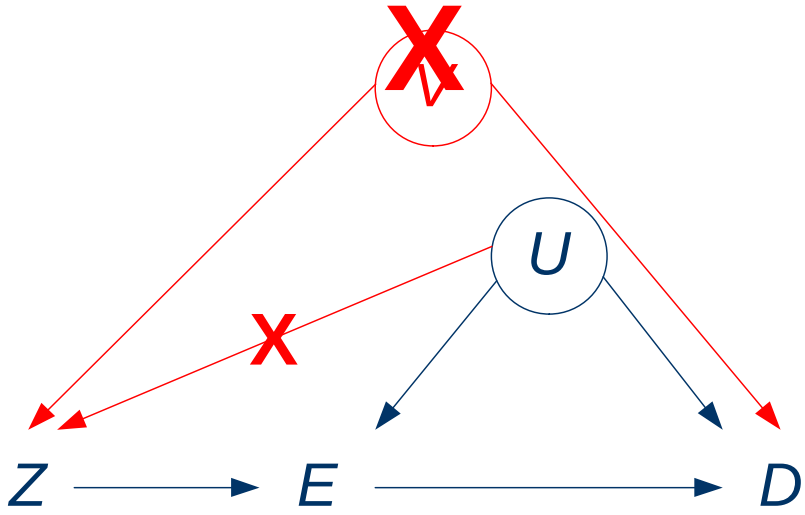
- It must NOT affect the outcome except through its effect on the exposure.

Assumptions (4)



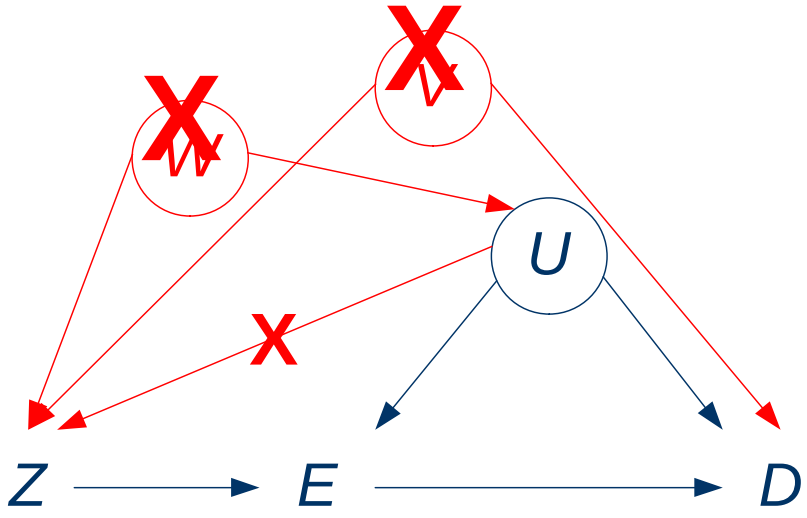
- It must NOT share any common causes with the outcome.

Assumptions (5)



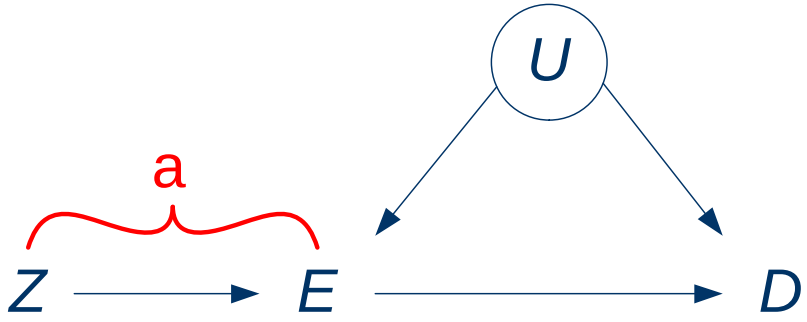
- It must NOT share any common causes with the outcome.

