Robustness and Efficiency of Covariate Adjusted Linear Instrumental Variable Methods

Vanessa Didelez

School of Mathematics

University of Bristol

joint work with Stijn Vansteelandt, Ghent

Copenhagen — April 2015

Overview

- Basic Idea of IVs, e.g. Mendelian Randomisation
- (Marginal) Instrumental Variables
- Conditional Instrumental Variables
- Model Assumptions and Robustness / Efficiency
- Efficiency and Robustness of 2-Stage Methods
- G–Estimators, Double-Robustness, and Bias Reduction
- Conclusions

Motivation

Epidemiology interested in effect of interventions ('drink less alcohol', 'eat folic acid' etc.)

Observational studies are inevitable: preliminary research, but also assessment of effects in general population.

Obvious problem is **confounding:** effects of interest are entangled with many other effects — this can never be fully excluded.

Instrumental variables allow *some* inference on effects of interventions in the presence of confounding.

Problem with this is: how to find a suitable instrument? It has recently become popular to look for a genetic variant as IV — **Mendelian randomisation**.

Mendelian Randomisation: Basic Idea

If we cannot randomise, let's look for instances where NATURE has randomised, e.g. through genetic variation.

Example: Alcohol Consumption

Genotype: ALDH2 determines blood acetaldehyde, the principal metabolite for alcohol.

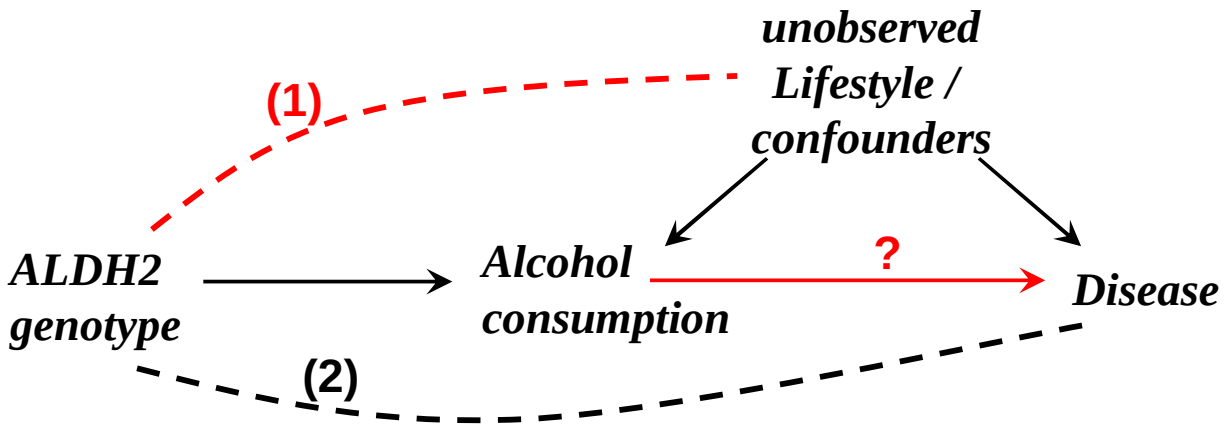
Two alleles/variants: wildetype *1 and "null" variant *2.

*2*2 homozygous individuals suffer facial flushing, nausea, drowsiness and headache after alcohol consumption.

$\Rightarrow$ *2*2 homozygous individuals have low alcohol consumption *regardless* of their other lifestyle behaviours

IV–Idea: check if these individuals have a different risk than others for alcohol related health problems!

Example: Alcohol Consumption

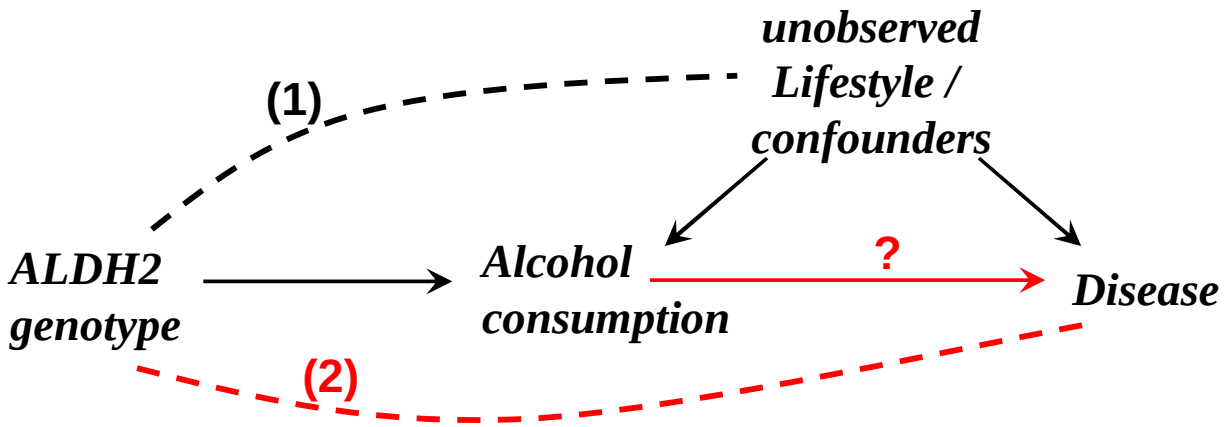


Note 1: due to random allocation of genes at conception, can be fairly confident that genotype is not associated with unobserved confounders (subpopulation structure can be a problem).

Further evidence: in extensive studies no evidence for association with *observed* confounders, e.g. age, smoking, BMI, cholesterol.

(see also Davey Smith et al., 2007)

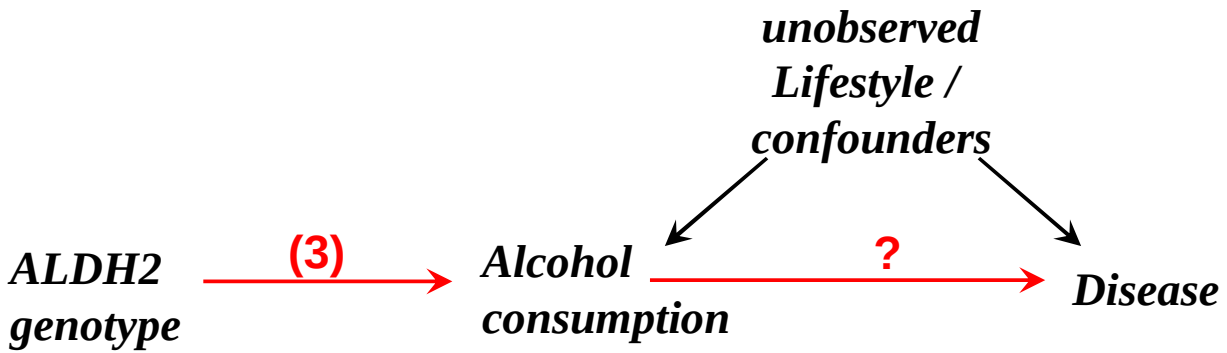
Example: Alcohol Consumption



Note 2: due to known 'functionality' of ALDH2 gene, we can exclude that it affects the typical diseases considered by *another* route than through alcohol consumption.

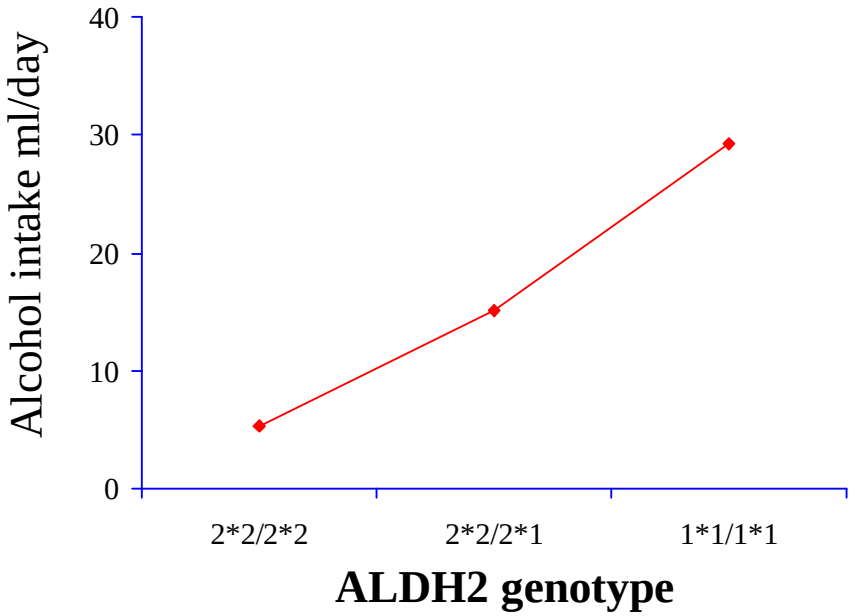
$\Rightarrow$ important to use well studied genes as instruments!

Example: Alcohol Consumption



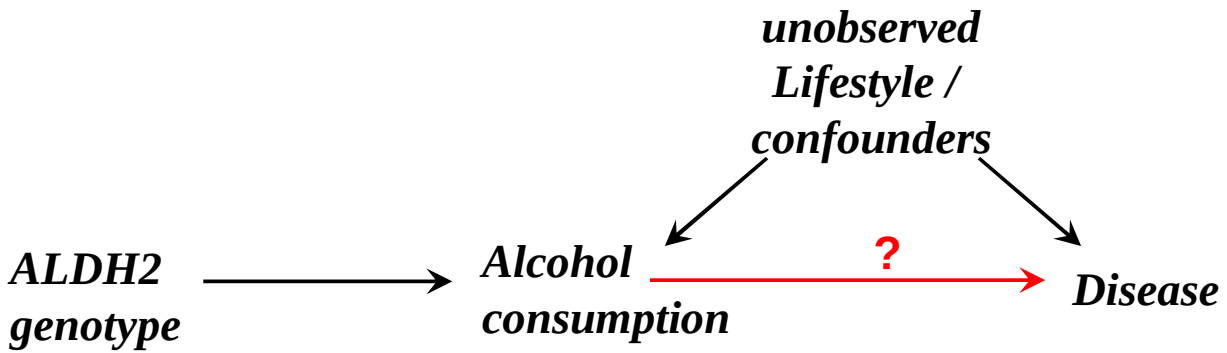
Note 3: association of ALDH2 with alcohol consumption well established, strong, and underlying biochemistry well understood.

