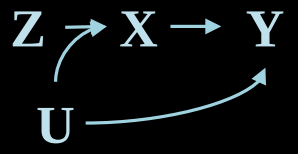
Epi 265: Introduction to Instrumental Variables Analyses: Intuition, Limitations, Applications, and Extensions

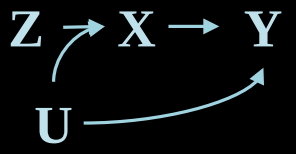
Many slides adapted from Stefan Walter & Eric Tchetgen Tchetgen SER 2014/2015 workshops

Outline



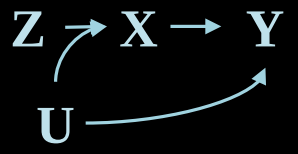
- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications

Motivation



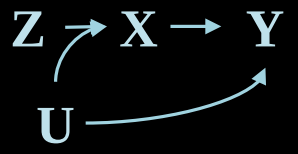
- Studies in the social and health sciences commonly aim to determine the causal effect of a point exposure.
- Although the double-blind randomized study design remains the gold standard for unbiased evaluation of the effects of an exposure, observational studies are often conducted for practical or ethical reasons.
- The main challenge with drawing causal inferences from observational studies stems from their inability as a study design, to categorically rule out the possibility that differences in outcome measures between exposed and unexposed persons, may be due to systematic background differences in selecting exposure status, that also predict the outcome.

Motivation



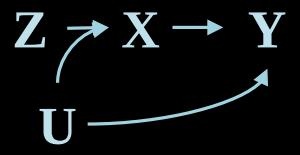
- Such confounding can compromise conclusions drawn from an observational study, but may likewise operate in a randomized trial with non-compliance, where individuals may choose for reasons unknown to the investigator, not to adhere to their treatment assignment.
- The development of methodology to adequately address this issue remains a priority for several disciplines, including biostatistics, epidemiology, econometrics and sociology.

Introduction

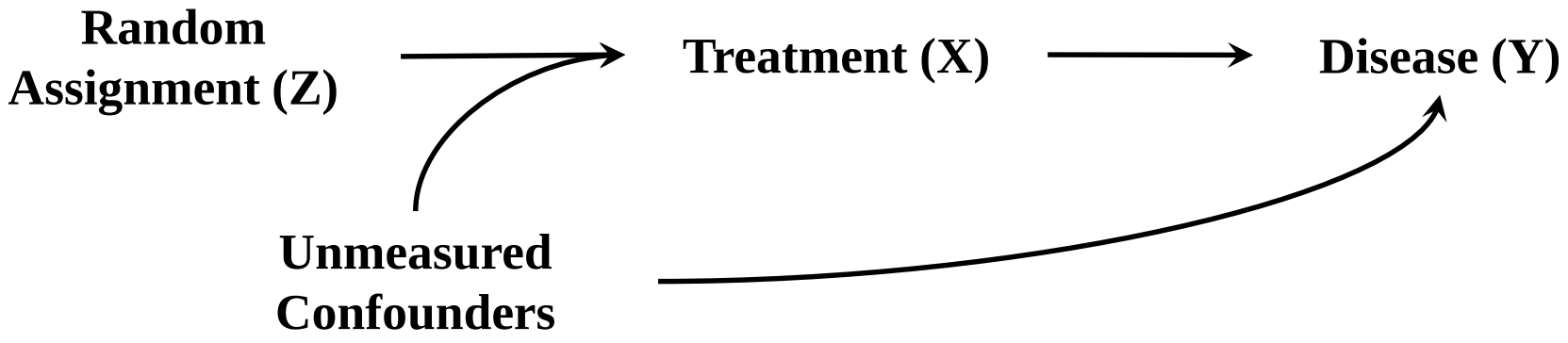


- The instrumental variable (IV) approach refers to a particular set of methods that allow one to recover, under certain assumptions, a causal effect of an exposure in the presence of unmeasured confounding.
- The IV approach has a longstanding tradition in econometrics going back to the original work of Wright (1928) and Goldberger (1972) who initially developed the approach in the context of linear structural equation modeling
- These ideas have more recently been formalized using potential outcomes or counterfactual variables, by Angrist (1994), Robins (1994), Angrist, Imbens and Rubin (1996) and Heckman (1997).
- A key contribution of the counterfactual language to the IV approach is that it allows one to formally define the causal effect of interest and to clearly articulate assumptions needed to identify this effect.

Instrumental Variables Analyses:
leverage (pseudo-)randomization



Use randomization as an *instrument* to estimate the effect of an exposure/treatment on the outcome.

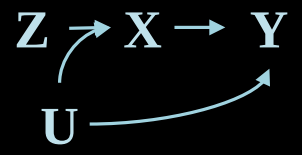


Instrumental variables (IV) analysis can be used to estimate the effect of X on Y when this causal structure applies – even if there is an unmeasured common cause of X and Y. Key in this causal structure is the "exclusion restriction"

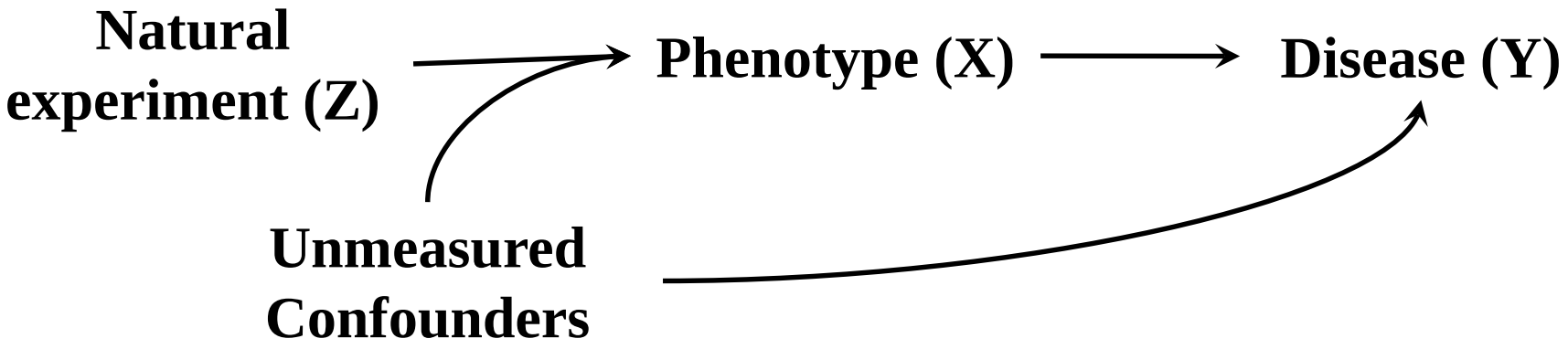
IV is appealing when causality is uncertain:

1) Unavoidable confounding
2) Reverse causation
3) Uncertainty: biomarker/endophenotype or cause
4) Effect "at the margins" is of greatest interest

Instrumental Variables Analyses:
leverage (pseudo-)randomization



Use pseudo-randomization as an *instrument* to estimate the effect of a phenotype on the outcome.



Example instruments:

- Before/after policy change, e.g., labeling rules, pharmacy rules, especially if not implemented universally
- Physician preference
- Local area treatment patterns
- Distance to service provider
- Characteristic that makes patient ineligible for treatment

Intuition for IV for Epidemiologists: RCT

- Randomization can be thought of as a special case of an instrumental variable:

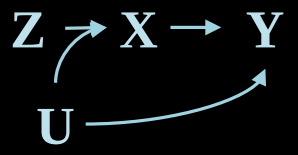
Random assignment→X→Y

  - We wish to test whether X affects Y
  - We randomly assign people to receive treatment or exposure X
  - We compare the outcome Y across levels of randomization (ITT), rather than across levels of exposure or take-up
  - With imperfect compliance, we assume ITT is an underestimate of the causal effect of X on Y

Intuition for IV for Epidemiologists: RCT

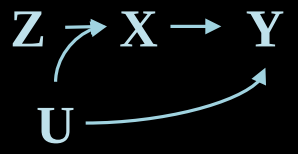
- Randomization can be thought of as a special case of an instrumental variable
- But IVs can arise from natural experiments, e.g., policy changes
- Controversy over whether, compared to conventional associational studies IVs are:
  - Stronger- "natural randomized trial"
  - Weaker – "exclusion restriction is implausible"
  - Complementary – "triangulation strengthens causal arguments"
- My view: case by case
  - Education and dementia: prefer IV with CSLs

Intuition for IV



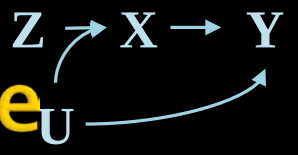
- The ITT is a valid test for the null hypothesis of no average causal effect of X on Y if:
  - There is at least some take-up (randomization affects exposure)
  - Randomization is fair (no common cause of randomization and the outcome)
  - Randomization influences the outcome *only* via the treatment X (not via related treatment X′ or via compensatory pathways in the controls)
- These criteria for a valid RCT correspond exactly with the criteria for a valid IV/MR analysis

Outline



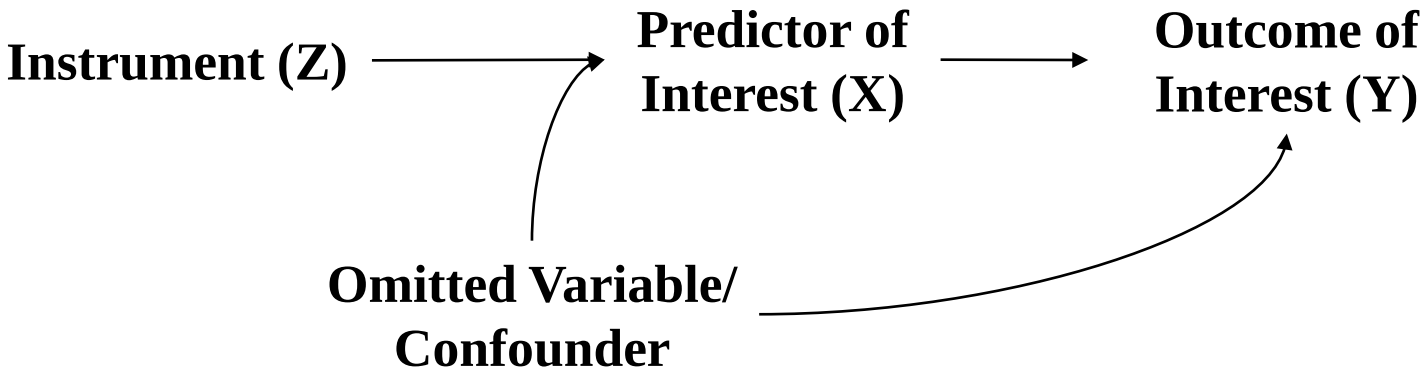
- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications

Instrument Assumptions: in intuitive terms

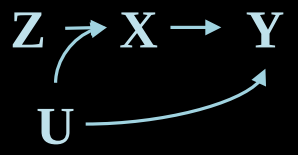


Z is an instrument for the influence of X on Y if:

1. Z predicts X
2. Z has no effect on Y unless the effect is mediated by X
3. No other variables influence both Z and Y
4. X does not affect Z (redundant but people are often confused about this)

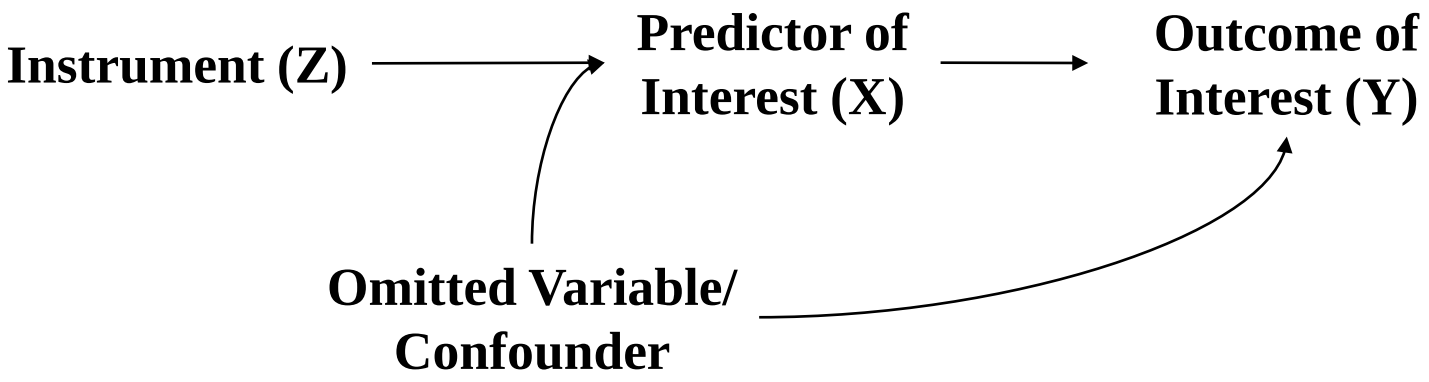


Instrument Assumptions: in DAG Terminology

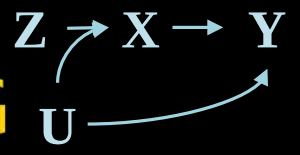


Z is an instrument for the influence of X on Y if:

- Z is not an effect of X
- Z is not independent of X (typically: Z affects X)
- Every unblocked path between Z and Y contains an arrow pointing into X

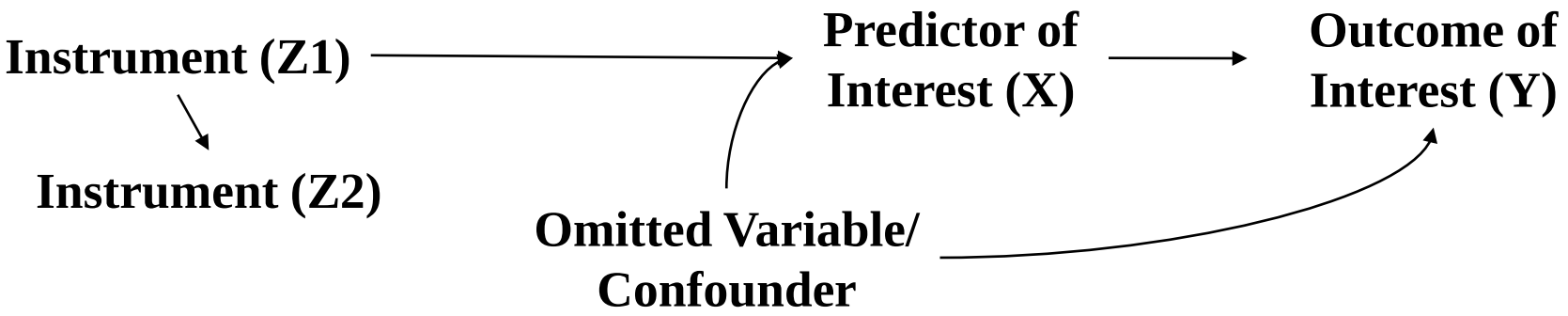


Instrument Assumptions: alternative DAG (more relevant for Mendelian Randomization)

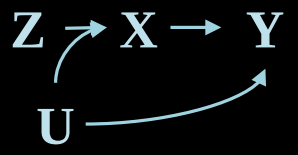


Z is an instrument for the influence of X on Y if:

- Z is not an effect of X
- Z is not independent of X (typically: Z affects X)
- Every unblocked path between Z and Y contains an arrow pointing into X



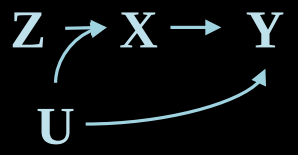
D-separation



- A path between two variables is blocked by conditioning on a set of variables {C} if:
  1. The path contains a non-collider that is in {C}, or
  2. The path contains a collider which is not in {C} , and no descendent of the collider is in {C} .

- If there is an unblocked path linking X and Y, then X and Y will typically be statistically dependent (unless there is a perfectly offsetting balance between two paths).