Assumptions (6)



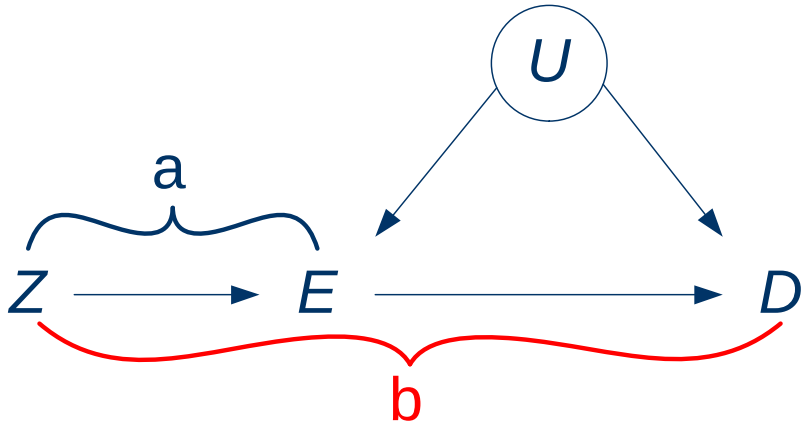
- It must NOT share any common causes with the outcome.

Analysis by two-stage least squares (1)



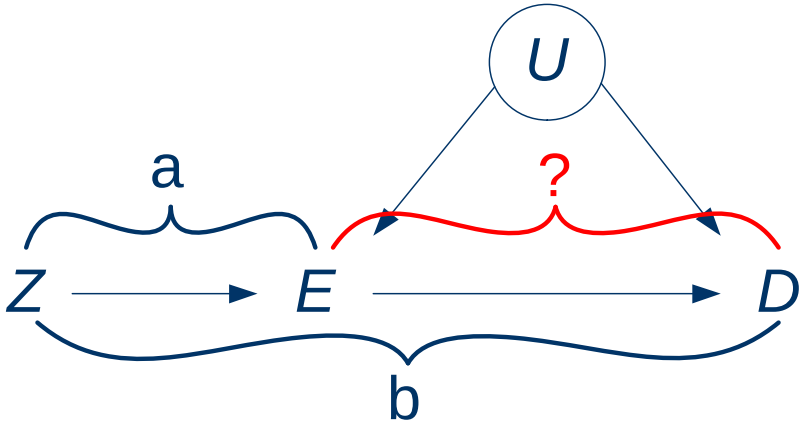
- Informally, how does it work?
- We can estimate a by measuring the association between Z and E (by randomisation, association is causation here).

Analysis by two-stage least squares (2)



- We can similarly estimate b by measuring the association between Z and D (again by randomisation, association is causation here).

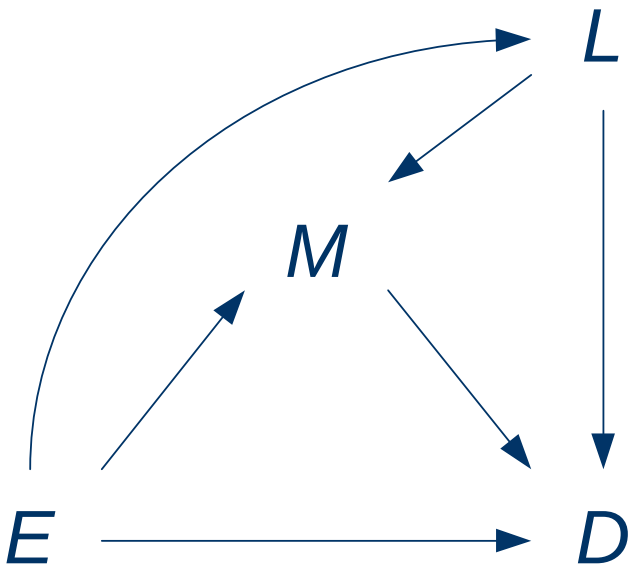
Analysis by two-stage least squares (3)



- We can then infer our causal effect of interest indirectly as the ratio of b and a.

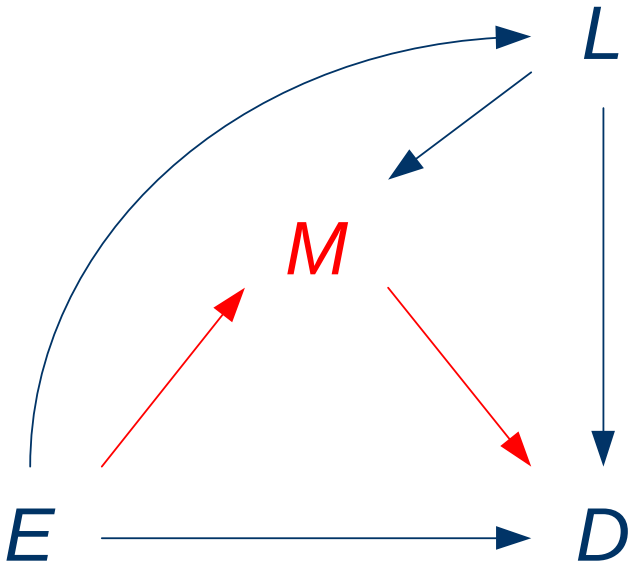
Outline

Pathway-specific effects (1)