Example: Alcohol Consumption



Note 3: association of ALDH2 with alcohol consumption well established, strong, and underlying biology well understood.

Example: Alcohol Consumption



Test for Causal Effect? under IV assumptions, the null–hypothesis of no causal effect of alcohol consumption, should imply no association between ALDH2 and disease;
While if alcohol consumption has a causal effect we would expect an association between ALDH2 and disease.

Example: Alcohol Consumption

Findings: (Meta-analysis by Chen et al., 2008)

Blood pressure on average 7.44mmHg higher and
risk of hypertension 2.5 higher
for ALDH2*1*1 than for ALDH2*2*2 carriers (only males).
$\Rightarrow$ mimics the effect of *large versus low* alcohol consumption.

Blood pressure on average 4.24mmHg higher and
risk of hypertension 1.7 higher
for ALDH2*1*2 than for ALDH2*2*2 carriers (only males).
$\Rightarrow$ mimics the effect of *moderate versus low* alcohol consumption.

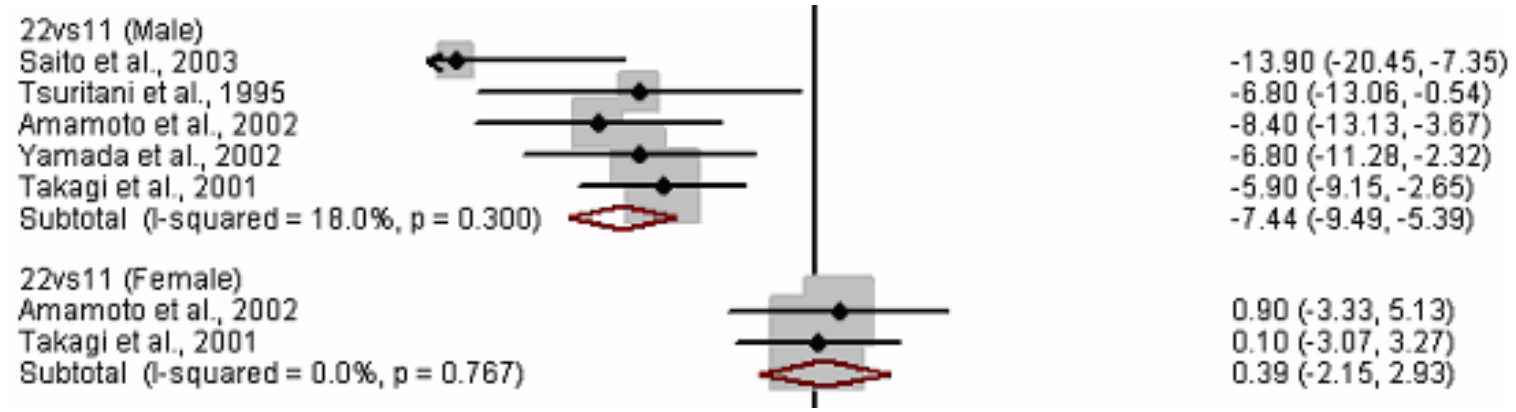
$\Rightarrow$ it seems that **even moderate** alcohol consumption is **harmful**.

Note: studies mostly in Japanese populations (where ALDH2*2*2 is common), where women drink only little alcohol in general.

Example: Alcohol Consumption

(Chen et al., 2008)

Can somewhat **check assumptions:**



Some indication

Women in Japanese study population do not drink. ALDH2 genotype in women not associated with blood pressure $\Rightarrow$ there does not seem to be another pathway creating a G–Y association here.

Why does IV Help with Causal Inference?

Testing:

check if IV and outcome are oassociated — this is (roughly) testing whether there is a causal effect of exposure on outcome at all.

Estimation:

(1) when all observable variables are discrete, we can obtain bounds on causal effects without further assumptions.

(cf. `Stata` package, Palmer et al., 2011)

(2) for point estimates need some (semi–)parametric / structural assumptions, as well as clear definition of target causal parameter.

Motivation ctd.

IV analysis always less precise $\Rightarrow$ larger st.errors, CIs, low power etc.

Especially in **Mendelian Randomisation**: weak IVs (gene-exposure association is weak relative to sample size)

Wanted: exploit all available information & be as efficient as possible, e.g.

– use multiple IVs (e.g. multiple SNPs)

– use **observed** covariates

Examples of IV Applications

- Randomised trials with partial compliance: $Z =$ 'randomisation', $X =$ 'actual treatment taken'.
- Mendelian randomisation: $Z =$ 'genotype(s)', $X =$ 'phenotype', e.g. ALDH2 and alcohol intake.
- Pharmaco-epidemiology: $Z =$ 'physician's / hospital's preferred prescription', $X =$ 'actual prescription'.
- Econometrics: $Z =$ 'area of residence', $X =$ 'access to different types of schools'.

Note: in all of these it is common to measure some **covariates**, at least baseline (age, sex, SES, etc.)

(Marginal) Instrumental Variable

Wanted: effect of X on Y.

Problem: confounding by (set of) **unobs. variables** U.

If you're lucky: instrumental variable Z can help.

Assumptions:

(A1) $Z \perp\!\!\!\perp U$

(A2) $Z \not\perp\!\!\!\perp X$

(A3) $Z \perp\!\!\!\perp Y \mid (X, U)$

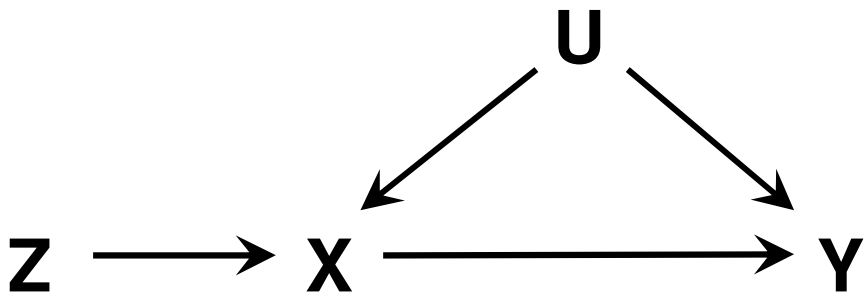
$+$ structural assumptions about interventions

$\Rightarrow$ enables testing for causal effect & under further parametric assumptions estimation of causal parameter(s).

Note: (A1) and (A3) not (generally) testable.

Assumptions as Directed Acyclic Graph (DAG)

DAG shows conditional independence as missing edge (d–separation).



Assumption (A1) corresponds to absence of Z—U edge.

Assumption (A3) corresponds to absence of Z—Y edge.

Note: need to justify the *absence* of edges.

(Marginal) Instrumental Variable Revisited

Z is marginal IV wrt. confounding by unobserved U and observed C if

Assumptions:

(A1*) $Z \perp\!\!\!\perp (U, C)$

(A2*) $Z \not\perp\!\!\!\perp X$

(A3*) $Z \perp\!\!\!\perp Y \mid (X, U, C)$

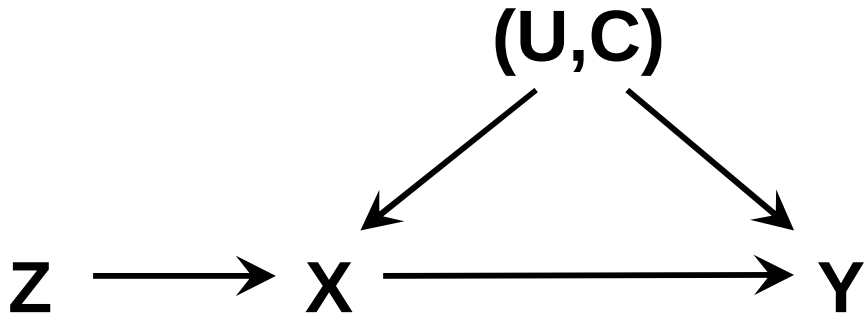
$+$ structural assumptions

$\Rightarrow$ enables inference as before, where C could *in principle* be ignored.

Note: (A1*) implies $Z \perp\!\!\!\perp C$ which can be tested (cf. Davey Smith et al., 2007)

Marginal IV — Assumptions as DAG

DAG shows conditional independence as missing edge (d–separation).



Assumption (A1*) corresponds to absence of Z—(U, C) edge.

Assumption (A3*) corresponds to absence of Z—Y edge.

Note: need to justify the *absence* of edges.

Marginal IV — Issues

Ignore or include covariates C?

- danger of misspecifying models with covariates, especially if high-dimensional;
- potential for efficiency gains when more information is used
  $\Rightarrow$ seems important in view of weak IVs in MR studies;
- must include covariates if structural model requires it, i.e. effect modifier (can also be problem with U of course).

Marginal IV — Example for Efficiency Gains

Generated data:

U, C independent $\sim N(0, 1)$

Z binary, $p(Z = 1) = 0.27$

$X|Z, U, C \sim N(\mu_X, 1)$ with $\mu_X = 0.75Z + U + C$

$Y|X, U, C \sim N(\mu_Y, 1)$ with $\mu_Y = 0.5X - U - 2C$

$n = 500$, $F \approx 20$; $R^2 \approx 3\%$, reps $= 1000$.