IVs "discovered" multiple times: multiple definitions



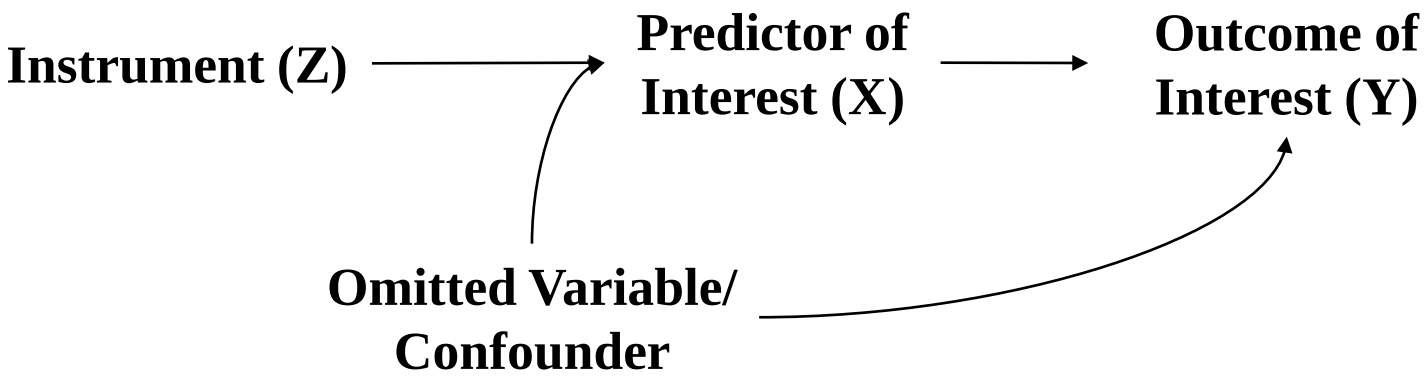
From Pearl *Causality*

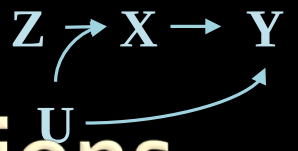
Economists:

Z is an instrument relative to the pair (X,Y) if (i) Z is independent of all error terms that have an influence on Y that is not mediated by X and (ii) Z is not independent of X

Counterfactual:

Z is independent of Y_X but not independent of X



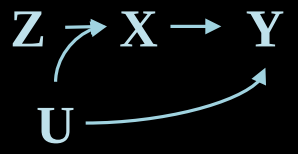


Classic Econometrics IV Assumptions

$Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$

This [instrument] is a new independent variable which must have two characteristics. First, it must be contemporaneously uncorrelated with the error; and second, it must be correlated (preferably highly so) with the regressor for which it is to serve as an instrument.

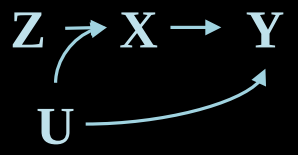
Angrist Imbens Rubin IV Assumptions



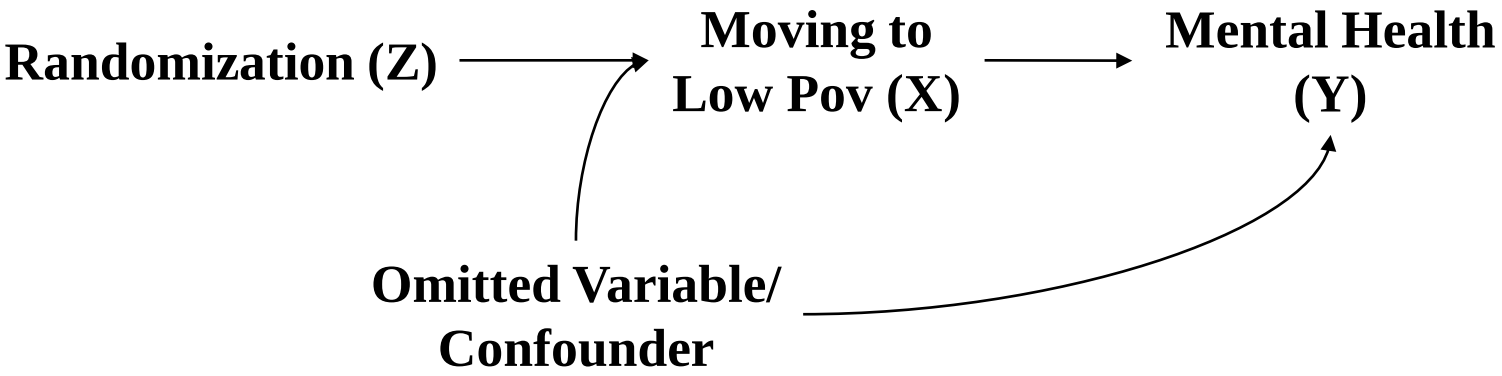
AIR propose a causal interpretation of an IV estimate as the "effect of the exposure on the responders" under:

- Randomization of instrument (=no common causes)
- Exclusion restriction (=no diect pathways)
- Nonzero average causal effect of G on X
- SUTVA
  - If $G_i=G'_i$ then $X_{iG}=X_{iG'}$
  - If $G_i=G'_i$ and $X_i=X'_i$ then $Y_{iG,X}=Y_{iG',X'}$
- Monotonicity (more later)

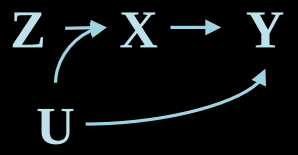
Motivating Example: Moving To Opportunity



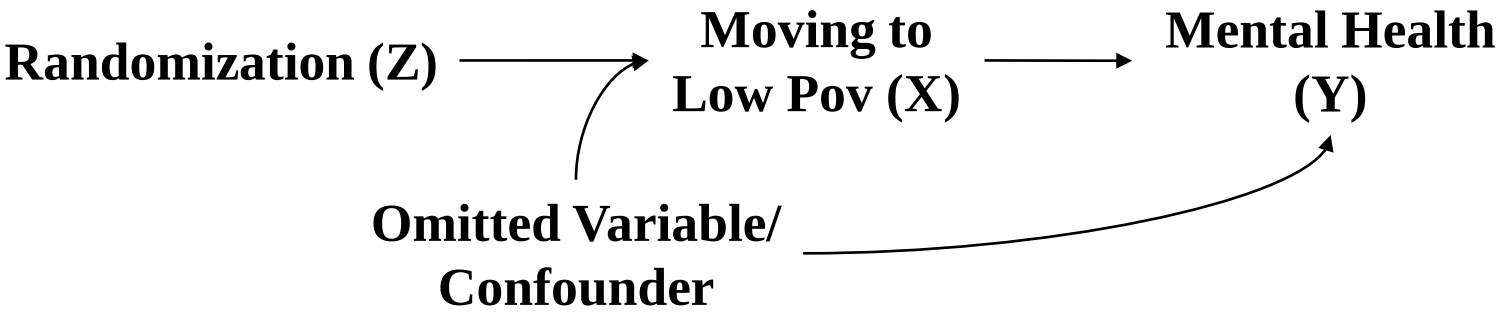
- Causal question:
  - Does moving from very high poverty public housing developments benefit the health of mothers or their children?
- A randomized experiment: Moving To Opportunity (MTO) study:
  - Z =families randomized to receive a voucher to alleviate the cost of moving from high poverty to low poverty neighborhood vs. no voucher.
  - X = Moved to low poverty neighborhood
  - Outcome Y:Psychological distress of the kids measured after 10 yrs of follow-up
  - Since X not randomized, cannot rule out U unobserved confounding effects of moving from high poverty to low poverty neighborhood



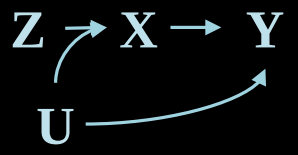
Motivating Example: Moving To Opportunity



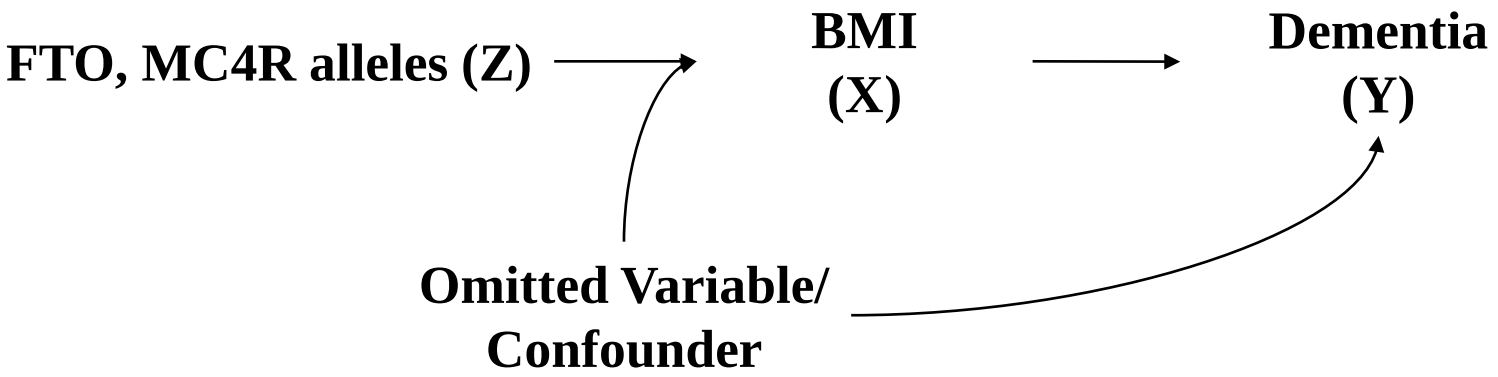
- Intent to treat analysis: Ignore X and focus on Z-Y relation.
- Preferred approach in most placebo controlled RCT because under the sharp null of no causal effect of the intervention, the type 1 error rate is guaranteed.
- In this example, only test for an association between assignment to receive a voucher and psychological distress
- Per protocol analysis compares only treated and untreated among those who complied to their assigned treatment, ie the effect of Z on Y restricted to the sample with X=Z.
- As treated analysis ignores the randomization and focuses on the association between X and Y.
- Per protocol and as treated may be subject to confounding bias- no improvement on usual methods.



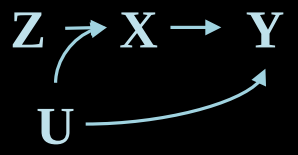
Example: Mendelian Randomization Study of BMI and Dementia



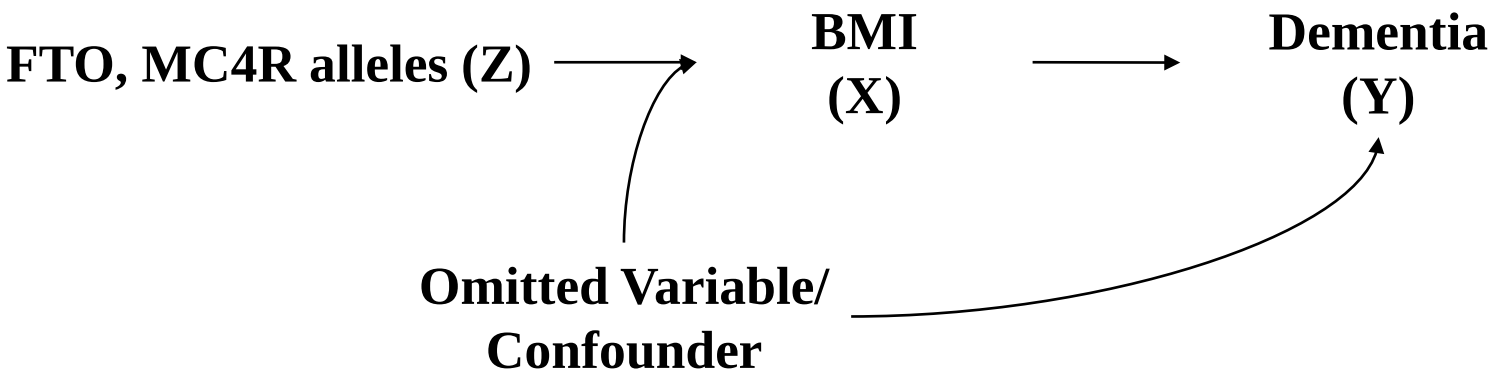
- Causal question:
  - Does increasing BMI increase risk of dementia?
- A Mendelian randomization study:
  - Z =polygenic risk score composed of genes linked to BMI
  - X = BMI
  - Outcome Y:Dementia diagnosis by age 80
  - Since BMI is not randomized, cannot rule out U unobserved confounding effects that influence BMI and dementia risk, e.g., education



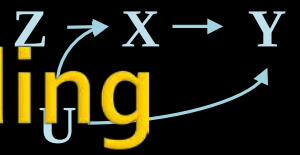
Example: Mendelian Randomization Study of BMI and Dementia



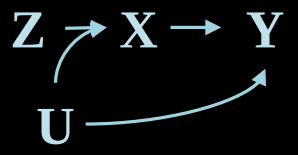
- Causal question:
  - Does increasing BMI increase risk of dementia?
- Example entails special considerations (more later)
  - What age is the BMI variable represented in this DAG?
  - Does BMI fulfill the "consistency" assumption for causal inference?
  - Pleiotropic pathways (corollary for the MTO instrument)
  - Montonicity assumption (hold on that one)



Examples of IVs and Corresponding DAGs?

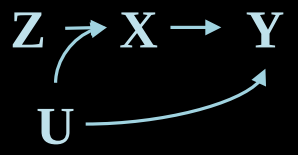


Outline



- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications
- Estimation approaches- control function
- IV for a binary or survival outcomes
- Inverse variance weighted IV
- Coding examples in R

Remember the assumptions?



Z is an instrument for the influence of X on Y if:

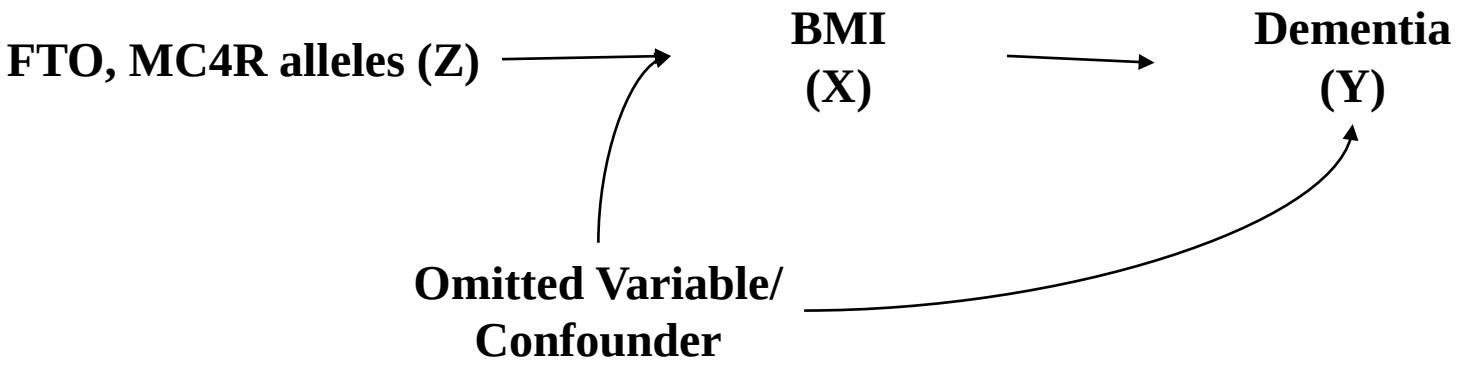
1. Z is not an effect of X
2. Z is not independent of X (typically: Z affects X)
3. Every unblocked path between Z and Y contains an arrow pointing into X

#3 cannot be verified. The assumption implies:

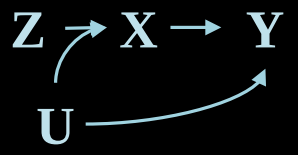
$Y \perp\!\!\!\perp Z \mid U, X$

But also:

$Y \not\perp\!\!\!\perp Z \mid X$



Remember the assumptions?