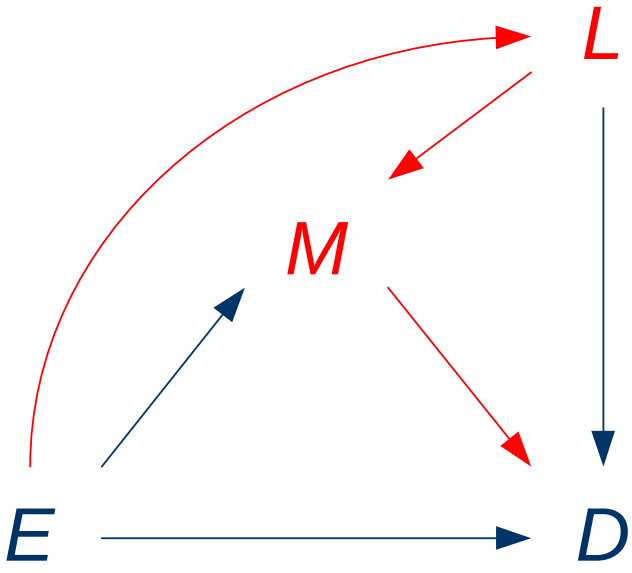
- Sometimes we are interested in more complex causal questions than simply "what is the (total) causal effect of (one exposure) E on (one outcome) D?"
- Take for example this somewhat complicated 'network'.
- E affects D 'directly' and 'indirectly', through M and L.

Pathway-specific effects (2)



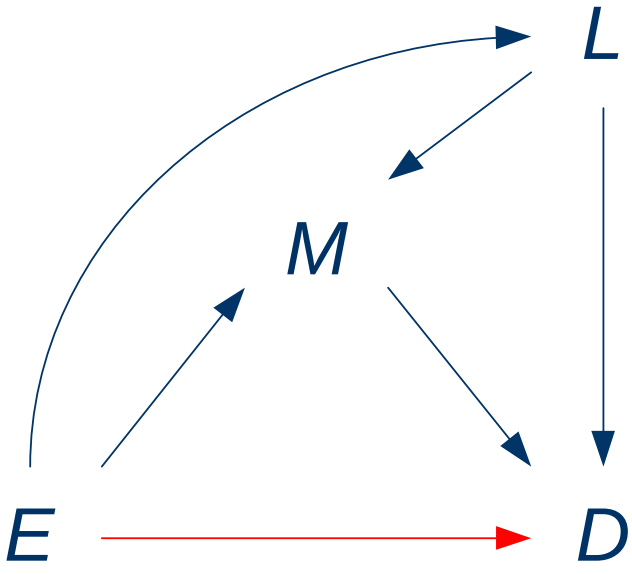
- For example, how much of the causal effect of E on D is mediated by M?

Pathway-specific effects (3)



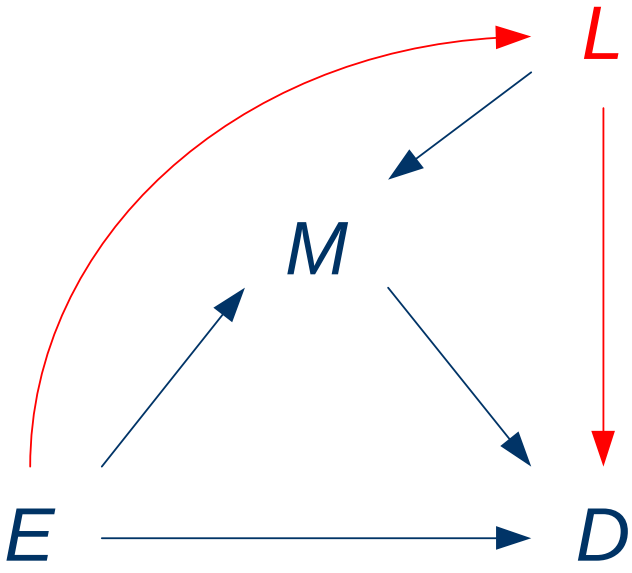
- For example, how much of the causal effect of E on D is mediated by M?