Marginal IV — Example for Efficiency Gains

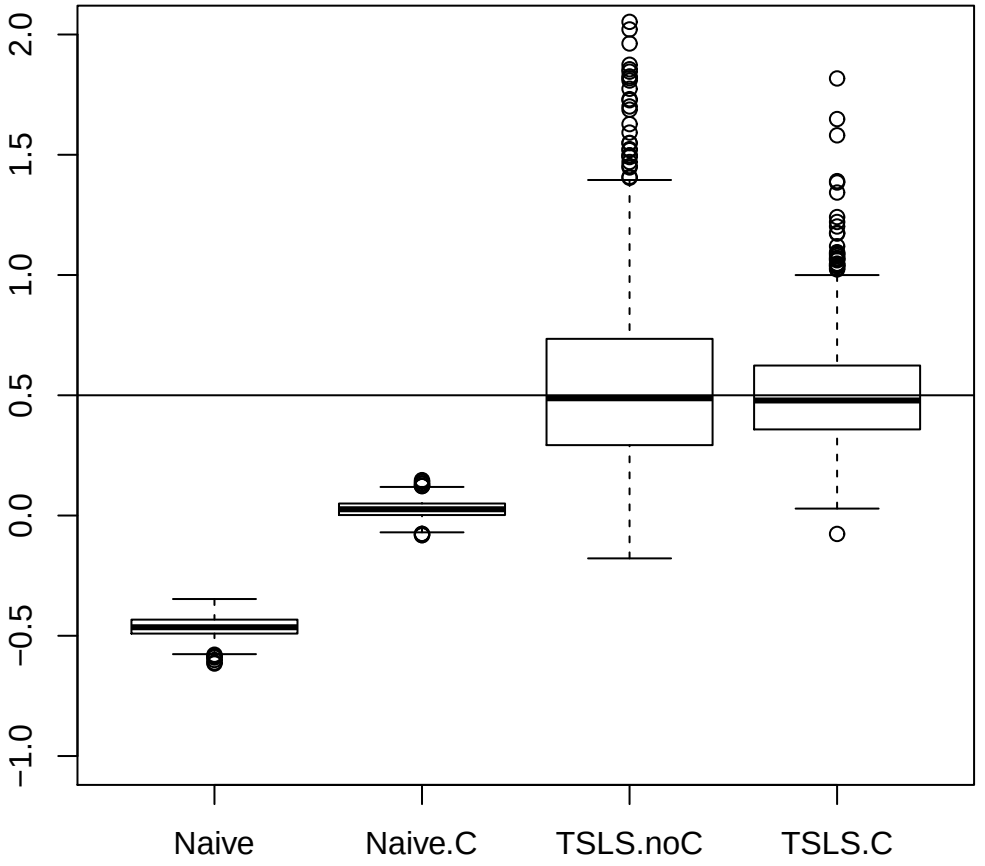
IV unadjusted: RMSE = 0.68

IV adjusted: RMSE = 0.55

Power testing 'no causal effect'

– unadjusted: 46%

– adjusted: 84%



Conditional Instrumental Variables

Definition:

Z is a conditional IV for effect of X on Y relative to confounding by U given observed covariates C if

(C1) $Z \perp\!\!\!\perp U \mid C$

(C2) $Z \not\perp\!\!\!\perp X \mid C = c$, for all c

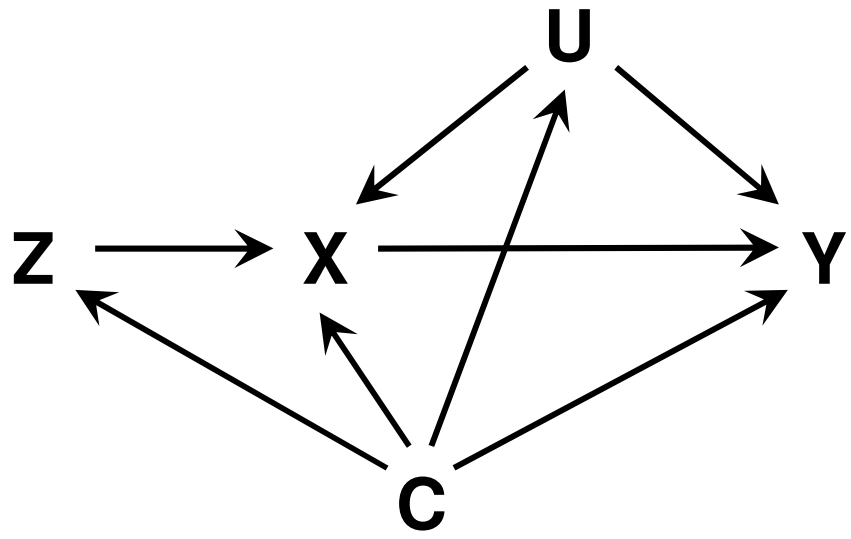
(C3) $Z \perp\!\!\!\perp Y \mid (X, U, C)$

$+$ structural assumptions

$\Rightarrow$ In principle: carry out inference within levels of / stratified by C.

Conditional IV — Assumptions as DAG

DAG shows conditional independence as missing edge (d–separation).
(Other DAGs possible.)



Assumption (C1) all Z—U paths blocked by C.

Assumption (C3) all Z—Y paths blocked by (X, U, C).

Note: need to justify the *absence* of edges.

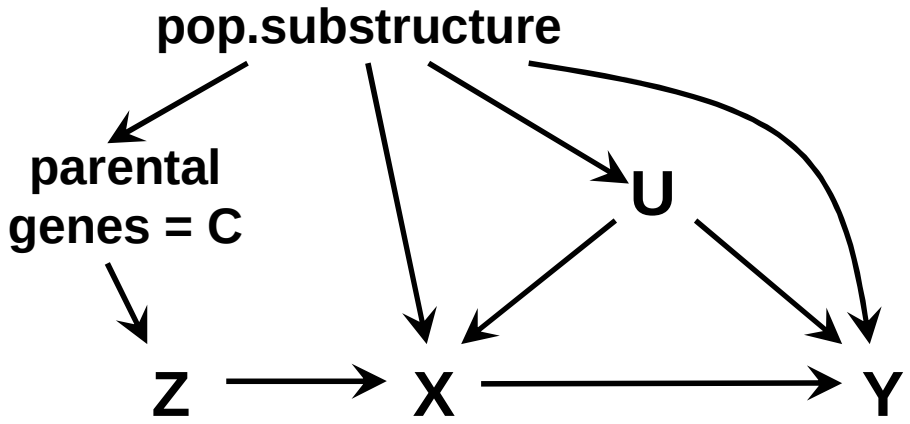
Conditional IV — Examples

Population substructure in Mendelian randomisation studies.

$X =$ modifiable phenotype

$Y =$ health outcome

$Z =$ genotype for X



$\Rightarrow$ conditional on either pop. indicator or on parental genes Z is valid conditional IV.

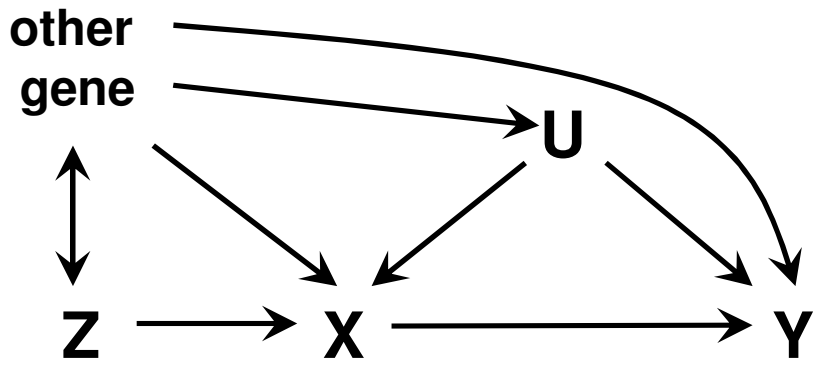
Conditional IV — Examples

LD in Mendelian randomisation studies.

$X =$ modifiable phenotype

$Y =$ health outcome

$Z =$ genotype for X



$\Rightarrow$ conditional on "other gene" Z is valid conditional IV.

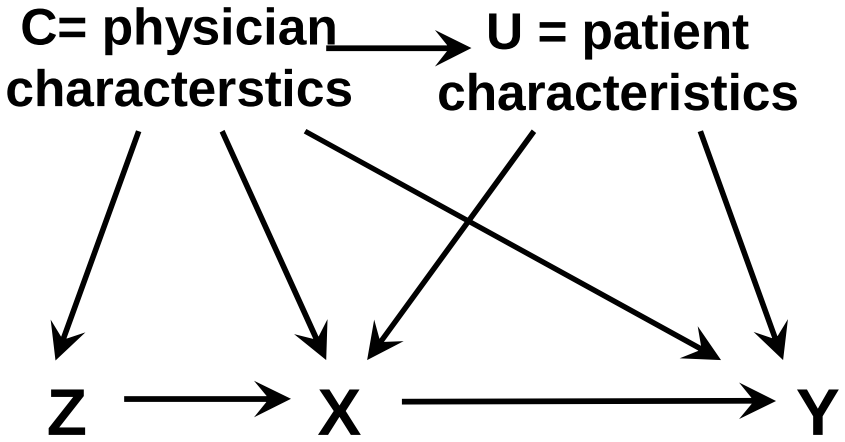
Conditional IV — Examples

Physician's preference used as IV in Pharmaco-epidemiology.

$X =$ actual version of drug taken

$Y =$ health outcome

$Z =$ physician's preference for a version of the drug



$\Rightarrow$ plausible that in official databases we find more information on physician's charateristics $(= C)$ than on patient's characteristics $(= U)$.

Conditional IV — Issues

Ignore or include covariates C?

- ignoring them will lead to bias;
- again danger of misspecifying models with covariates, especially if high-dimensional.

Robustness and Efficiency of IV Estimators

Robustness: some aspects of model can be misspecified, but estimator still consistent.

Efficiency: want to achieve less variability of consistent estimators.

Two-Stage Estimators

Popular: Two-stage-least-squares (TSLS) / other two-stage estimators

– first fit exposure model (1st stage)

– plug into outcome model (2nd stage).

Notes:

(i) separate fitting not always *efficient*;

(ii) misspecifying 1st stage could lead to bias (lack of *robustness*).