Z is an instrument for the influence of X on Y if:

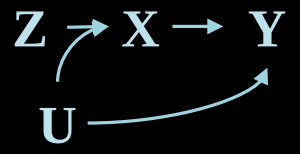
1. Z is not an effect of X
2. Z is not independent of X (typically: Z affects X)
3. Every unblocked path between Z and Y contains an arrow pointing into X

We also need another assumption – an assumption that cannot be expressed in the DAG because it relates to heterogeneity of treatment effects - in order to provide a specific causal interpretation to the IV estimate, but you have some choices about what you use for this last assumption.

Most common additional assumption:

4. Monotonicity: The effect of Z on X does not have the opposite sign for any two people in the sample. If Z increases X for some people, it must not decrease X for anyone else.

Obtaining IV Effect Estimates

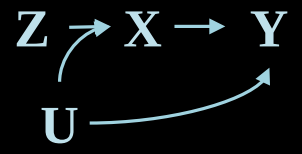


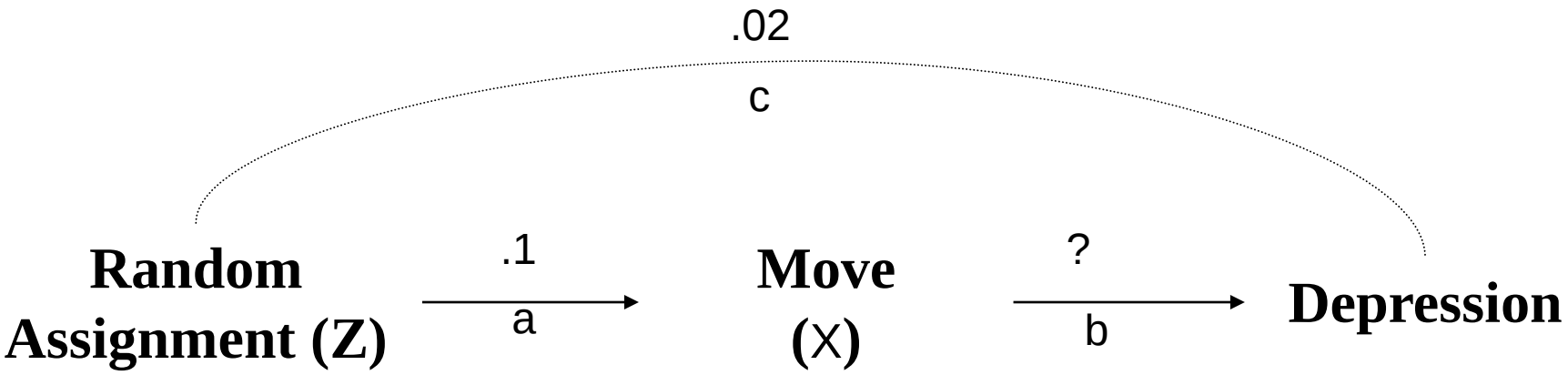
Random Assignment (Z) —.1→ (a) **Move (X)** —?→ (b) **Depression**

Wald Estimates:

$$\text{IV Estimate} = \frac{\Pr(\text{Depression}|Z=1)-\Pr(\text{Death}|Z=0)}{\Pr(\text{Move}|Z=1)-\Pr(\text{Move}|Z=0)}$$

Obtaining IV Effect Estimates

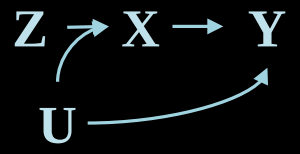


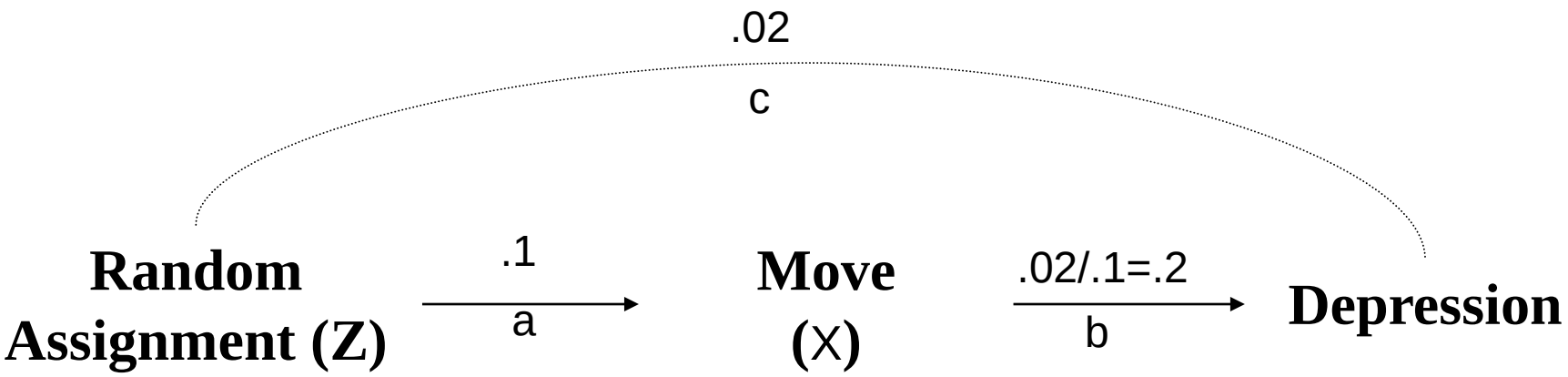


Wald Estimates:

$$\text{IV Estimate} = \frac{\Pr(\text{Depression}|Z=1)-\Pr(\text{Death}|Z=0)}{\Pr(\text{Move}|Z=1)-\Pr(\text{Move}|Z=0)}$$

Obtaining IV Effect Estimates

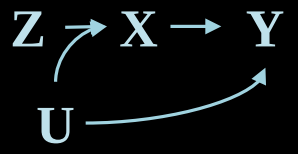




Wald Estimates:

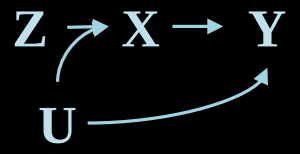
$$\text{IV Estimate} = \frac{\Pr(\text{Depression}|Z=1)-\Pr(\text{Death}|Z=0)}{\Pr(\text{Move}|Z=1)-\Pr(\text{Move}|Z=0)}$$

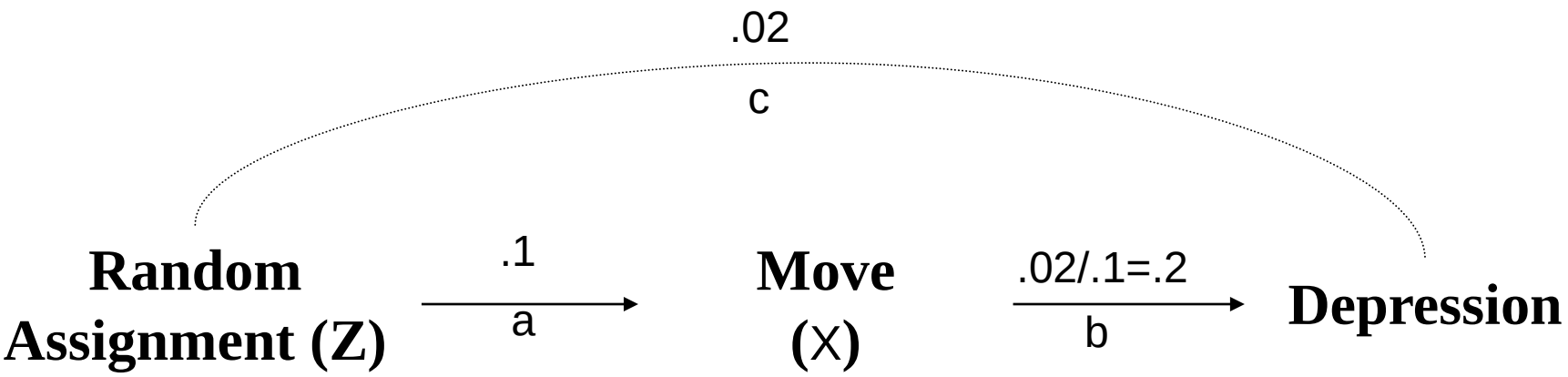
Outline



- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications

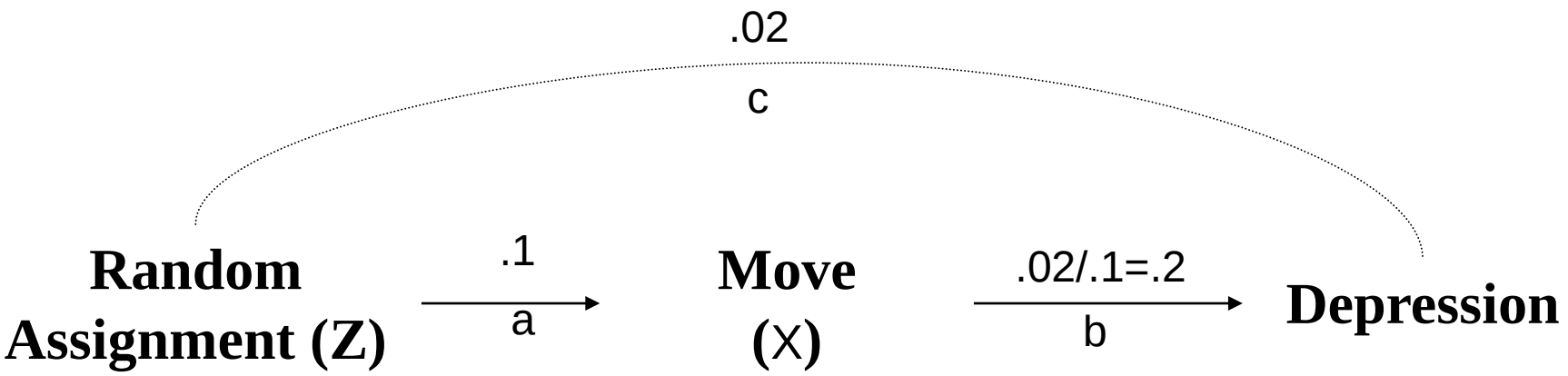
Obtaining IV Effect Estimates





IV developments focus on binary variables (instrument yes/no; phenotype present/absent; disease present/absent), so conventional estimator is based on risk differences but correlations are fine if you assume everything is linear: c=a*b and therefore b=c/a

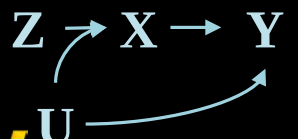
Extend to continuous variables, with linear models



$E(Y| X, U, Z)=\beta_0+\beta_x X+\beta_u U+\varepsilon_y$
Where ε_y is independent of (X,U)

Note that Z is not in the right hand side of the equation. By assumption, Z is not a determinant of Y except via X (the exclusion restriction) thus $E(Y| X, U, Z)= E(Y| X, U)$

Extend to continuous variables, with linear models



$E(Y|X, U, Z)=\beta_0+\beta_x X+\beta_u U$
Where ε_y is independent of (X,U)

Note that Z is **not** in the right hand side of the equation. By assumption, Z is not a determinant of Y except via X (the exclusion restriction), thus $E(Y|X, U, Z)= E(Y|X, U)$

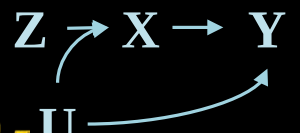
But consider how to write Y as a function of Z (replace X with the formula for X as a function of Z)
U is unrelated to Z but recall Z is a cause of X, say:
$E(X|Z)=\alpha_0+\alpha_x Z$

Substitute the equation for X as a function of Z instead of X itself.

$E(Y|Z) = E(\beta_0+\beta_x(\alpha_0+\alpha_x Z)+\beta_u U)$
$= E(\beta_{new}+\beta_x\alpha_x Z+\beta_u U)$

The coefficient for the effect of Z on Y is the product of the effect of Z on X and the effect of X on Y.

Extend to continuous variables, with linear models

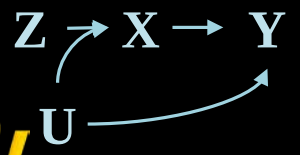


$$E(Y|Z) = E(\beta_0+\beta_x(\alpha_0+\alpha_x Z)+\beta_u U)$$
$$= E(\beta_0+\beta_x\alpha_x Z+\beta_u U)$$
$$= E(\beta_0+\gamma_Z Z)$$

Under this perfectly linear model, if Y is estimated as a function of Z, the effect of Z on Y is $\gamma_Z=\beta_x\alpha_x$, so if you estimate α_x (by regressing X on Z), then $\beta_x=\gamma_Z/\alpha_x$

This should be making you think about how we decompose a total effect into a direct and indirect effect. With IV, we assume that the total effect of Z on Y equals the indirect effect of Z on Y (look at the DAG). If we were using the product method, how would we estimate the indirect effect of Z on Y? First, the effect of Z on X; then the effect of X on Y.
In IV, we impose a strong assumption of full mediation and then we work backwards to estimate the effect of X on Y given that assumption.

Extend to continuous variables, with linear models



U is unrelated to Z but recall Z is a cause of X, say:

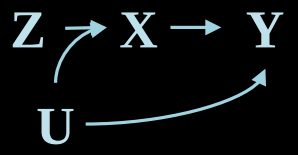
$E(X|Z)=\alpha_0+\alpha_x Z$

$$E(Y|Z) = E(\beta_0+\beta_x(\alpha_0+\alpha_x Z)+\beta_u U)$$
$$= E(\beta_0+\beta_x\alpha_x Z+\beta_u U)$$
$$= E(\beta_0+\gamma_Z Z)$$

Under this perfectly linear model, if the effect of Z on Y is $\gamma_Z=\beta_x\alpha_x$, so if Z), then $\beta_x=\gamma_Z/\alpha_x$

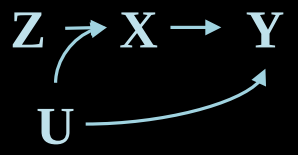
The regression of Y on Z is sometimes called the "reduced form"
Note that γ_Z is also the ITT effect estimate.

Linearity assumptions



- The linear models above make strong homogeneity assumptions.
- The linearity assumptions made in the previous slide are often overly restrictive, fortunately, these assumptions are not necessary and one may consider alternative assumptions, which may lead to identification of alternative causal effects.
- The assumption of monotonicity (detailed more later) is sufficient for nonparametric identification of the LATE or CACE.
- Separately, Robins specified assumptions that would support interpretation of the IV as effect of treatment on the treated.

