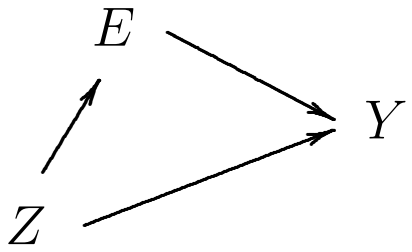
Brief Notes on Confounding

Basic notation/terminology:

- outcome Y (of any type, but continuous for now),
- exposure E (of any type),
- extraneous factor of interest Z (measured or unmeasured; of any type),
- causal diagram:



- confounder (or lurking variable) Z: extraneous factor that exerts confounding of the relation $E \rightarrow Y$.[1]

3 necessary conditions for Z to confound the relation $E \rightarrow Y$:

- Z must be a risk factor for Y; more precisely:
  - at the reference level of E, i.e. within "exposure-negative subjects" (because the risk must not be caused by a link with E)
- Z must be associated with E in the source population; specifically,
  - cohort study: Z and E must be associated at the start of the follow-up period,[2]

[1] Note that a confounder is always tied to both outcome and exposure.

[2] If E is constant during follow-up, this can be assessed by the unconditional association between Z and E in the data.

  - ∗ case-control study: Z and E must be associated in the controls,[3]

- Z must not be affected by E (which would make Z an *intervening variable* between E and Y), and Z must not be an effect of Y.

Definition of confounding:

- mathematically not easy; best attempt uses counterfactual arguments, but often infeasible in practice,
- literature agrees (and focuses) on *necesssary* (but not sufficient) conditions for confounding (previous page).

Pragmatic solution: define confounding of Z for the relation $E \rightarrow Y$ as present when both the conditions (i)−(ii) below are met:

(i) Z meets the 3 necessary conditions to be a confounder,

(ii) the difference between a crude ("total") measure of association/effect[4] and a confounding-adjusted measure of association/effect[5] is "substantial", i.e.[6]

  - ∗ a bias above 20−30% (arbitrary cut-off set in VER) measured relative to the crude estimate.

[3] If there is no selection bias (for E and Z) in the control population, the association in the source population can be estimated based on the controls alone.

[4] In regression models: a model with X as predictor for Y but Z not included.

[5] In regression models: a model with both X and Z as predictors for Y.

[6] Rothman *al* (2008), p. 262, note that usually 50% would be considered substantial, and 5% would not...