Formally: two-stage methods solve two separate estimating equations

$\Rightarrow$ check when equivalent to efficient / robust joint estimating equations.

Modelling Assumptions

Formal approach:

- what type of modelling assumptions can be made?
- are they needed / desirable?
- what happens under misspecification?
- can choices be made to reduce bias?

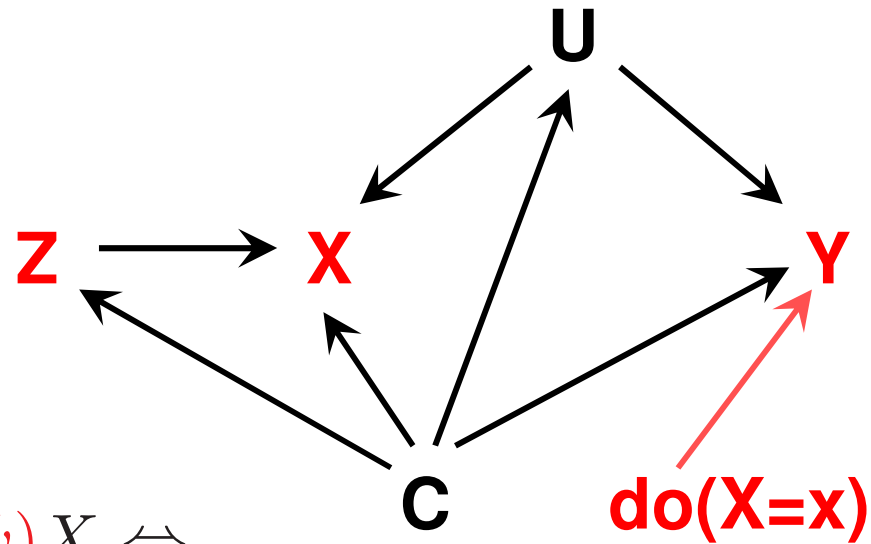
Modelling Assumptions

$\mathcal{M}$ structural model, e.g. $\mathcal{M}_{lin}$ = linear SMM

$\mathcal{A}_y$ model for effects of C on Y

$\mathcal{A}_x$ exposure model given C, Z

$\mathcal{A}_z$ instrument model given C



e.g. $\mathcal{M}_{lin}$

$E(Y|X,Z,U,C) = \omega(C,U) + m_y(C;\psi)X \Leftrightarrow$

LSMM: $E(Y|X,Z,C) - E(Y_0|X,Z,C) = m_y(C;\psi)X$

for instance choose $m_y(C;\psi) = \psi$.

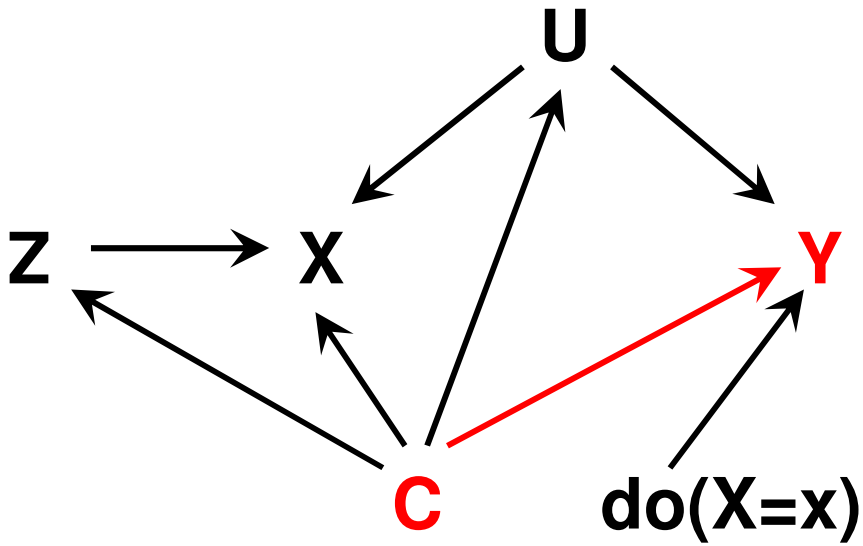
Modelling Assumptions

$\mathcal{M}$ structural model, e.g. $\mathcal{M}_{lin}$ = linear SMM

$\mathcal{A}_y$ model for effects of C on Y

$\mathcal{A}_x$ exposure model given C, Z

$\mathcal{A}_z$ instrument model given C



e.g. $\mathcal{A}_y$

$\omega(C;\beta) = m(C;\beta)$

for instance choose $m(C;\beta) = \beta^\top C$

$$\mathcal{M}_{lin} \cap \mathcal{A}_y \Rightarrow E(Y|Z,C) = m(C;\beta) + m_y(C;\psi)E(X|Z,C).$$

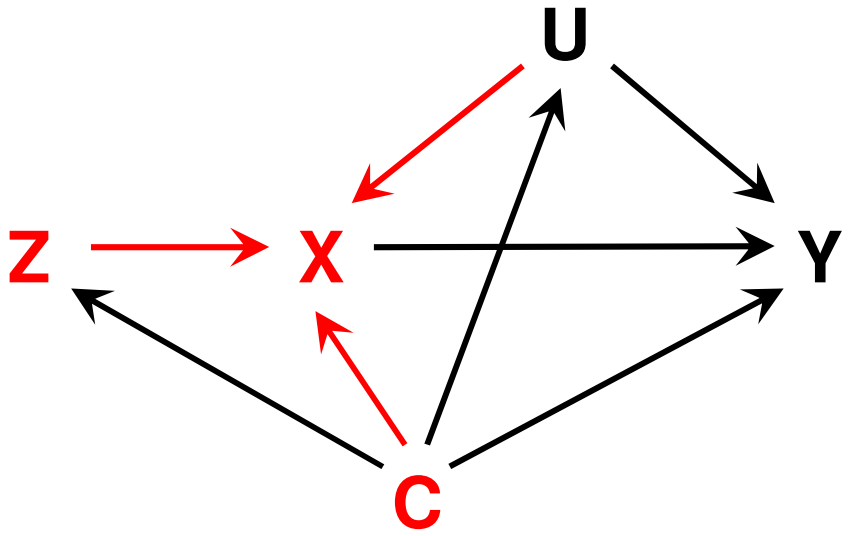
Modelling Assumptions

$\mathcal{M}$ structural model, e.g. $\mathcal{M}_{lin}$ = linear SMM

$\mathcal{A}_y$ model for effects of C on Y

$\mathcal{A}_x$ exposure model given C, Z

$\mathcal{A}_z$ instrument model given C



e.g. $\mathcal{A}_x$

$E(X|Z,C) = m_x(Z,C;\alpha)$

for instance choose $m_x(Z,C;\alpha) = \alpha_1^\top Z + \alpha_2^\top C$

Stand.cond.mean.model: $E(Y|Z,C) = m(C;\beta) + m_y(C;\psi)m_x(Z,C;\alpha)$.

(Chamberlain, 1987)

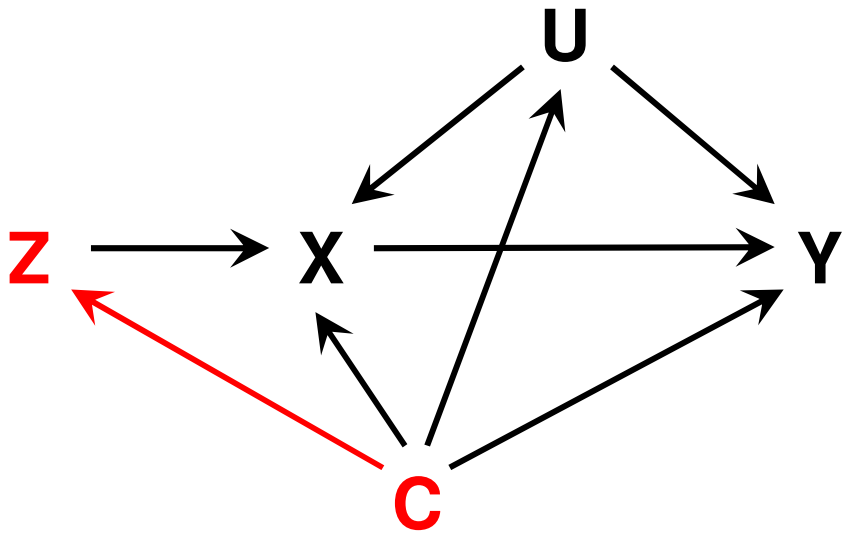
Modelling Assumptions

$\mathcal{M}$ structural model, e.g. $\mathcal{M}_{lin}$ = linear SMM

$\mathcal{A}_y$ model for effects of C on Y

$\mathcal{A}_x$ exposure model given C, Z

$\mathcal{A}_z$ instrument model given C



e.g. for marginal IV $Z \perp\!\!\!\perp C$ known.

not explicitly used in two-stage methods;

but relevant for **robust** estimation.

Effciency of Two-Stage Estimators

Under $\mathcal{M}_{lin} \cap \mathcal{A}_y \cap \mathcal{A}_x$: simultaneous fitting most efficient

But: can investigate when equivalent! (Details omitted...)

... for above choices (typical for TSLS):

$m_y(C;\psi) = \psi$ (no effect modification by C!)
$m(C;\beta) = \beta^\top C$
$m_x(Z,C;\alpha) = \alpha_1^\top Z + \alpha_2^\top C$ (linear)
and $Cov(X,Y|C,Z)$ constant.

In particular: always more efficient to include C than not.

Nothing similar when exposure model non-linear (for 2-stage methods).

Robustness of Two-Stage Estimators

Misspecification of $\mathcal{A}_x$ and/or $\mathcal{A}_y$?

Investigate when influence function independent of m_x and/or m_y.

(Details omitted...)

(i) robust wrt. $\mathcal{A}_x$: again for above model choices.