Obtaining IV Estimates



Wald Estimates:

$$\text{Effect} = \frac{\text{(Difference in Avg Exposure between IV)}}{\text{(Difference in Avg Exposure between IV)}}$$

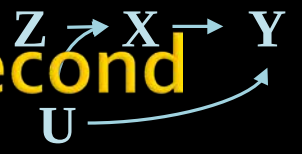
Two Stage Least Squares Estimates:

$E(X)=\beta_0+\beta_1 IV + \beta_k$ Other Covariates

Outcome$=\gamma_0+\gamma_1 E(X) + \gamma_k$ Other Covariates

γ_1 is a consistent estimate of the causal effect (e.g., LATE, ATE, ETT, depending on assumptions)

Control Function Approach: Adjust the second stage for the residuals from the first stage



(Tchetgen Tchetgen and Vansteelandt, 2013)

- Assume X is binary and estimate:
- $\Pr(X|IV, Covs)=\beta_0+\beta_1 IV + \beta_k$ Other Covariates
- Now estimate the residual:
  - X-Pr(X|IV, Covs)
  - The residual represents all the things that predict X other than the IV and the measured covariates
  - That includes the unmeasured confounders
- Now include the residual in the outcome model :
- $\text{Outcome}=\gamma_0+\gamma_1 X +\gamma_2 X*Z+\gamma_3 (X-\Pr(X|IV,Covs))+ \gamma_k$ Other Covariates

Variations on IV Estimators

Wald Estimate: $= \frac{\text{(Difference in Avg Outcome between IV values)}}{\text{(Difference in Avg Treatment between IV values)}}$

Two Stage Least Squares (2SLS):

- Multiple instruments
- Control for covariates

Separate Sample IV:

- 1st and 2nd stage of a 2SLS are from different data sets
- Often relevant in MR: assumes 2 samples are "equivalent"

- Generalized method of moments
- Residual control approaches

The role of covariates

Suppose that together with (X, Z, Y) we alo observe a vector C of all common causes of (X, Z, Y) , then the IV approach must account for C, in order to

- Avoid confounding if there is a common cause of Z and Y
- Explain variation in the outcome Y to improve efficiency

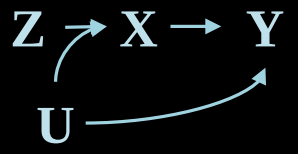
The role of covariates

Common approach based on assumed parametric model for the conditional average potential outcome among compliers with covariate C, e.g.,

$E(Y_x|x_{z=1}=1, x_{z=0}=0,C)=\beta_0+\beta_1*x+\beta_2*C$

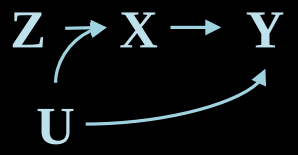
A linear structural equation model, describing counterfactual outcome among compliers

Outline



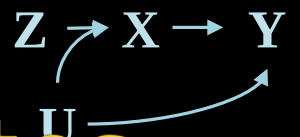
- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications

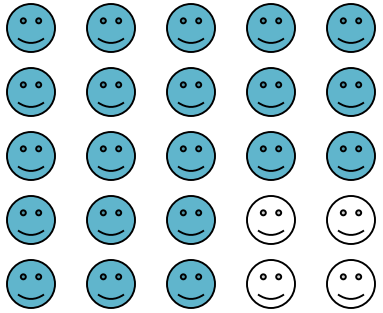
Testing the Null vs Estimating the Parameter



- Structural assumptions + faithfulness (no perfectly canceling pathways) enough to test use the IV to test the null
- More assumptions are needed to give a precise interpretation to the IV effect estimate.
- Often, testing the null is enough
  - Show instrument-outcome association (regress Y on Z)
  - Like an ITT, but effects will generally be quite small
  - Often have an instrument, but no measure of endogenous variable

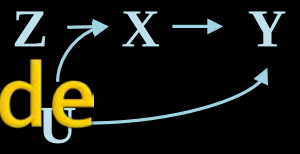
Interpreting IV Effect Estimates





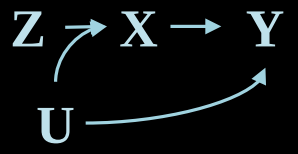
- Any given exposure may have different effects on different people in the population
- IV effect estimates are not necessarily the same as population average treatment effect (i.e., what would happen to the average value of Y if I treated everyone in the whole population).
- Alternative assumptions are required to support any specific interpretation of the IV estimate, the most common is "treatment effect on compliers".

Adopting a new assumption to provide an interpretation of the IV estimate



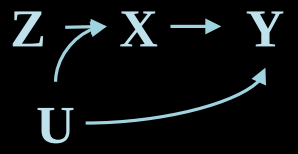
- Lots of alternative accounts of the interpretation of the IV estimate, resting on stronger or weaker assumptions
- Often, the "strong" assumptions are things we make in interpretation of estimates from other study designs (e.g. RCTs) without concern.
- Under linear, constant relationships, interpretation is easy: causal effect of X on Y.
- Angrist, Imbens, Rubin proposed a weaker set of assumptions and this framework has been widely adopted.

AIR Framework for Interpreting IV Estimate



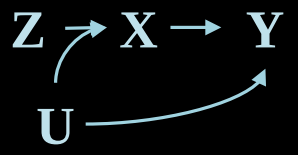
- 4 types of responses to two values of randomizing variable (Z=0 vs Z=1)
  1. Never has the exposure X=1(regardless of random assignment)
  2. Always has the exposure X=1 (regardless of random assignment)
  3. Has exposure X=1 *only* if has Z=1
  4. Has the exposure X=1 *only* if has Z=0

AIR Framework for Interpreting IV Estimate



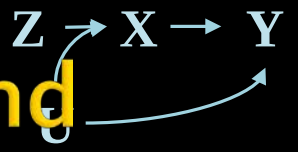
- 4 types of responses to two values of randomizing variable (Z=0 vs Z=1)
  1. Never has the exposure X=1(regardless of random assignment) never takers
  2. Always has the exposure X=1 (regardless of random assignment) always takers
  3. Has exposure X=1 *only* if has Z=1 compliers
  4. Has the exposure X=1 *only* if has Z=0 defiers/contrarians

AIR Assumptions (in addition to main IV assumptions)



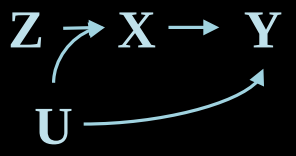
- No defiers: for every individual in the population, the gene either increases or has no effect on the probability of the phenotype.
- It must not increase the probability of the phenotype for some people but decrease it for others
- AIR called this monotonicity, relates to Z-X association, not X-Y association.
- Note power consequences: compliers may be only a small fraction of the sample, and compliers are all that contribute to the effect estimate

Interpretation Under "No Defiers" and Standard IV Assumptions



- The IV estimate is the effect of X on Y among those whose value of X was determined by Z.
- Sometimes called the local average treatment effect (LATE)
- Or the complier average causal effect (CACE)
- IV estimate may not equal the population average treatment effect
- In counterfactuals: $E(Y_{x=1}-Y_{x=0} | X_{z=1}=1, X_{z=0}=0)$

Whose Causal Effect?



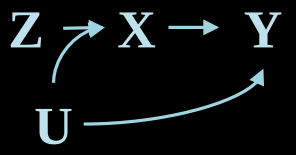
Classify people based on their treatment under either value of the IV/randomly assigned treatment.

What the person will do if assigned to experimental treatment:

What the person will do if assigned to control treatment:

	Take control	Take experimental
Take control	Never-Takers	Compliers
Take experimental	Contrarians/ Defiers	Always Takers

Whose Causal Effect?



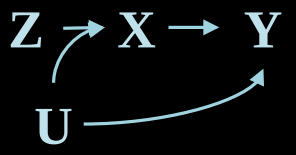
Classify people based on their treatment under either value of the IV/randomly assigned treatment.

What the person will do if assigned to control treatment:

What the person will do if assigned to experimental treatment:

	Take control	Take experimental
Take control	Never-Takers	Compliers
Take experimental	Contrarians/ Defiers	Always Takers

Whose Causal Effect?



Classify
on their t
either va
IV/rando
treatmen

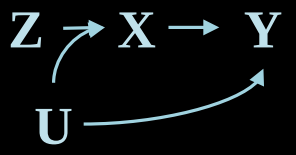
Never-takers do not contribute to any outcome differences between the IV=0 and IV=1 group.

What the person will do if assigned to experimental treatment:

What the person will do if assigned to control treatment:

	Take control	Take experimental
Take control	Never-Takers	Compliers
Take experimental	Contrarians/ Defiers	Always Takers

Whose Causal Effect?



Classify
on their t
either va
IV/rando
treatmen

Never-takers do not contribute to any outcome differences between the IV=0 and IV=1 group.

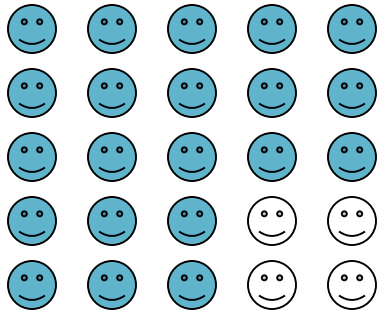
What the person will do if assigned to experimental treatment:

What the person will do if assigned to control treatment:

	Take control	Take experimental
Take control	Never-Takers	Compliers
Take experimental	Contrarians/ Defiers	Alw Tak

Always-takers do not contribute to any outcome differences between the IV=0 and IV=1 group.

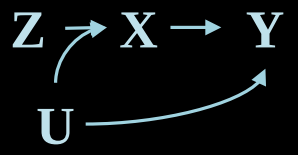
Interpreting IV Effect Estimates



- The effect of exposure (X) on outcome (Y) among those people whose exposure was determined by the IV.
- This may or may not be the same as the effect of the exposure on *other* people, much less the whole population

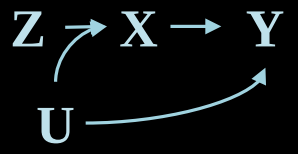
- It may not even be the same *sign* as the effect of the exposure on the whole population
- Impossible to know precisely whose exposure was determined by the IV
- But you *can* know what % of sample were compliers: E(X|Z=1)- E(X|Z=0)

Is the LATE useful?



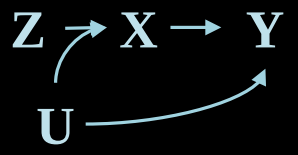
- The IV estimate is the effect of X on Y among those whose value of X was determined by Z.
- $E(Y_{x=1}-Y_{x=0} | X_{z=1}=1, X_{z=0}=0)$
- Criticized because one can never know who is in the group (you only see one value of exposure under the actual value of the IV)
- But sometimes of more interest than the PATE
  - Physician preference IV
  - Judge lottery IV
  - Other examples?

Alternative Assumptions You Might Prefer to Monotonicity



- Control group cannot access the treatment, i.e., $Pr(X_{Z=0})=0$
- Then monotonicity holds $X_{Z=0} <= X_{Z=1}$
- And the Wald estimand also identifies the effect of treatment on the treated (the ETT or TOT)
- In other words, if the controls cannot access treatment, can interpret the IV as the LATE or the ETT
- In MTO, most papers do this but it is a little sleight of hand.
- Define treatment not as "moving" but "moving with a voucher"

Another Alternative: No current treatment value interaction

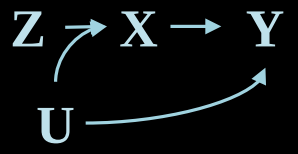


- I am not going to explain this because we would never reach the next slide. Know that this is an option.

- Effect of treatment is the same for exposed people with IV=1 as for exposed people with IV=0
- $E(Y_{X=1} - Y_{X=0} | X=1, Z=1) = E(Y_{X=1} - Y_{X=0} | X=1, Z=0)$

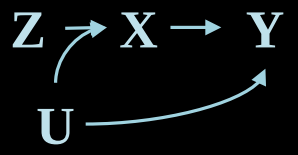
- Under structural assumptions and "no current treatment value interaction", then Wald estimand is the effect of treatment on treated
- Bonus contest: any better names for this assumption?

Is the ETT useful?



- Effect of treatment on the treated has a long tradition in epidemiology
- Sometimes it is the only relevant effect, i.e., the effect of exposure on those people who were actually exposed.
- In the case of a harmful exposure, this is the estimand the judicial court would care about.
- The advantage of the approach is that it does not require monotonicity: the relation between IV and exposure can be increasing for some individuals and decreasing for others.
- But does require ruling out the possibility of effect heterogeneity of exposure on outcome among treated.

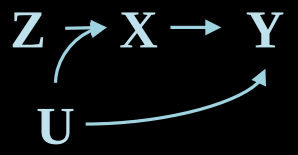
One more alternative: homogeneous selection bias



- Consider the contrast: $E(Y_{x=0}|X=1, Z=z)-E(Y_{x=0}|X=0, Z=z)$
- Note this would be zero if there were no confounding (recall the definition of confounding= association between the counterfactual values of Y given any particular value of X and the actual value of X received).
- The amount of confounding may differ in the active arm (Z=1) from the amount of confounding in the control arm (Z=0) that is:
- $E(Y_{x=0}|X=1, Z=1)-E(Y_{x=0}|X=0, Z=1)$

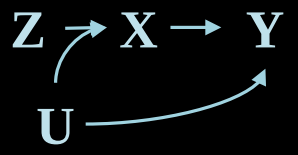
$$\neq E(Y_{x=0}|X=1, Z=0)-E(Y_{x=0}|X=0, Z=0)$$

One more alternative: homogeneous selection bias



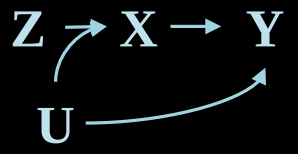
- An alternative assumption is that there is homogeneous selection bias (Tchetgen Tchetgen and Vansteelandt, 2013), so that:
-
- $E(Y_{x=0}|X=1, Z=1)-E(Y_{x=0}|X=0, Z=1)$
  $=E(Y_{x=0}|X=1, Z=0)-E(Y_{x=0}|X=0, Z=0)$
- ie degree of unobserved confounding is constant across values of Z.
- Under this assumption, the ETT is identifiable, but the Wald estimator does not produce it.

Take home re estimation and assumptions



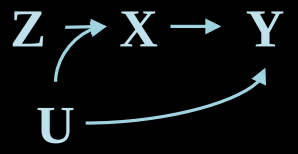
- You will mostly use 2SLS or a variation
- There are many options for the 4th 'non-structural' assumption and they imply slightly different interpretations.
- In most cases, this doesn't matter enough to worry about, but if you have to argue for your IV, you may want to think about preferable assumptions.

Outline

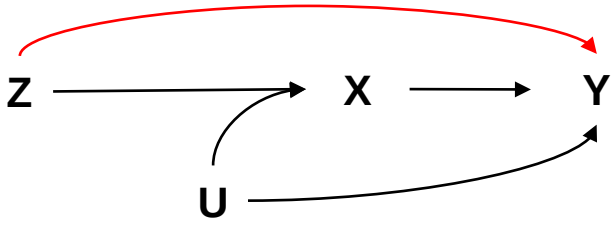


- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications

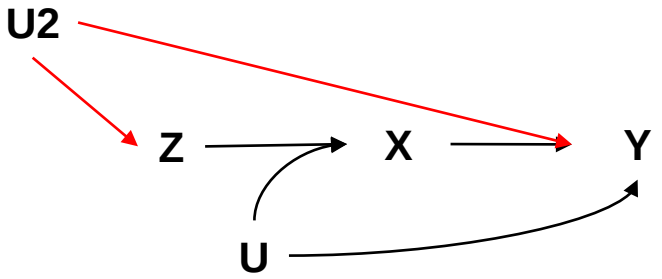
Doubting Instruments: major biases similar to critiques of RCTs



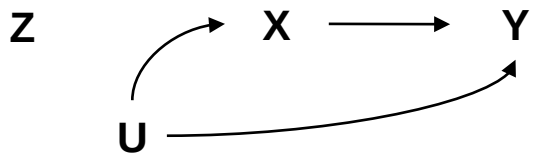
- Do they have other pathways to the outcome?
  - *Unblinded trials: controls become demoralized*



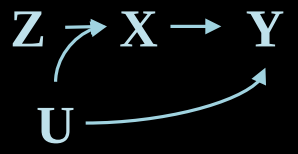
- Is there a common cause of the instrument and the outcome?
  - *Trials: unfair random assignment*



- Do they actually affect anyone's exposure?
  - *Trials: nobody adheres to assignment*



Violations of IV assumptions: No instrumental variable

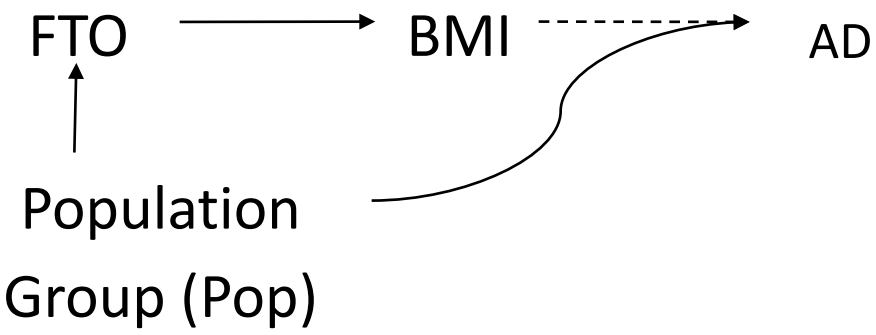


FTO → BMI ⇢ Dementia

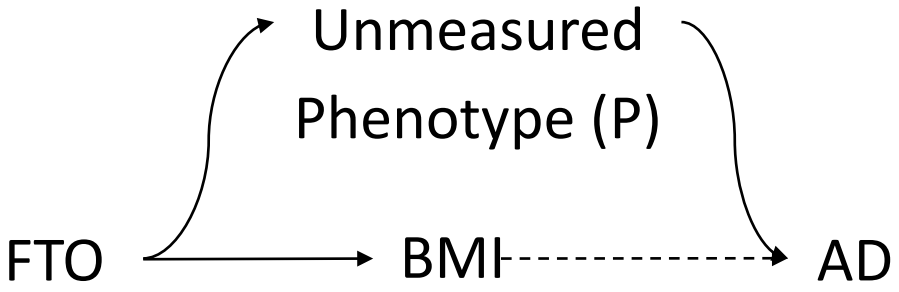
- No effect of the gene on the phenotype