Pathway-specific effects (4)



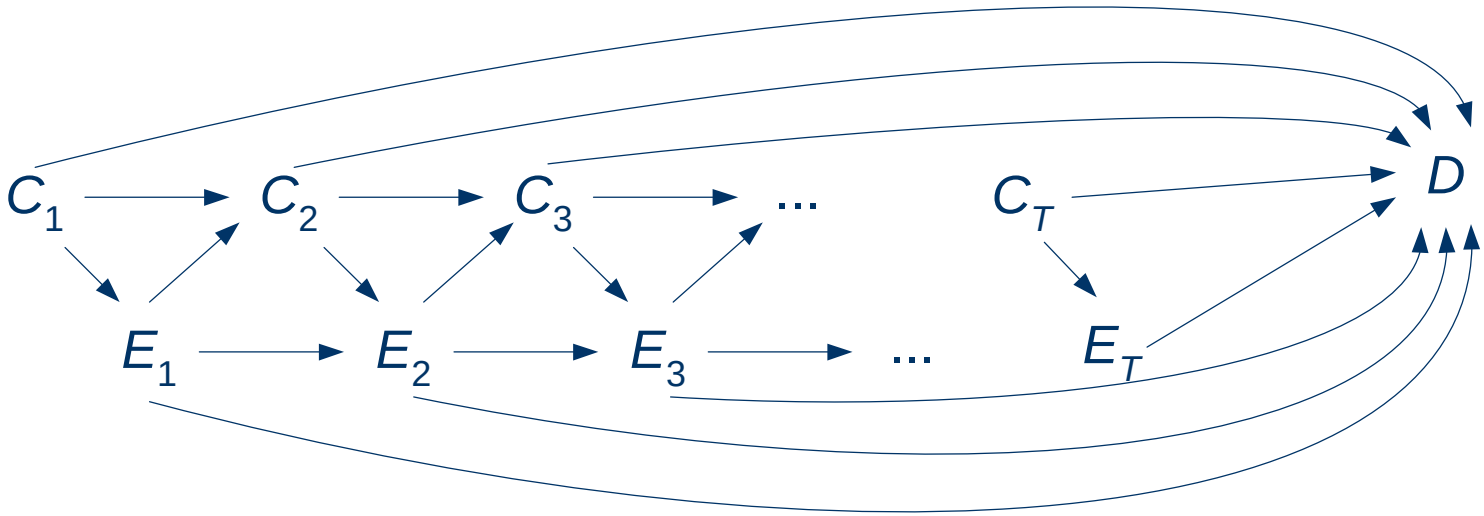
- And how much is not mediated by M?

Pathway-specific effects (5)



- And how much is not mediated by M?

Longitudinal exposures



- Or, what is the joint causal effect of the time-varying exposures E_1, E_2, ... on D when there is causal feedback between E_1, E_2, ... and C_1, C_2, ...
- These are all questions that can (only) be addressed (given certain assumptions) using **novel statistical methods**.

Outline

Summary (1)

What's the point?

- **There's no such thing as a causal statistical method; statistical methods measure associations.**
- So, what's the point of all this causal language, causal assumptions, causal thinking . . . ?

Summary (2)

The point is...

- By expressing the question and our assumptions clearly, we hope to identify either:
  - a particular conditional association (simple case)
  - a combination of (conditional) associations (complex case)

  that answer(s) our causal question under these assumptions.
- Or we see that we can't realistically manage it! (But better to know this...)
- Without a careful causal language and its ability to express our causal assumptions, spotting the appropriate association(s) to estimate would be impossible.
- Clearly communicating (and criticising) the assumptions under which our answer to the causal question is valid would also be impossible.

Want to know more?

- Visit the causal inference theme page on the LSHTM's Centre for Statistical Methodology website: `http://csm.lshtm.ac.uk/themes/causal-inference/`.
- A few places are still available for the 2011 short course *Causal Inference in Epidemiology: Recent Methodological Developments* to be held at LSHTM in the November reading week. For more info, see: `http://www.lshtm.ac.uk/prospectus/short/causal_inference.html`.
- All the methods mentioned briefly today (regression, propensity scores, IV methods, IPW-estimation of marginal structural models, the g-computation formula) and much more, are covered in detail on the short course!