More generally: $\mathcal{A}_y$ must 'dictate' $\mathcal{A}_x$

(ii) robust wrt. $\mathcal{A}_y$: when $E(Z|C) = \gamma^\top C$, in particular if $Z \perp\!\!\!\perp C$

i.e. under particular $\mathcal{A}_y$.

Notes: see for (i) Wooldridge (2002), and for (ii) Robins (2000);
in (i) efficiency in particular subclass achieved.

Marginal IV — Example for Robustness

Here: $\mathcal{A}_x$, $\mathcal{A}_y$ misspecified, but $\mathcal{A}_z : Z \perp\!\!\!\perp C$.

Generated data:

U, C independent $\sim N(0,1)$

Z binary, $p(Z=1) = 0.27$

$X|Z,U,C \sim N(\mu_X, 1)$ with $\mu_X = Z - ZC + C + C^2 + U$

$Y|X,U,C \sim N(\mu_Y, 1)$ with $\mu_Y = 0.5X - U - 2C + 2C^2$

$n = 500$, $F \approx 22$; $R^2 \approx 4\%$, reps $= 10000$.

Marginal IV — Example for Robustness

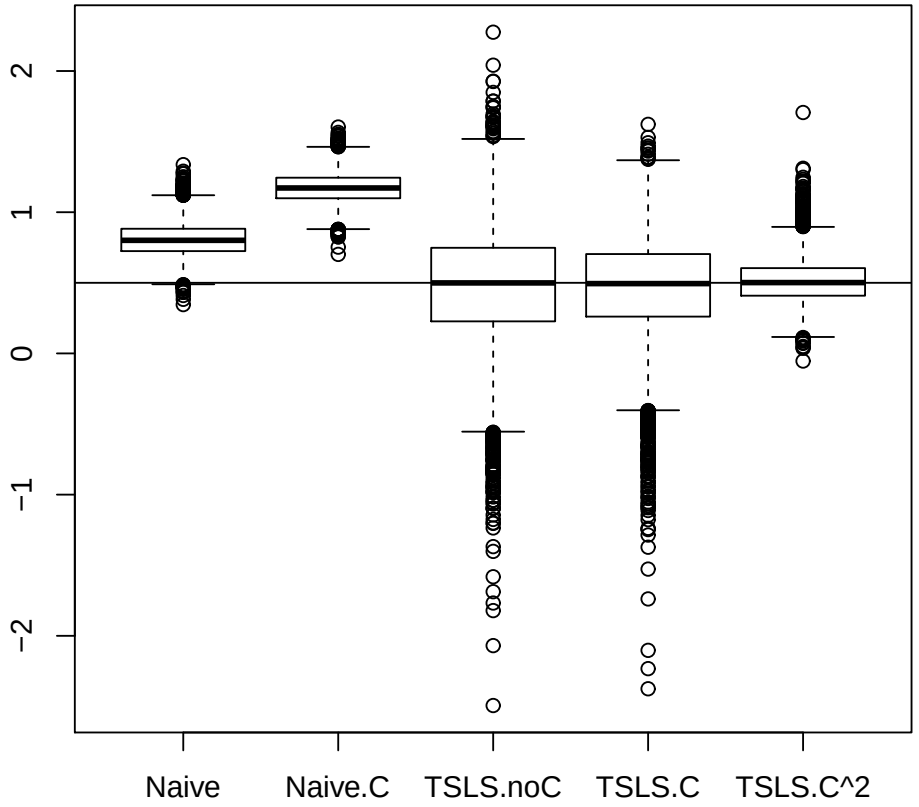
IV unadjusted: RMSE $= 0.63$

IV adjusted: RMSE $= 0.58$

Power testing 'no causal effect'

– unadjusted: 23%

– adjusted: 26%



Note: $\mathcal{A}_y$ correct with C^2 — more efficient.

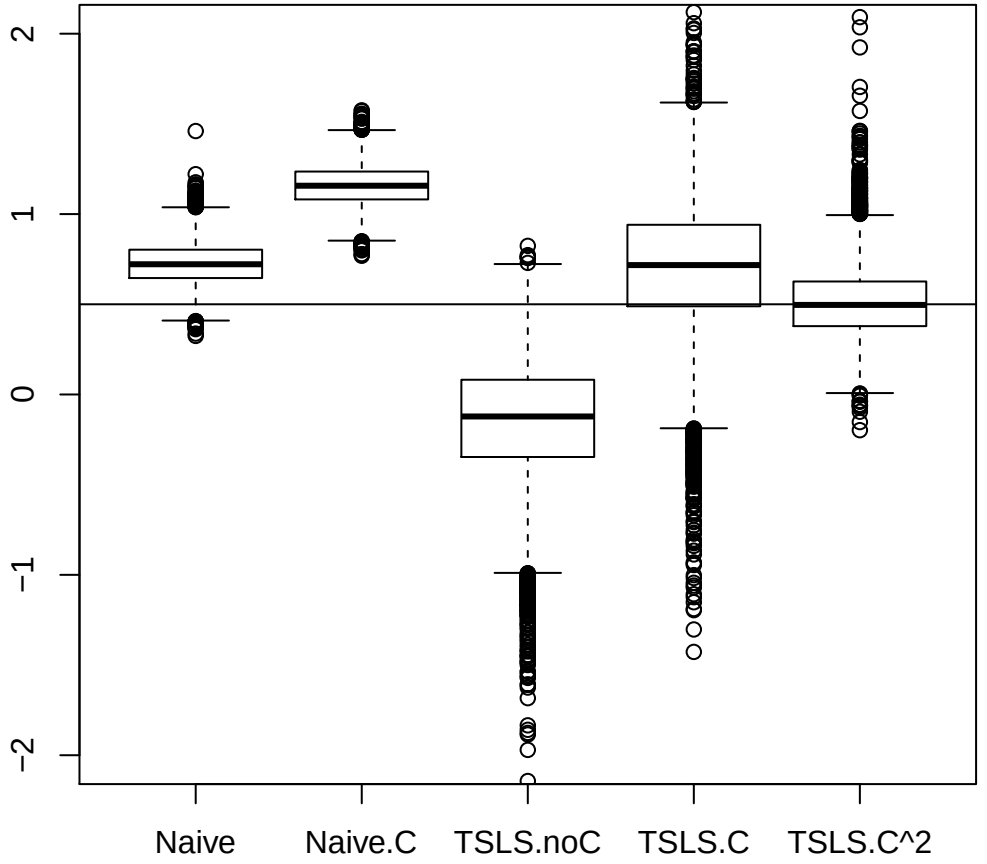
Not Marginal IV — Example for Lack of Robustness

Same setting but now

$E(Z|C) \neq \gamma^\top C$ as

$p(Z = 1) = \text{expit}(-1 + C/2)$

(not independent nor linear)



More Generally: G–Estimation

(Okui et al., 2012)

G–Estimator $\hat{\psi}$ solves (with consistent $\hat{\gamma}$, $\hat{\beta}$)

$$0 = \sum \underbrace{[e(Z,C) - E\{e(Z,C)|C;\hat{\gamma}\}]}_{\text{IV: centred fct. of } Z,C} \underbrace{\left[Y - m(C;\hat{\beta}) - m_y(C;\psi)X\right]}_{\text{residual}}$$

$\Rightarrow$ consistent under $\mathcal{M}_{lin} \cap (\mathcal{A}_y \cup \mathcal{A}_z)$

— either $\mathcal{A}_y$ (i.e. $m(C;\hat{\beta})$) correct

— or $\mathcal{A}_z$ (i.e. model that gives $E\{e(Z,C)|C;\hat{\gamma}\}$) correct,

e.g. if $Z \perp\!\!\!\perp C$ is known (marginal IV). $\Rightarrow$ "double robustness"

Note: TSLS is G–estimator if $\mathcal{A}_z$ is $E(Z|C) = \gamma^\top C$.

Locally Efficient G–Estimation

(Robins, 1994)

Choose $e(Z, C)$ optimally under working models

$\Rightarrow$ requires working model for $\mathcal{A}_x$

$\Rightarrow$ can behave erratically if working model for $\mathcal{A}_x$ wrong.

Empirical Efficiency Maximisation (EEM)

(see Rubin & Van der Laan, 2008; Cao et al., 2009)

In estimating equation

$$0 = \sum \left[\underbrace{e(Z,C)}_{\text{parameterise}} - E\{e(Z,C)|C\} \right] \left[Y - \underbrace{m(C)}_{\text{parameterise}} - m_y(C;\psi)X \right]$$

– parameterise $e(Z,C)$ and $m(C)$;

– find parameter values by minimising asymptotic variance of estimator for ψ (this is a little tedious);

– variance calculations rely on model for $\mathcal{A}_z$.

Note: when $Z \perp\!\!\!\perp C$ known, EEM can only improve efficiency;

Similar argument: adjusting for C in TSLS can only improve efficiency!

Focused Estimation of Nuisance Parameters

(Vermeulen & Vansteelandt, 2014)

– designed to minimise bias if both $\mathcal{A}_z$ and $\mathcal{A}_y$ misspecified;

– i.e. set gradient (wrt. β or γ) of influence function to zero (valid estimating equations, independent of truth);

– either focused on γ (indexing $\mathcal{A}_z$) or on β (indexing $\mathcal{A}_y$).

Notation: FE-γ, FE-β.

Simulation Study

Generated data:

U, C independent $\sim N(0, 1)$

Z binary, $p(Z = 1|C) = \text{expit}(-1 + C/2 + \lambda_z C^2/3)$ (not linear!)