- Bad news: can't use IV
- Good news: you know it right away
- Bad news: if you have 100,000 polymorphisms that don't affect the phenotype, IV will reproduce the conventional effect estimate (weak IV bias)

Causal Structures Violating IV Assumptions: Population stratification and Pleiotropy

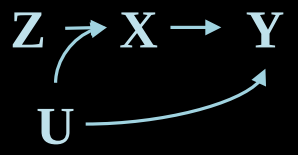


Population Stratification



Pleiotropy

IV Estimates



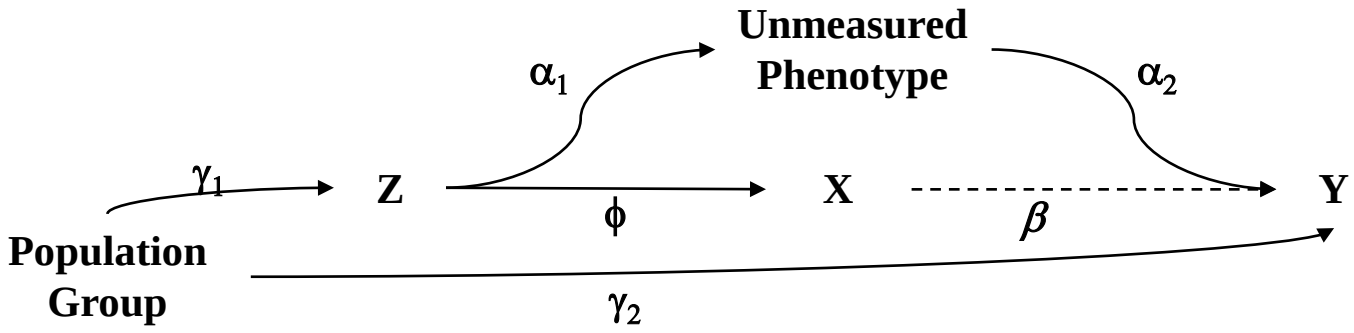
β_{MR}= (E(Y|Z=1)-E(Y|Z=0)) / (E(X|Z=1)-E(X|Z=0))

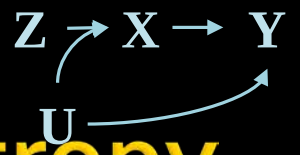
- If unbiased:

$\beta_{MR} = (\phi\beta) / (\phi) = \beta$

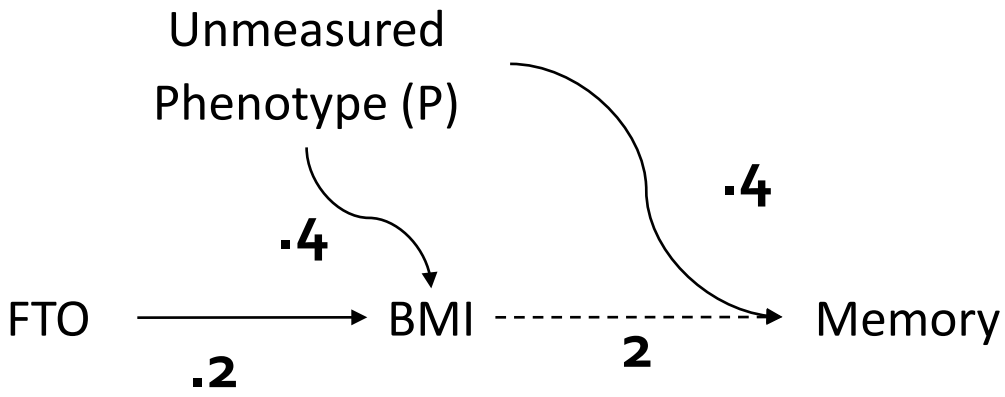
- If biased:

$\beta_{MR} = (\phi\beta+\gamma_1\gamma_2+\alpha_1\alpha_2) / \phi \neq \beta$

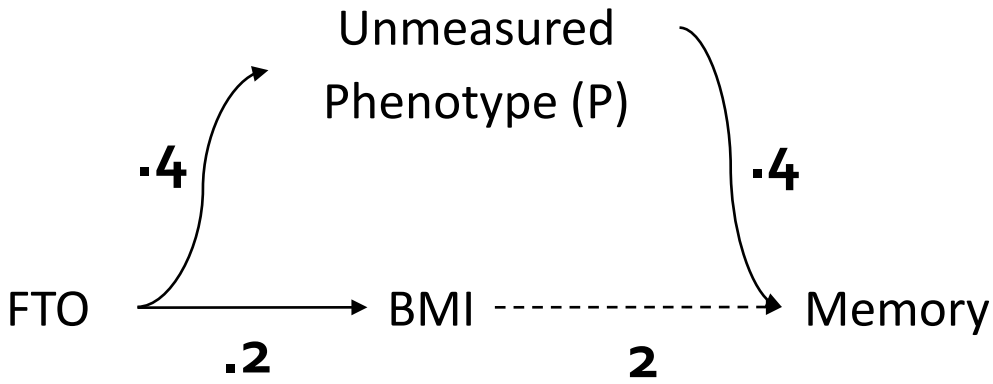




MR is sensitive to violations: pleiotropy

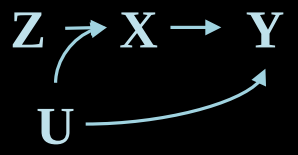


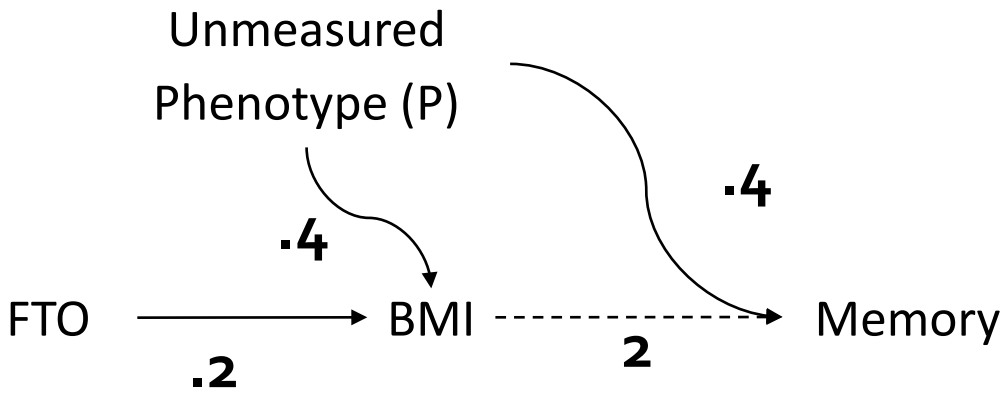
OLS estimate: 2.16



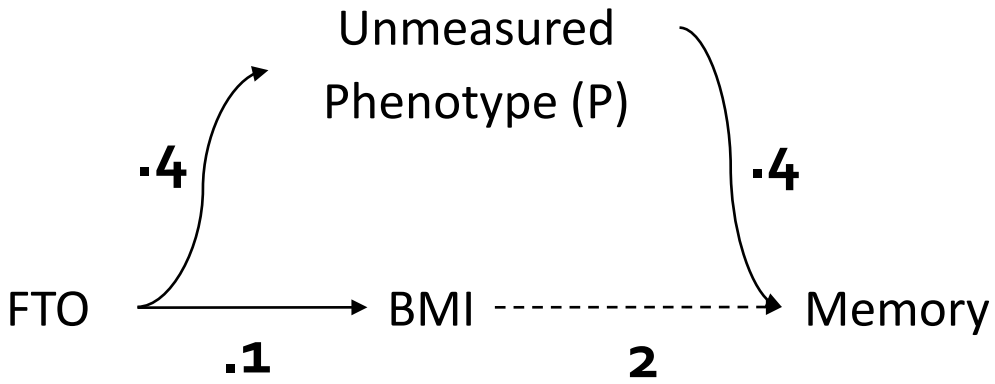
IV estimate: 2.8

MR is sensitive to violations: pleiotropy with a weaker instrument



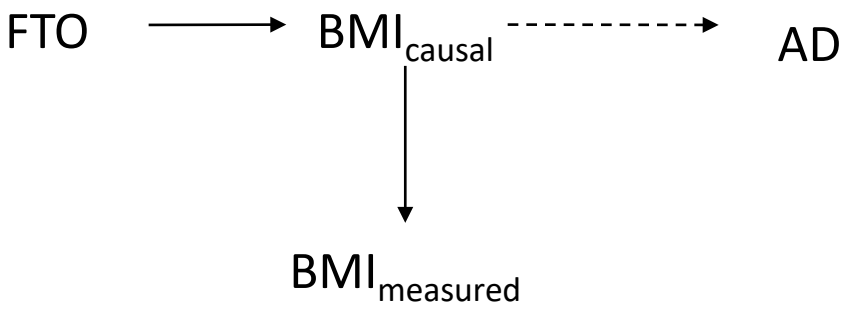


OLS estimate: 2.16

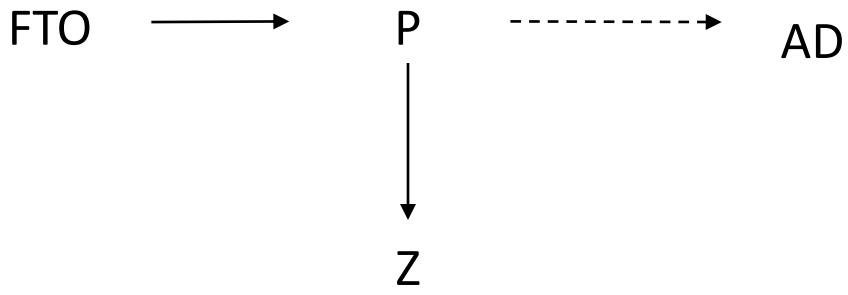


IV estimate: 3.6

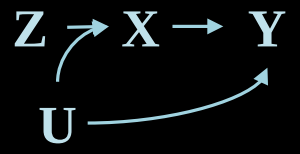
Violating IV Assumptions: mismeasuring the causal phenotype

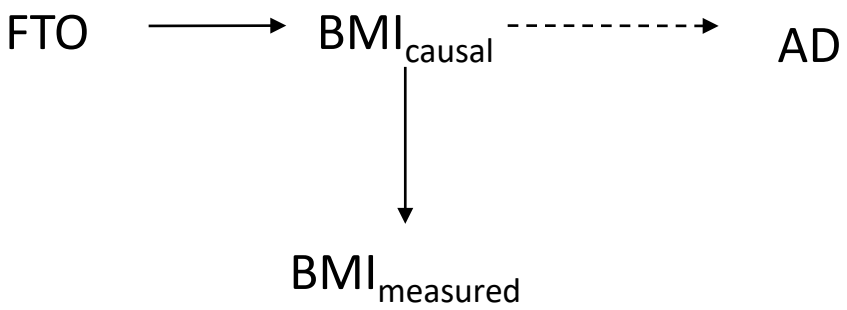


Structurally related to "pleiotropy": IV estimate of the effect of ***Z*** on AD corresponds to effect of ***P*** on AD only under special circumstances



Mismeasuring the causal phenotype



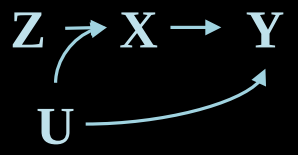


This structure will almost always be relevant in MR, because the gene affects lifelong values of the phenotype, but we measure at only one or a few moments.

Appropriate to test the null hypothesis: BMI has no effect on AD. Therefore, helpful to present the "ITT" estimate of the association between the gene and the outcome.

IV estimate with $BMI_{measured}$ can be the same as, larger, or smaller than the IV estimate with BMI_{causal}

Mismeasuring the causal phenotype

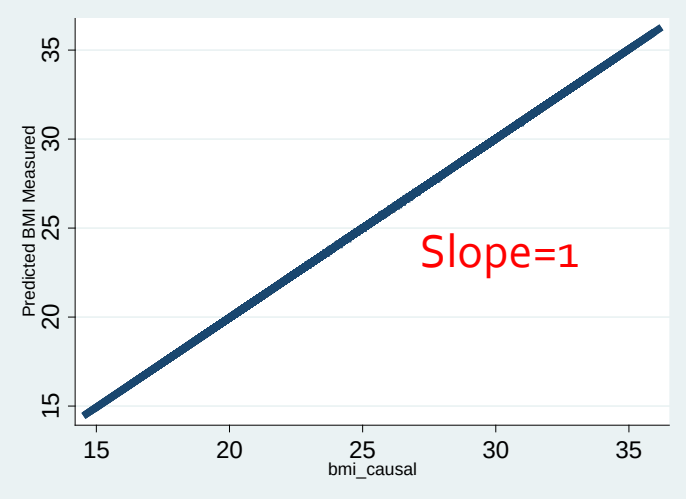


How does $IV_{measured}$ relate to IV_{causal}?

FTO ⟶ BMI_{causal} ⇢ AD

↓

$BMI_{measured}$

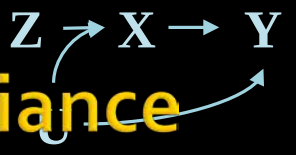


Classical measurement error:

$BMI_{measured} = BMI_{causal} + e$

$IV_{measured} = IV_{causal}$

Mismeasuring the causal phenotype: if variance of measured > variance of causal variable

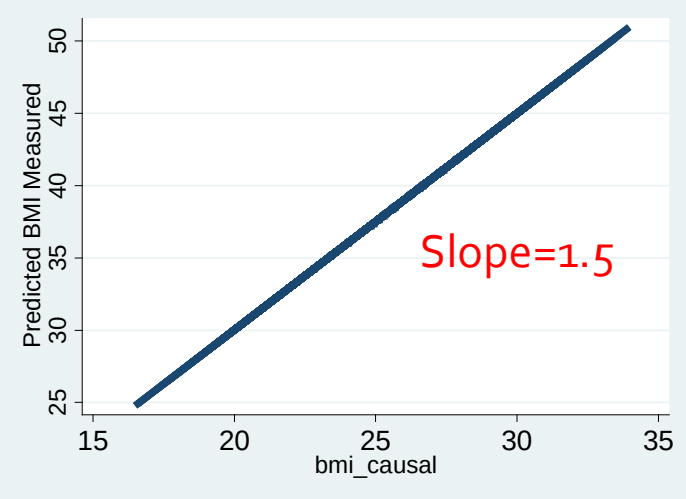


How does $IV_{measured}$ relate to IV_{causal}?

FTO ⟶ BMI_{causal} ⇢ AD

BMI_{causal} ↓

$BMI_{measured}$

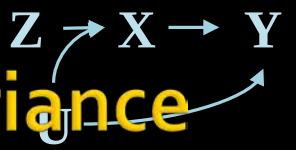


$BMI_{measured}=1.5*BMI_{causal}+e$

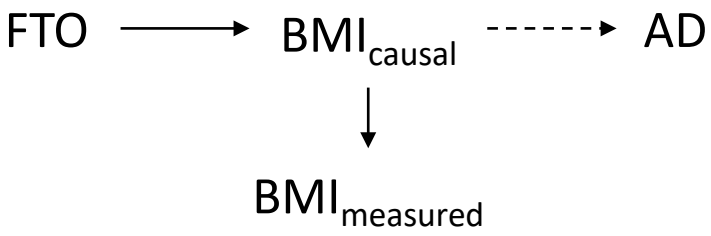
If the IV estimate with BMI_{causal} >0 then:

$IV_{measured} < IV_{causal}$

Mismeasuring the causal phenotype: if variance of measured < variance of causal variable



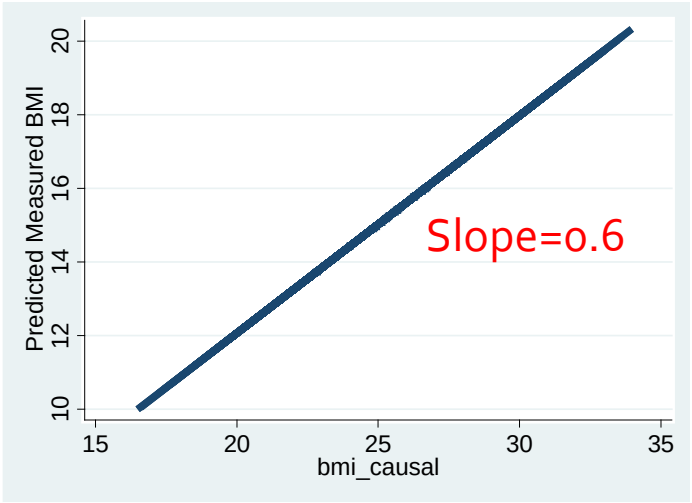
How does $IV_{measured}$ relate to IV_{causal}?



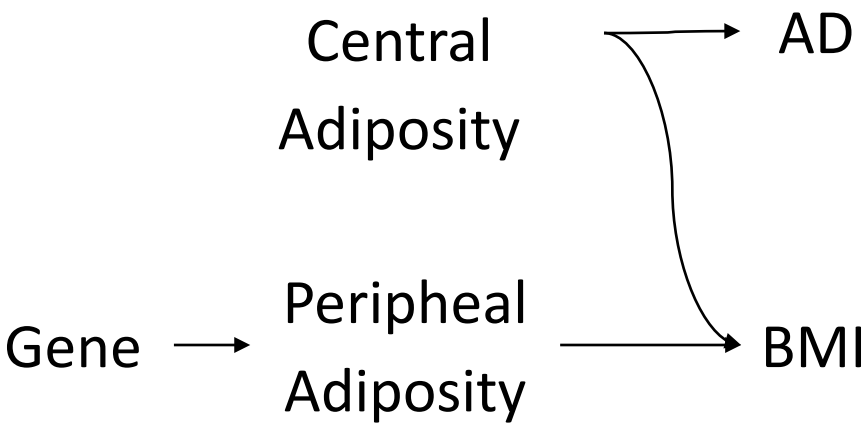
$BMI_{measured}=0.6*BMI_{causal}+e$

If the IV estimate with BMI_{causal} >0 then:

$IV_{measured} > IV_{causal}$



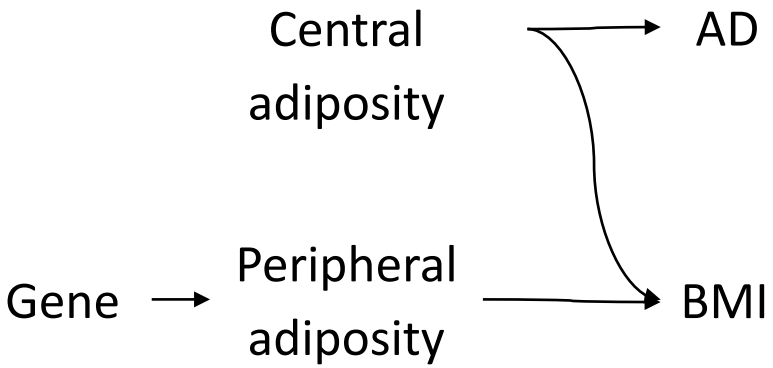
Causal Structures that (may) Violate IV Assumptions: multi-component phenotype



Phenotypes with multiple versions or components

- Gene predicts BMI
- No other pathways link the gene to AD
- A component of BMI (central adiposity) affects AD
- Gene will be independent of AD
- IV analysis will suggest the incorrect inference that BMI does not affect AD

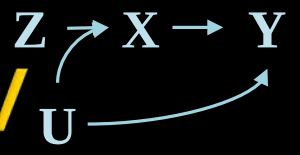
Causal Structures that (may) Violate IV Assumptions: multi-component phenotype

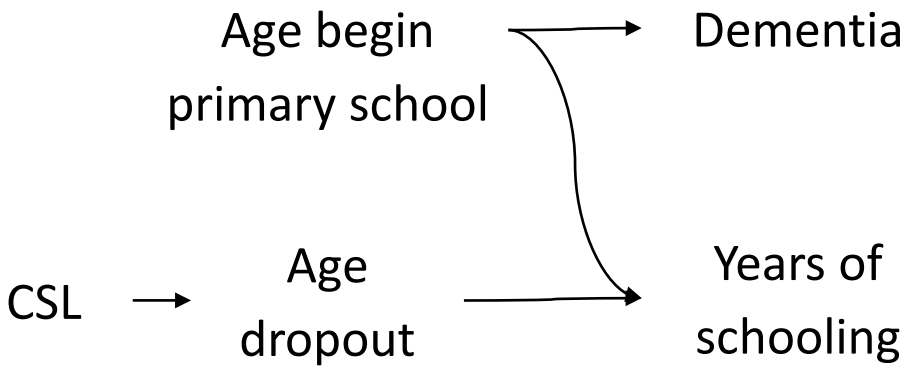


Phenotypes with multiple versions or components

- IV analysis will suggest the incorrect inference that BMI does not affect AD
- Correct inference: the phenotype affected by the gene (peripheral adiposity) does not affect AD.
- In theory, could use this to help identify causally relevant variations on the phenotype (i.e. is it important to distinguish between central or peripheral adiposity? How about knee fat and ankle fat?)

Causal Structures that (may) Violate IV Assumptions: multi-component phenotype

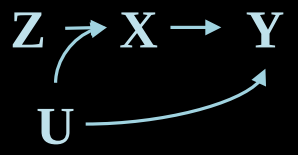




Exposures with multiple versions or components

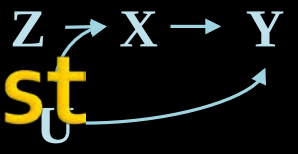
- Can also arise with non-genetic IVs.

Improving IV Studies

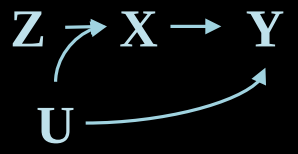


- Stronger IVs with larger effects on treatment
  - In MR: better genetic risk scores for phenotypes
  - In PP: larger practice effect differences, stronger encouragements
- Control for most plausible violations
- Sensitive tests for assumptions
- Sample size improves situation because tests of violations are informative and allows for more control variables.

Violation assumptions of greatest concern with your proposed IV?

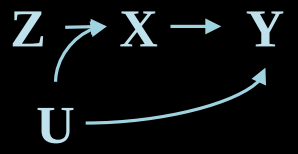


Outline



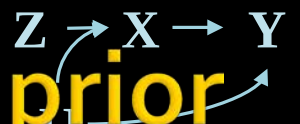
- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Promising applications

Evaluating the assumptions



1. Constraints implied by theory
2. Over-identification tests
3. IV inequality constraints
4. Stratification-based tests (similar to over-identification)
5. Negative controls
6. Independent from known confounders

1 Evaluating MR Studies: leverage prior assumptions about confounding

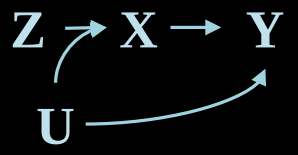


- Compare IV effect estimate to conventional effect estimate: if conventional effect estimate is positively confounded, IV<conventional
- Relies on assumption regarding the direction of confounding: if you don't know the direction, the test is not informative
- 4 *equivalent* versions of this test: no more convincing to show all 4.
- Often, field is only interested in knowing if a conventional effect estimate is biased up
- Other direction of bias not of interest, or not considered plausible
- Not guaranteed to be consistent in non-linear causal structures (but doesn't completely rely on linearity).

1 Evaluating MR Studies: leverage prior assumptions about confounding

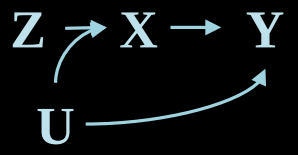
1. If the conventional estimate is biased upwards, and the IV is unbiased, the IV estimate should be smaller than the conventional estimate
2. If the IV effect is fully mediated by the phenotype, adjusting the regression of the outcome on the instrument for the phenotype should attenuate the effect estimate for the instrument
3. Adjusting a biased conventional estimate for a valid instrument *exacerbates* the bias
4. The residual between the predicted value of the outcome based on the IV estimate and the actual value of the outcome is correlated with the phenotype.

2 Evaluating IV Studies: Over-identification tests



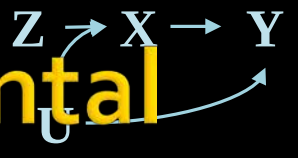
- Use multiple instrumental variables to conduct over-identification tests.
  - Physician preference versus regional preference versus formulary changes
  - In MR: Other genes.
- Cannot detect violations of the IV assumptions if all instruments have identical biasing pathways.
- May also reject even when all instruments are valid if the model is incorrectly assumed to be linear or the phenotype is composite. These tests generally have low statistical power.

Limitations of Over-ID Tests



- Power challenges: lots of uncertainty around 2 IV estimates will fail to reject
- May reject even if both instruments are valid, but have different subpopulations of "compliers", with different causal effects of X on Y (i.e. different LATEs)
- It's hard to get just 1 good instrument, much less 2 or 3.
- Data from GWAS may help - debatable

3 Evaluating IV Studies: instrumental inequality tests



- These tests are applicable only when the causal variable is known to be categorical (drug, no drug).
- Certain inequalities are impossible given the IV assumptions: if you see them, IV must not be right

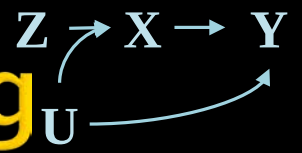
$$\max_i [\Pr(X=0, Y=1|G=i)] \leq \min_i [1 - \Pr(Y=0, X=0|G=i)]$$

$$\max_i[\Pr(X=0, Y=1|Z=i) + \Pr(X=1, Y=1|Z=i)] + \max_i[\Pr(X=0, Y=1|Z=i) + \Pr(X=1, Y=0|Z=i)] + \max_i[\Pr(X=0, Y=0|Z=i)] \leq 2$$

$$\max_i [\Pr(X=1, Y=1|Z=i)] \leq \min_i [1 - \Pr(Y=0, X=1|Z=i)]$$

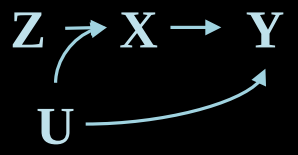
- Detects extreme violations of the assumptions
- With more instrumental variables, more opportunities for violations of tests.

4 Evaluating IV Studies: Modifying factors



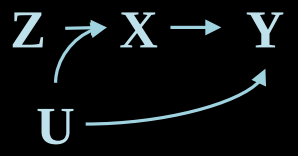
- Identify factors that modify the IV-treatment association.
- Compare the IV effect estimate across groups in which the population association between the instrument and the treatment is either silenced or reversed.
- This test could identify a biased instrument if the biasing pathway is active in both subgroups.

Stratification Tests: Gene Environment Interactions



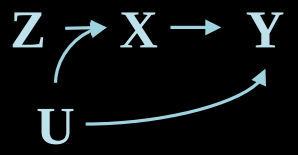
- An environmental context changes the relationship between the genotype and the phenotype (the first stage of the IV analysis).
  - "Silencing": genotype and phenotype are independent in some environments.
  - "Modifying": genotype changes the probability of the phenotype by a different extent in different environmental contexts.
  - "Reversing": genotype increases probability of the phenotype in some environmental contexts and *decreases* probability of the phenotype in other environments (violations of monotonicity).

How can this help with MR?



- Assumptions of MR are untestable when they stand alone.
- Combined with additional causal knowledge, these assumptions have testable implications.
- GxE interactions add new causal knowledge – may be able to tell whether associations are inconsistent with hypothesized MR-valid causal structure.

MR Stratified by Gene Silencing Environmental Characteristic



If unbiased:

E=1

Gene (G) $\xrightarrow{\phi}$ **Phenotype (X)** $\overset{\beta}{\dashrightarrow}$ **Health (Y)**

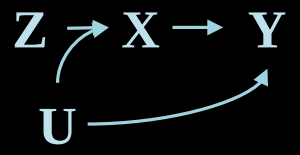
$\beta_{MR} = \phi * \beta / \phi = \beta$

E=0

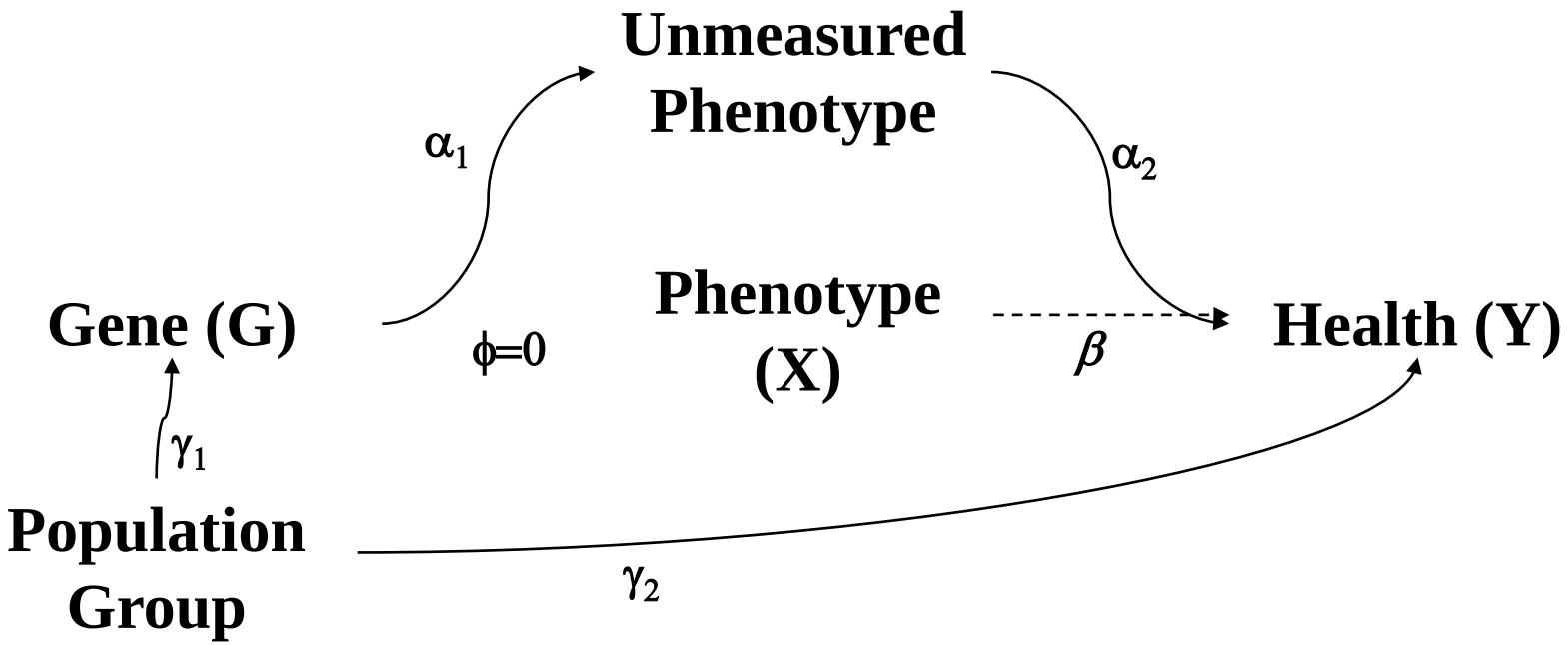
Gene (G) $\overset{\phi=0}{\not\longrightarrow}$ **Phenotype (X)** $\overset{\beta}{\dashrightarrow}$ **Health (Y)**

$\beta_{MR} = \phi * \beta / \phi = 0/0$

MR Stratified by Gene Silencing Environmental Characteristic

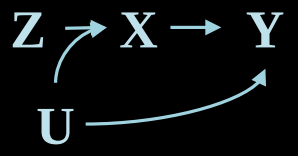


E=o (biased case)



$$\beta_{MR} = (\phi\beta + \gamma_1\gamma_2 + \alpha_1\alpha_2) / \phi = (\gamma_1\gamma_2 + \alpha_1\alpha_2) / 0$$