$X|Z, U, C \sim N(\mu_X, 1)$ with $\mu_X = Z + U + C - ZC + \lambda_x C^2$

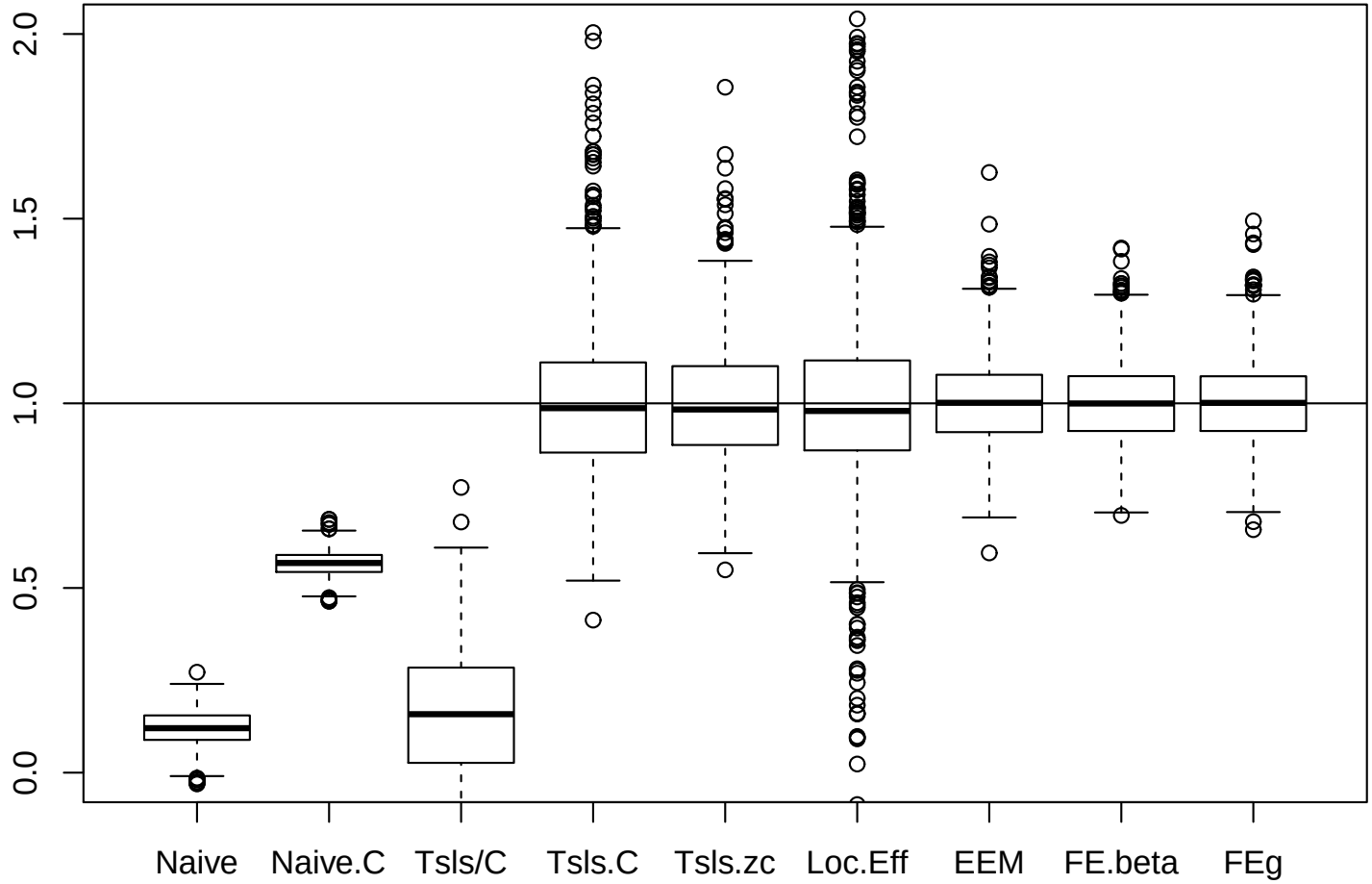
$Y|X, U, C \sim N(\mu_Y, 1)$ with $\mu_Y = X - U - C + \lambda_y C^2$

Note: $\mathcal{A}_x$, $\mathcal{A}_y$ and $\mathcal{A}_z$ assume zero $\lambda_x, \lambda_y, \lambda_z$; .

$n = 500$, $F \approx 49\%$; $R^2 \approx 8.7\%$, reps $= 1000$.

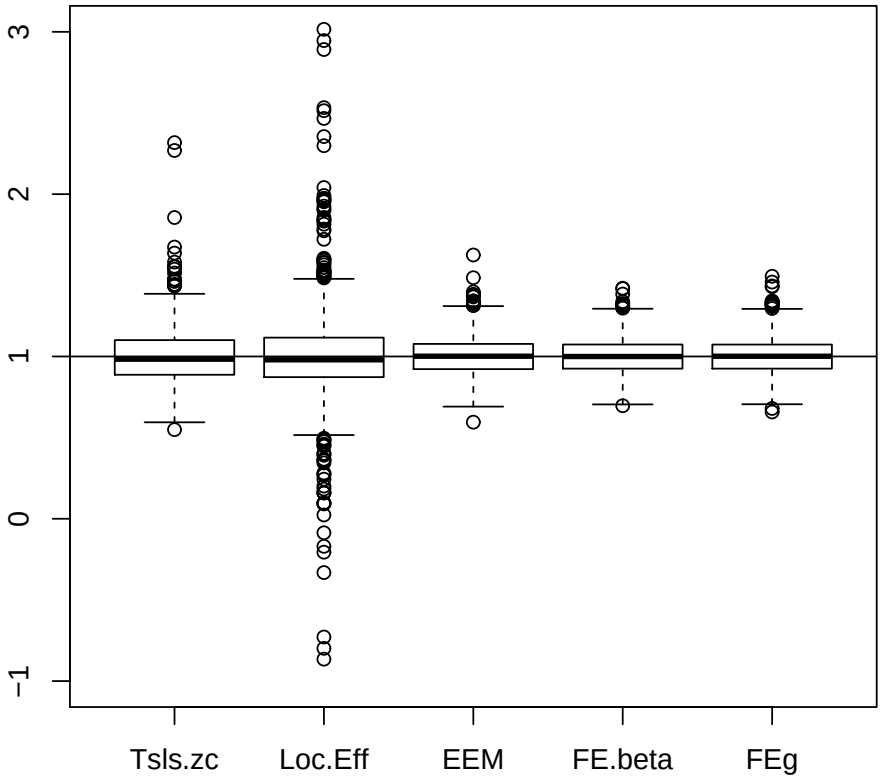
Simulation Study — Some Results

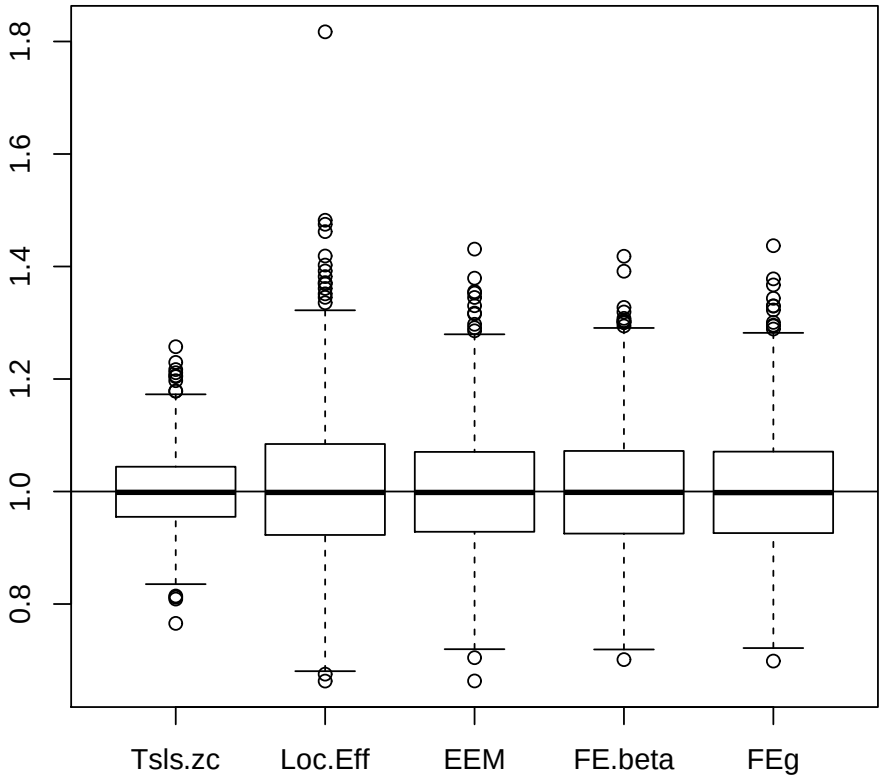
No C^2 terms $(\lambda_x = \lambda_y = \lambda_z = 0)$



Simulation Study — Some Results

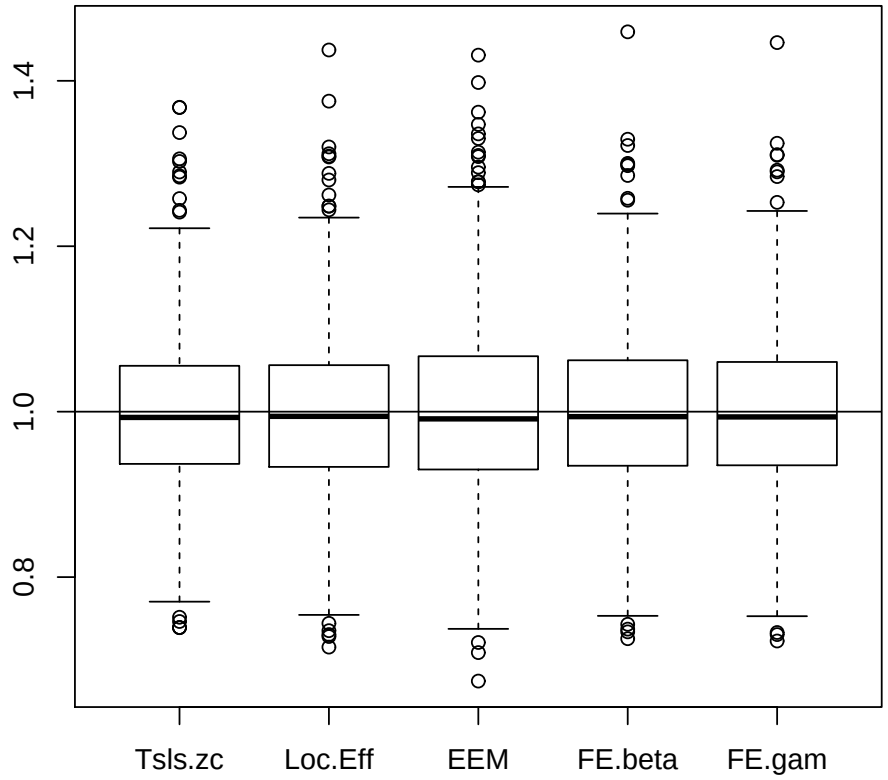
Only exposure model $\mathcal{A}_x$ wrong $(\lambda_x = 1, -1)$