MR Stratified by Gene Reversing Environmental Characteristic



If unbiased:

E=1

Gene (G) $\xrightarrow{\phi_{E=1}>0}$ **Phenotype (X)** $\overset{\beta}{\dashrightarrow}$ **Health (Y)**

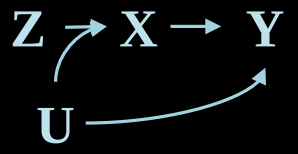
$\beta_{MR} = \phi_{E=1} * \beta / \phi_{E=1} = \beta$

E=0

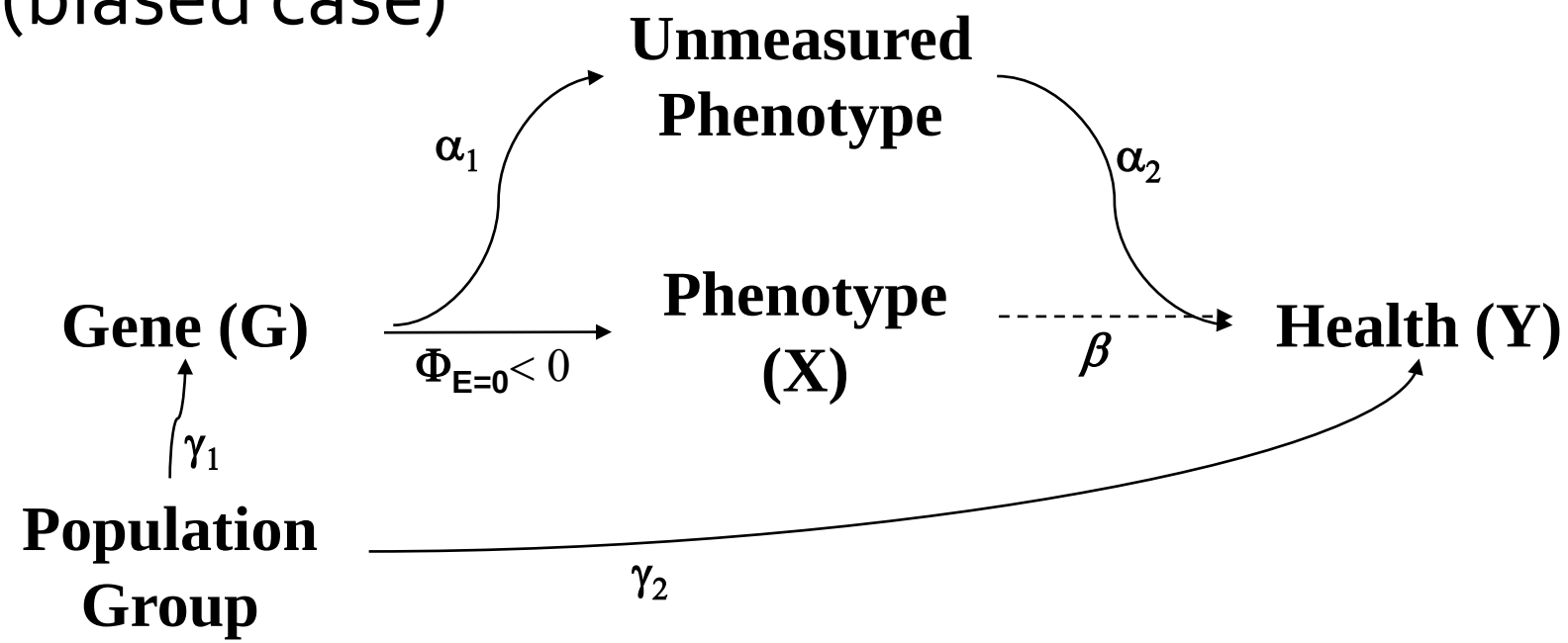
Gene (G) $\xrightarrow{\Phi_{E=0}< 0}$ **Phenotype (X)** $\overset{\beta}{\dashrightarrow}$ **Health (Y)**

$\beta_{MR} = \phi_{E=0} * \beta / \phi_{E=0} = \beta$

MR Stratified by Gene Reversing Environmental Characteristic

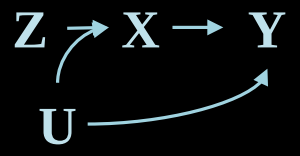


E=o (biased case)

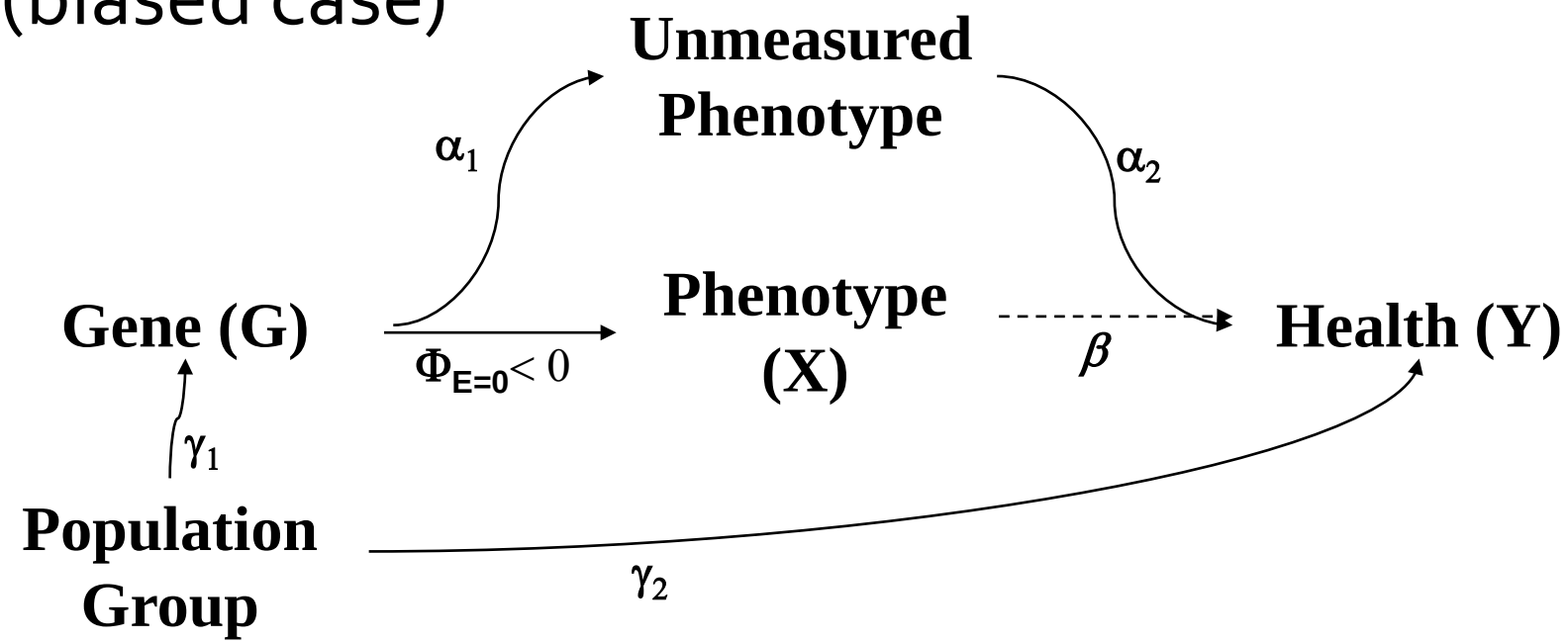


$$\beta_{MR} = (\phi_{E=o}\beta + \gamma_1\gamma_2 + \alpha_1\alpha_2)/\phi_{E=o}$$

MR Stratified by Gene Reversing Environmental Characteristic

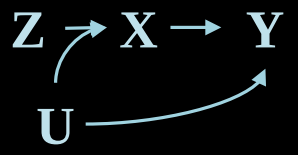


E=o (biased case)



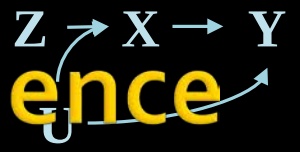
$$\beta_{MR} = (\phi_{E=o}\beta + \gamma_1\gamma_2 + \alpha_1\alpha_2)/\phi_{E=o} \neq \beta \neq (\phi_{E=1}\beta + \gamma_1\gamma_2 + \alpha_1\alpha_2)/\phi_{E=1}$$

5 Evaluating IV Studies: Negative controls



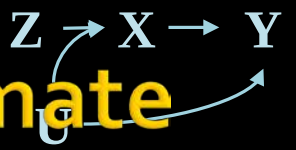
- Identify outcomes with similar confounding structure but that are definitely not affected by exposure.

6 Evaluating IV Studies: Independence from measured confounders



- Identify measured confounders of exposure-outcome association
- Test whether these variables are independent of the IV
- If not, you can control for these confounders, but must worry about unmeasured confounders
- Sometimes most coherent to examine independence conditional on a core set of confounders, e.g., age, population eigenvectors

7 Egger Regression: treat each IV estimate as an element in a meta-analysis



- Under the InSIDE (instrument strength independent of direct effects) assumption, bias converges to zero.
- Regress the Z-Y associations on the Z-X associations.

The intercept is an estimate of average pleiotropy and the slope is an estimate of the true causal effect under InSIDE.

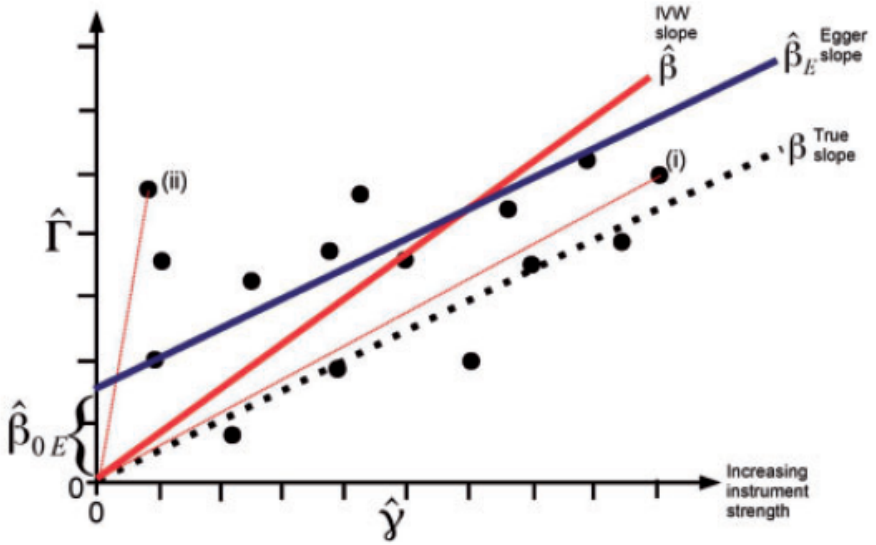
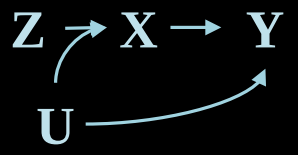


Figure 2. Plot of the gene–outcome ($\hat{\Gamma}$) vs gene–exposure ($\hat{\gamma}$) regression coefficients for a fictional Mendelian randomization analysis with 15 genetic variants. The true slope is shown by a dotted line, the inverse-variance weighted (IVW) estimate by a red line, and the MR-Egger regression estimate by a blue line. Refer to text for explanation of points

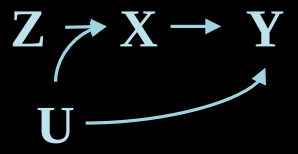
Bowden, Davey Smith, and Burgess, Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression, IJE 2015

Conclusions



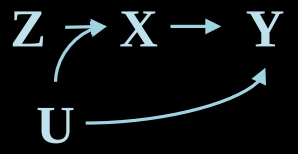
- Honest epidemiologists tend to be worried epidemiologists
- Abandon the difficult questions?
  - Or learn what we can from fraught methods?
- IV adds:
  - A way forward with observational data
  - Sometimes a parameter estimate of special interest
  - Pushes us to identify interventions that change exposures
- A set of related methods use pseudo-randomization, strongly preferred in econ, less common in epi

Outline

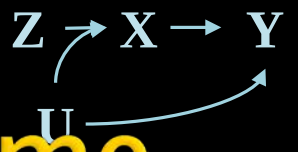


- Motivation/Intuition
- Defining an IV
  - Motivating examples: MTO and BMI→Dementia
- IV assumptions
- IV for continuous outcomes
- Interpretation of the IV effect estimate
  - Complier average causal effect versus effect of treatment on the treated
- IV assumption violations
- Evaluating the assumptions
- Dealing with different distributions

Overview

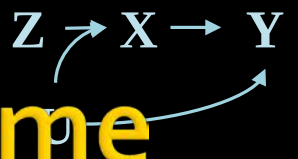


- IV analysis with outcome
- IV analysis in case-control studies
- IV analysis with survival outcomes
- IV analysis in R



IV Analysis with binary outcome

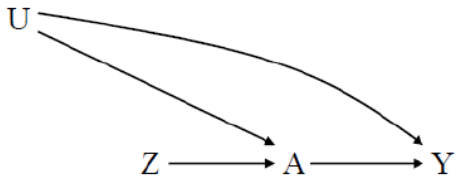
- Traditionally: use 2SLS with a linear probability model
  - Problem: no restriction on the space of a valid probability (0<=P<=1)
  - … might not be a problem when using genetic variants as instruments given that they explain so Little that the estimate will hardly ever be out off bounds
- Solution: use a link function: log, logit, probit



IV Analysis with binary outcome

IV Analysis with log link

- Suppose that the IV causal DAG holds



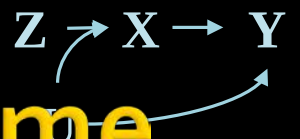
and

$$\begin{aligned} \log \Pr(Y = 1 | A, C, U, Z) &= \beta_0 + \beta_A A + \beta_U U, \\ A &= \alpha_0 + \alpha_U U + \alpha_Z Z + \delta_A \end{aligned}$$

where δ_A is independent residual error.

- The causal parameter is the causal risk ratio (conditional on U)

$$\frac{\Pr(Y = 1 | a + 1, C, U, Z)}{\Pr(Y = 1 | a, C, U, Z)} = \exp(\beta_A)$$



IV Analysis with binary outcome

IV Analysis with log link

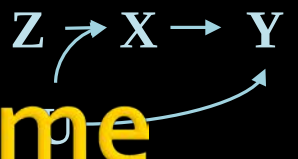
- One can then compute the implied model for $P(Y = 1|Z) =$

$$\begin{aligned}
& E\left\{\Pr\left(Y = 1|A, U, Z\right)|Z\right\} \\
= \; & E\left\{\exp\left(\beta_0 + \beta_A A + \beta_U U\right)|Z\right\} \\
= \; & E\left\{\exp\left(\beta_0 + \beta_A\left(\alpha_0 + \alpha_U U + \alpha_Z Z + \delta_A\right) + \beta_U U\right)|Z\right\} \\
= \; & \exp\left(\beta_0 + \beta_A \alpha_Z Z\right) \underbrace{E\left\{\exp\left(\beta_A\left(\alpha_0 + \alpha_U U + \delta_A\right) + \beta_U U\right)|Z\right\}}_{=\text{constant}} \\
= \; & \exp\left(\beta_0^* + \beta_A \alpha_Z Z\right) \\
= \; & \exp\left(\beta_0^* + \beta_Z^* Z\right)
\end{aligned}$$

  where we use the fact that U and δ_A are independent of Z.

- This implies that

$$\beta_A = \frac{\beta_Z^*}{\alpha_Z}$$



IV Analysis with binary outcome

IV Analysis with logit link

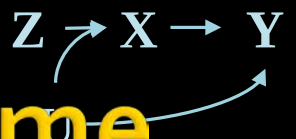
- Stage 1: Fit a linear regression of A on Z and C and compute the predicted value

$$\widehat{E}(A|Z,C) = \widehat{\alpha}_0 + \widehat{\alpha}_1 Z + \widehat{\alpha}_2' C$$

- Stage 2:Fit a log linear regression of Y on $\widehat{E}(A|Z,C)$ and C

$$\text{logitPr}(Y=1) = \mu_0 + \mu_1 \widehat{E}(A|Z,C) + \mu_2' C$$

- Then under the linear-log linear model, $\widehat{\mu}_1$ obtained in Stage 2 will converge to β_A.
- When the outcome is a rare disease, it is convenient to use standard logistic regression in Stage 2.