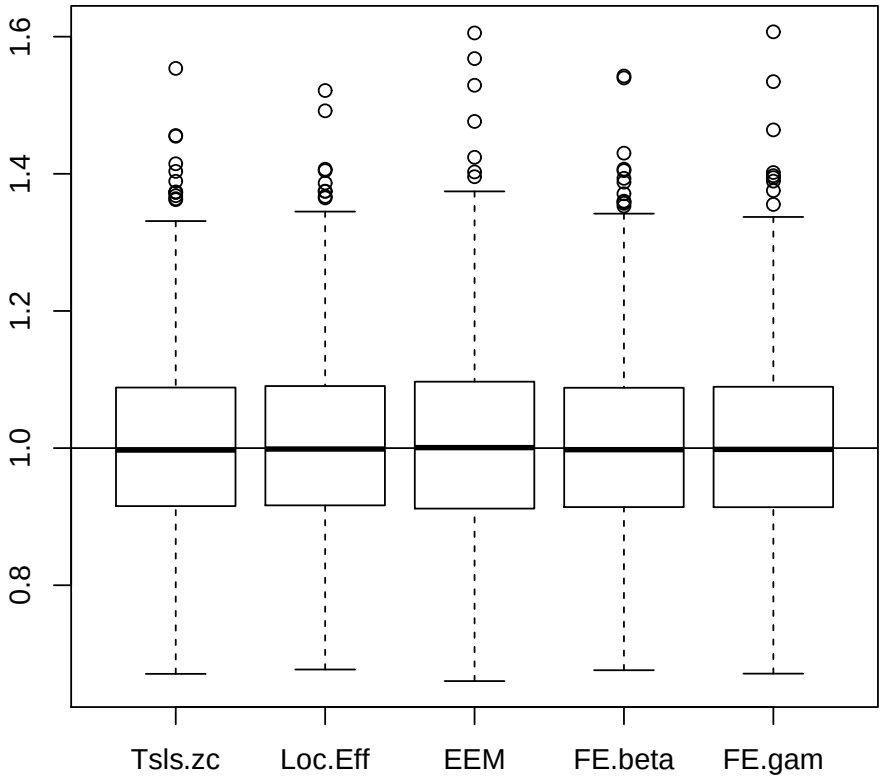




Simulation Study — Some Results

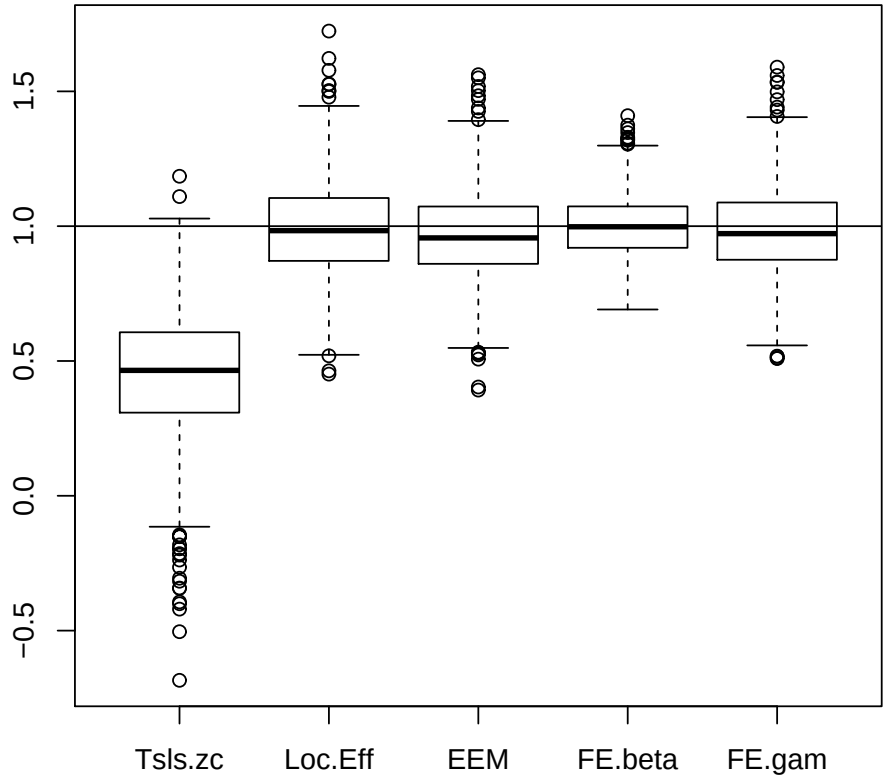
Only instrument model $\mathcal{A}_z$ wrong $(\lambda_z = 1, -1)$

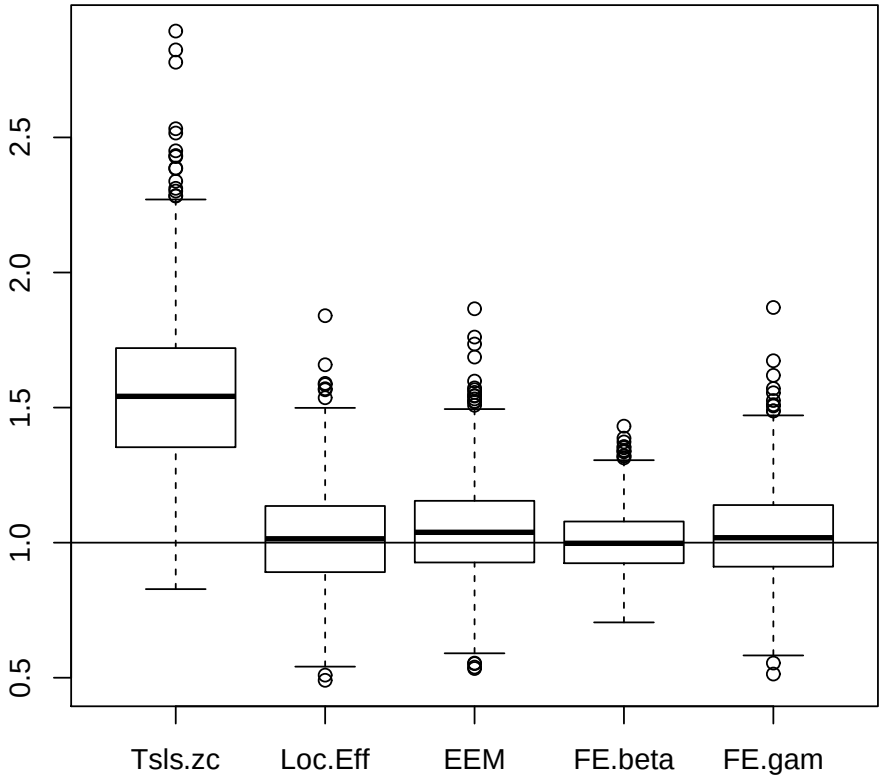




Simulation Study — Some Results

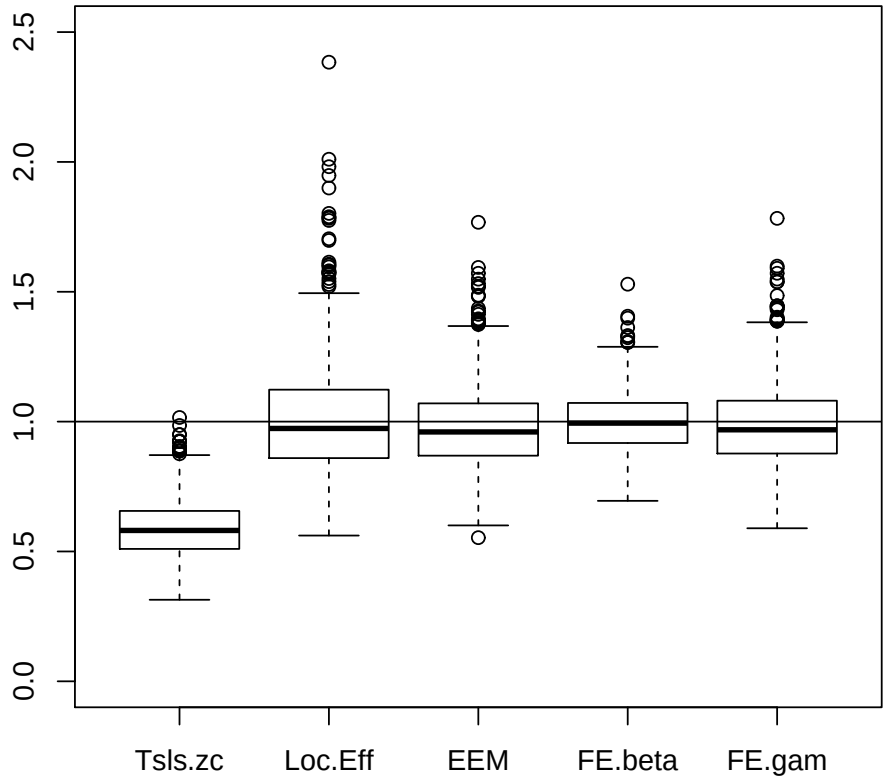
Only outcome model $\mathcal{A}_y$ wrong $(\lambda_y = 1, -1)$

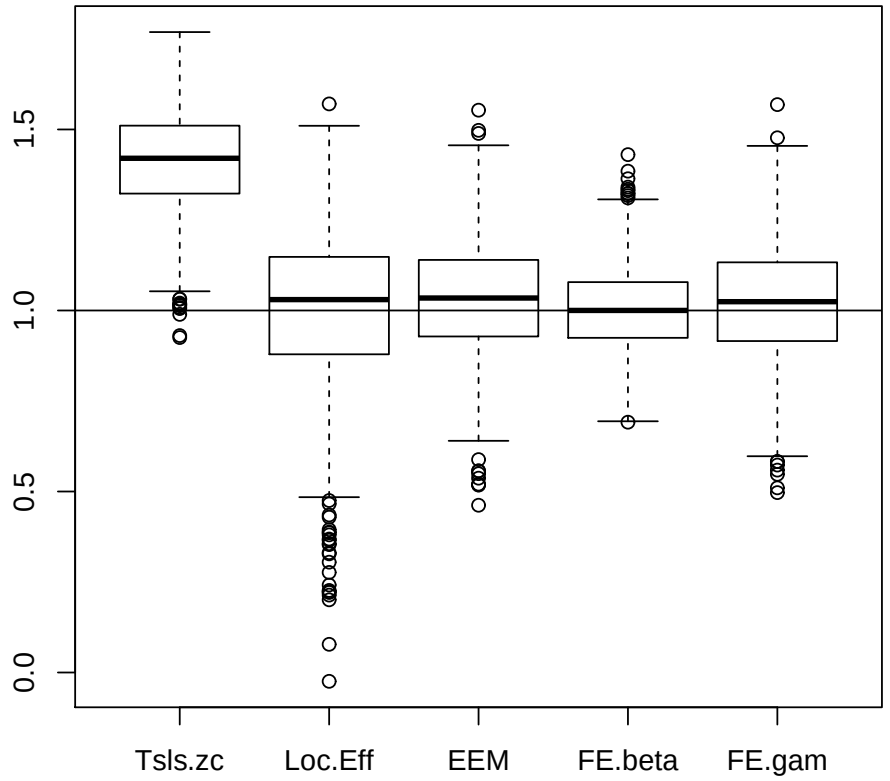




Simulation Study — Some Results

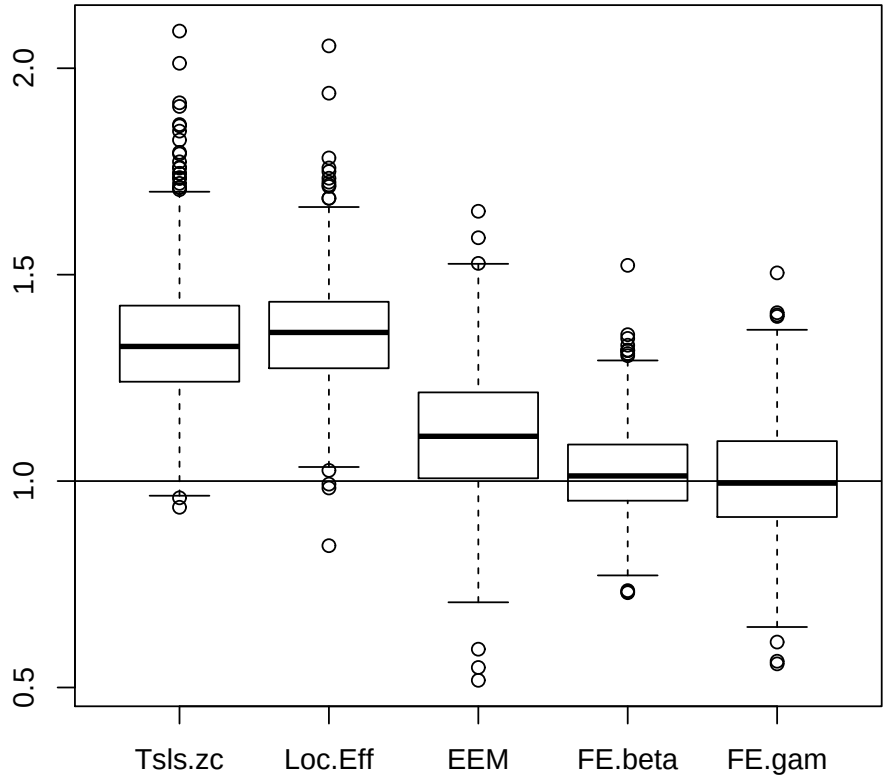
Both exposure and outcome $\mathcal{A}_x, \mathcal{A}_y$ wrong $(\lambda_x = -1, \lambda_y = 1, -1)$

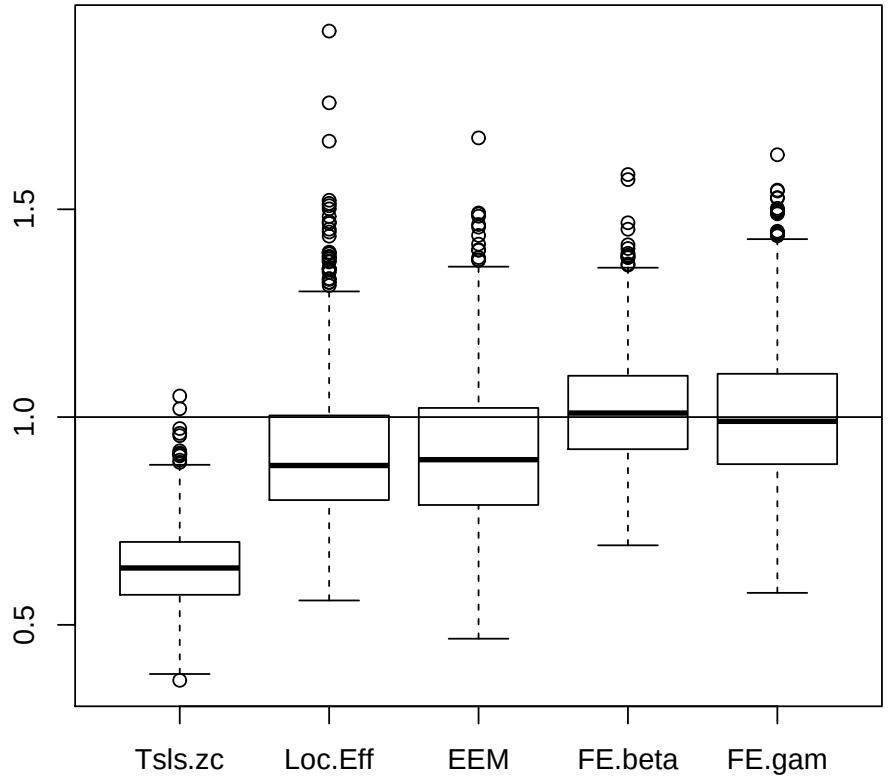




Simulation Study — Some Results

All models wrong $(\lambda_x = \lambda_z = 1, -1, \lambda_y = 1)$





Conclusions (1)

If $Z \perp\!\!\!\perp C$ plausible (as often in MR studies) and assuming $\mathcal{M}_{lin}$:

- TSLS robust towards misspecification of covariate model (and exposure model) & including covariates can only improve efficiency.

- Improvements with other g–estimators can be obtained if:

  (i) non-linear exposure model more plausible (e.g. X binary or count) $\rightarrow$ TSLS not efficient

  (ii) effect modification by covariates $\rightarrow$ TSLS not robust nor efficient

  (iii) if you want to safeguard against *all* models wrong (including possibly $Z \not\perp\!\!\!\perp C$).

Conclusions (2)

- Improvements on TSLS can especially be obtained by more complicated methods when $Z \not\perp\!\!\!\perp C$.

  In MR, if saturated model for $p(Z|C)$ possible $\Rightarrow$ robustness retained.

- Preliminary simulations with high-dim covariates very promising.

- Note: all results asymptotic, but simulations with $n = 500$ look good; more simulations needed.

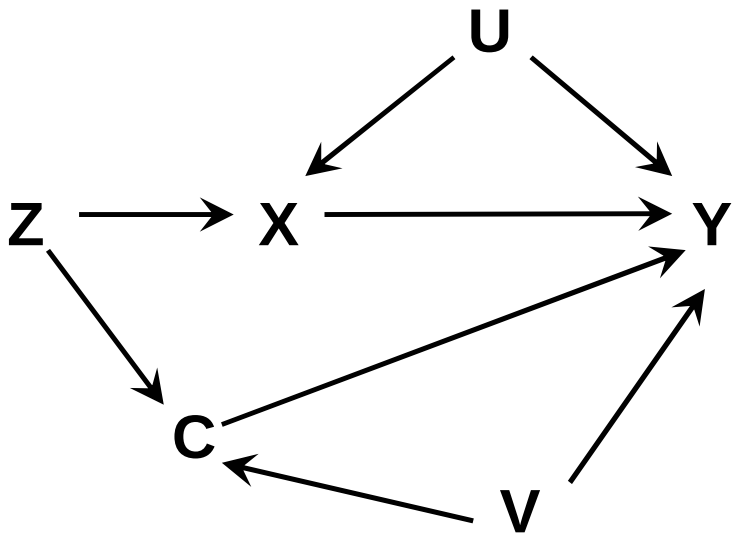
- Would like to investigate multiple instruments more carefully.

- Would like to consider survival outcomes — expect results for additive hazard models similar to TSLS (but not for Cox hazard).

Final Remark

Care has to be taken when covariates are **post-IV**.

Due to collider–bias, adjusting for C will induces new bias.



Example: Want effect of

$X =$ maternal adiposity on

$Y =$ offspring adiposity;

want to use $Z =$ maternal FTO as IV, but clearly Z affects $C =$ offspring FTO which affects Y.

However, offspring FTO is obviously affected by (unobserved) father's FTO (V) which may itself predict offspring adiposity.