IV Analysis with binary outcome

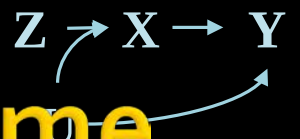
IV Analysis with logit link

- The estimated regression coefficient $\widehat{\mu}_1$ identifies the conditional causal effect

$$\frac{\Pr(Y_1 = 1|C)\Pr(Y_0 = 0|C)}{\Pr(Y_1 = 0|C)\Pr(Y_0 = 1|C)}$$

  only if :
- The disease outcome Y is rare in the underlying population and A is continuous.
- The unobserved confounder does not interact with (A, C) in the odds ratio scale in a model for the outcome
- The unobserved confounder does not interact with Z and C in predicting the exposure
- The regression model computed in the first stage is correctly specified.(note that this was not needed in the linear case)
- the residual for the first stage regression is homoscedastic

IV Analysis with binary outcome



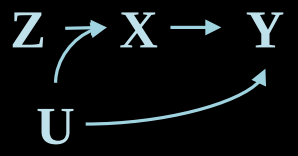
- There exist alternative methods for binary outcome, that relax some but not all of these rather stringent conditions, for example the probit model, which relies on a normal assumption about U.
- These methods essentially rely on making explicit assumptions about the unobserved confounders for interpretation.

- This may be reasonable if one knows what the unobserved confounder is, such that certain restrictions may be scientifically justified, otherwise, identification would be driven mainly by assumption about hidden variables (i.e. unobserved confounder)

- A major restriction of 2 stage linear logistic is that it cannot generally be used for binary exposure A.

IV for survival outcome

IV for Cox Proportional Hazards Model



- We observe data $(T^* = \min(T, W), A, Z, \Delta)$ where T is the censored outcome of interest, Δ is an indicator of whether T is censored, i.e. $\Delta = 1$ indicates $T^* = T$ and W is an independent censoring time
- Example: A is BMI, Z is FTO and T is time to death. We wish to estimate the causal effect of BMI on mortality.
- Consider the proportional hazards model

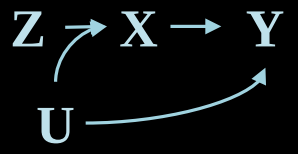
$$h(t|A, U, Z) = h_0(t) \exp(b_a A + b_u(U, t)) \tag{3}$$

where b_a is the log-hazards ratio encoding the effect of exposure, and $b_u(U, t)$ denotes the effect of an unmeasured confounder U at time t.

- Note that Z does not appear in the Cox model by IV exclusion restriction.

IV for survival outcome

IV for Cox Proportional Hazards Model



- Also assume that

$$A = \alpha_0 + \alpha_z Z + \Delta$$

and we have

$$COV(\Delta, U|Z) \neq 0,$$

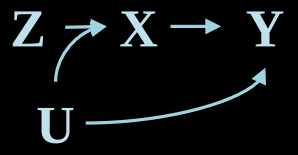
which implies that U confounds the effect of A on T.

- Tchetgen Tchetgen et al (2014) establish that if the outcome is rare,

$$\begin{aligned} h(t|Z) &= E\left[h(t|A, U, Z)\,|Z\right] \\ &\approx h_0^*(t)\exp\left(b_a \alpha_z Z\right) \end{aligned}$$

IV for survival outcome

IV for Cox Proportional Hazards Model



- This implies that(approximate) Wald estimand

$$b_a \approx \frac{b_z^*}{\alpha_z}$$

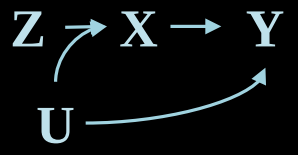
where b_z^* is obtained by fitting the proportional hazards model

$$h\left(t|Z\right) = h_0^*\left(t\right)\exp\left(b_z^* Z\right)$$

- This also suggests a 2 stage approach (incorporating covariates C)

IV for survival outcome

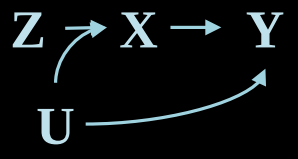
IV for Cox Proportional Hazards Model



- Stage 1: Fit $A = \alpha_0 + \alpha_z Z + \alpha_c' C + \Delta$ by standard OLS, obtain fitted values $\widehat{A} = \widehat{\alpha}_0 + \widehat{\alpha}_z Z + \widehat{\alpha}_c' C$
- Stage 2: Fit $h(t|Z) = h_0^{**}(t) \exp\left(b_a \widehat{A} + b_c' C\right)$ by standard Cox regression with covariates $(\widehat{A}, C)$, to obtain $\widehat{b}_a$ a consistent estimate of the causal log-hazard ratio.

IV for survival outcome

IV for Aalen Additive Hazards Models



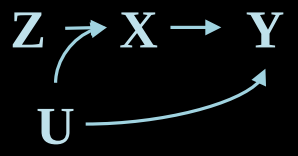
- Limitation of Cox IV is that it is only appropriate when the outcome is rare.
- An alternative approach is to consider Aalen's additive hazards model:

$$h\left(t|A,U,Z\right)=b_0\left(t\right)+b_a\left(t\right)A+b_u(U,t) \tag{4}$$

  where $\left(b_0,b_a,b_u\right)$ are unrestricted.
- The model states that conditional on U, the effect of A on T encoded on the additive hazards scale is linear in A for each t, however, the effect size $b_a\left(t\right)$ may vary with t.
- The null hypothesis of no causal effect of A (BMI) on T (mortality) is encoded by $b_a\left(t\right)=0$ for all t.

IV for survival outcome

IV for Aalen Additive Hazards Models



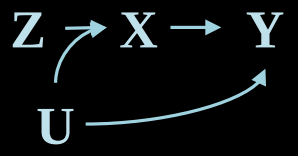
- An important submodel of interest is the constant hazards difference model obtained by setting

$$b_a(t) = b_a \tag{5}$$

  where b_a is an unknown constant.
- The model conditions on the unobserved confounders U, and allows its effect $b_u(U, t)$ to remain completely unrestricted, both in U and with respect to t.
- Crucially, the model assumes no interaction between A and U.
- The baseline hazard function $b_0(t)$ is apriori unrestricted.
- Finally, as in the Cox model, the right hand side of equation (4) does not depend on Z by the exclusion restriction.

IV for survival outcome

IV for Aalen Additive Hazards Models



- It then follows that assuming

$$A = \alpha_0 + \alpha_z Z + \Delta$$

with

$$COV(\Delta, U|Z) \neq 0,$$

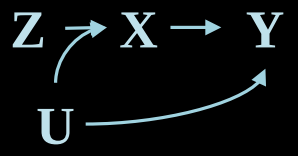
which implies that U confounds the effect of A on T,Tchetgen Tchetgen et al (2014) establish that

$$\begin{aligned} h(t|Z) &= E\left[h(t|A, U, Z)\,|Z\right] \\ &= h_0^*(t) + b_a(t)\,\alpha_z Z \end{aligned}$$

- Unlike for Cox regression, the result is not an approximation, it is exact!!!
- The result holds whether the outcome is rare or not.

IV for survival outcome

IV for Aalen Additive Hazards Models



- This implies the Wald estimand

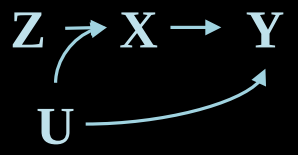
$$b_a(t) = \frac{b_z^*(t)}{\alpha_z}$$

where $b_z^*(t)$ is obtained by fitting the Aalen model

$$h(t|Z) = h_0^*(t) + b_z^*(t) Z$$

- This also suggests a 2 stage approach (incorporating covariates C)
  - Stage 1: Fit $A = \alpha_0 + \alpha_z Z + \alpha_c' C + \Delta$ by standard OLS, obtain fitted values $\widehat{A} = \widehat{\alpha}_0 + \widehat{\alpha}_z Z + \widehat{\alpha}_c' C$
  - Stage 2: Fit $h(t|Z) = h_0^{**}(t) + b_a(t)\widehat{A} + b_c'(t) C$ using standard software for additive Aalen model with covariates $(\widehat{A}, C)$.
- 2 stage approach results in valid estimate of $b_a(t)$ under our assumptions

Inverse Variance Weighted IV of separate samples (Burgess et al. 2013)



Burgess, Stephen, Adam Butterworth, and Simon G. Thompson. "Mendelian randomization analysis with multiple genetic variants using summarized data." *Genetic epidemiology* 37.7 (2013): 658-665.

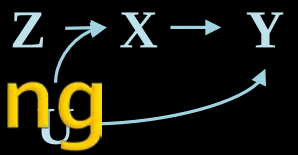
$$\hat{\beta}_{IVW} = \frac{\sum_k X_k Y_k \sigma_{Yk}^{-2}}{\sum_k X_k^2 \sigma_{Yk}^{-2}} \qquad (1)$$

The approximate standard error of the estimate is:

$$se(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_k X_k^2 \sigma_{Yk}^{-2}}} \qquad (2)$$

Genetic variant k, $k = 1, \ldots, K$ is associated with an observed Xk mean change in the risk factor per additional variant allele with standard error σXk and an observed Yk mean change in the outcome per allele with standard error σYk

Results for BMI on Dementia, including Inverse Variance Weighted IV

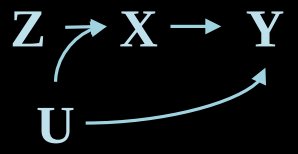


	Late onset Alzheimer's disease					
	ADGC		GERAD[†]		ADGC + GERAD[‡]	
	OR (95% CI)	*P*	OR (95% CI)	*P*	OR (95% CI)	*P*
BMI polygenic score	0.95 (0.90 to 1.01)	.091	0.96 (0.87 to 1.07)	.488	0.95 (0.91 to 1.00)	.058
Mechanism-specific BMI polygenic scores						
Adipogenesis	1.04 (0.87 to 1.24)	.700	1.26 (0.82 to 1.95)	.297	1.07 (0.90 to 1.27)	.450
Appetite	0.96 (0.89 to 1.03)	.280	0.94 (0.83 to 1.07)	.381	0.96 (0.90 to 1.02)	.160
Cardiopulmonary processes	1.17 (0.97 to 1.42)	.102	1.33 (0.90 to 1.97)	.152	1.21 (1.00 to 1.47)	.055
Unspecified cellular processes	0.82 (0.72 to 0.92)	.001	0.81 (0.62 to 1.06)	.131	0.81 (0.74 to 0.90)	<.001

Stefan Walter
Dept of Epidemiology and Biostatistics
UCSF
swalter@psg.ucsf.edu

IV Analysis

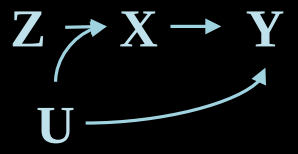
Practical Session: From Stefan Walter's SER practice



- Generate the Universe:

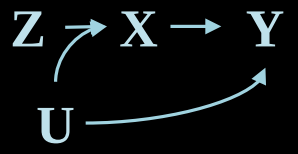
```
z1<-sample(c(0,0,0,0,0,1,1,1,2,2),10000, replace=T)
z2<-sample(c(0,0,0,0,0,1,1,1,2,2),10000, replace=T)
z3<-sample(c(0,0,0,0,0,1,1,1,2,2),10000, replace=T)
e1<-rnorm(10000, sd=3)
e2<-rnorm(10000, sd=3)
U<-rnorm(10000, sd=1)
A<-27+0.6*z1+0.3*z2+0.1*z3+2*U+e1
Y<-30+1.5*A+4*U+e2
```

2SLS



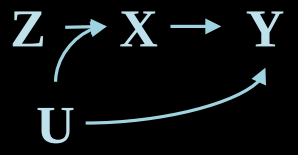

```
summary(lm(A~z1))   #beta = 0.6
summary(lm(A~z2))   #beta = 0.3
summary(lm(A~z3))   #beta = 0.1
summary(lm(A~z1+z2+z3))   #beta = 0.7
summary(lm(Y~A))   #beta = 2.2
#twostep
pred1<-predict(lm(A~z1))
summary(lm(Y~pred1))   #beta = 1.5
#Control Function IV
res1<-summary(lm(A~z1))$residual
summary(lm(Y~A+res1))   #beta = 1.5
```

2SLS



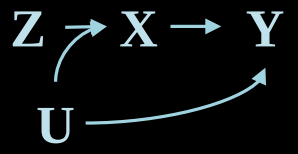

```
#ivreg 2 script
http://diffuseprior.wordpress.com/tag/over-identification/
#ivreg2(form,endog,iv,data,digits)
mroz<-as.data.frame(cbind(Y,A,z1,z2,z3))
ivreg2(form=Y ~ A
,endog="A",iv=c("z1","z2","z3"),data=mroz)
mrozs$res1<-
summary(lm(A~z1,data=mrozb))$residual
summary(lm(Y~A+res1, data=mrozs))   #beta = 1.5
```

Split Sample IV



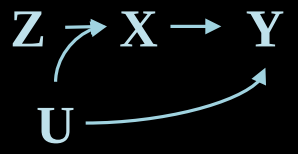

```
mrozb<-as.data.frame(cbind(Y,A,U, z1,z2,z3))
pred1<-predict(lm(A~z1))
summary(lm(Y~pred1))   #beta = 1.5
mrozvs<-mrozb[sample(1:10000, 500),]
a<-coef(lm(A~z1+z2+z3, data=mrozb))
mrozvs$GRS<-apply(sweep(mrozvs[c("z1",
"z2","z3")],MARGIN=2,c(a[2:4]),`*`),1,function(x
) sum(x, na.rm=T))
summary(lm(Y~GRS, data=mrozvs))   #beta =
0.77
```

Inverse Variance Weighted IV (external data)




```
### Burgess Approach
coeftest(lm(A~z1))  0.516322   0.046800
coeftest(lm(A~z2))  0.342791   0.046807
coeftest(lm(Y~z1))  0.84012    0.11495
coeftest(lm(Y~z2))  0.66891    0.11469
Xk<-c(0.516,0.343)
Xkse<-c(0.047,0.047)
Yk<-c(0.840,0.669)
Ykse<-c(0.115,0.115)
sum(Xk*Yk*Ykse^-2)/sum(Xk^2*Ykse^-2)
#Inverse Variance Weighted IV
(1/sum(Xk^2*Ykse^-2))^0.5
```

Inverse Variance Weighted IV (external data)



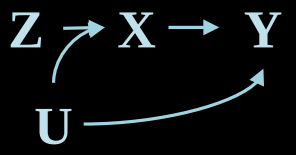

```
library(meta)

sum(Xk[1]*Yk[1]*Ykse[1]^-2)/sum(Xk[1]^2*Ykse[1]^-2)
#1.626
(1/sum(Xk[1]^2*Ykse[1]^-2))^0.5              #0.223

sum(Xk[2]*Yk[2]*Ykse[2]^-2)/sum(Xk[2]^2*Ykse[2]^-2)
#1.950
(1/sum(Xk[2]^2*Ykse[2]^-2))^0.5              #0.0.335

metagen(c(1.626,1.950),c(0.223, 0.335))   #identical -->
no heterogeneity, Instrument OK
```

Some background language



- Questions of interest are about "causation": "would intervening on exposure X change the value of outcome Y?" i.e. $Y_{X=1}$ vs $Y_{X=0}$
- Assume we never know this at the individual level, because we never find out what Y would have been if exposure had been different than it actually was
- Unfortunately $Y_{X=1} \neq Y|X=1$ because the people with X=1 may also tend to have some other feature, say U, that influences the value of Y.
- U is a confounder: a common prior cause of X and Y.
- We can estimate the effect of X on Y at a group level by randomization, pseudo-randomization, or controlling for common causes of X and Y.