Cornell University

Bias due to inclusion of auxiliary variables

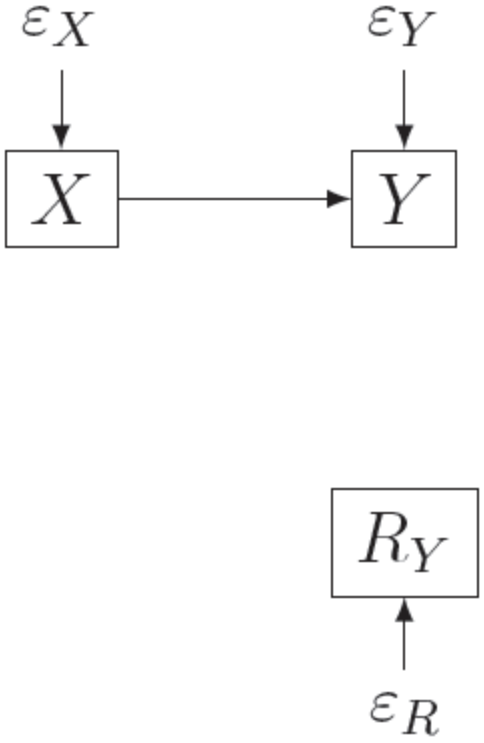
Felix Thoemmes

MCAR

$$P(R \mid Y) = P(R \mid Y_{obs}, Y_{mis}) = P(R)$$

$$R \perp (Y_{obs}, Y_{mis})$$

MCAR

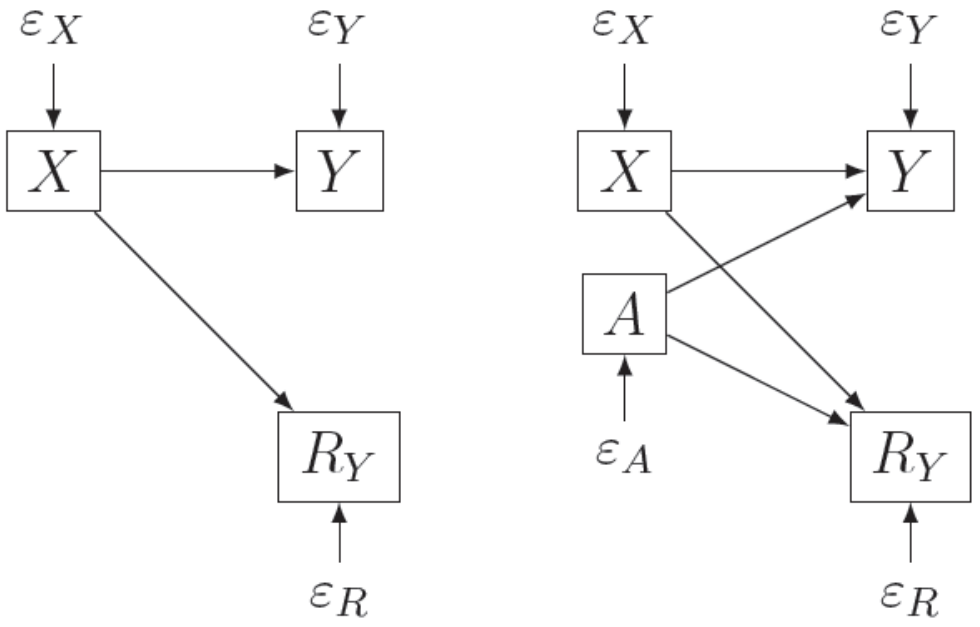


MAR

$$P(R \mid Y) = P(R \mid Y_{obs}, Y_{mis}) = P(R \mid Y_{obs})$$

$$R \perp Y_{mis} \mid Y_{obs}$$

MAR

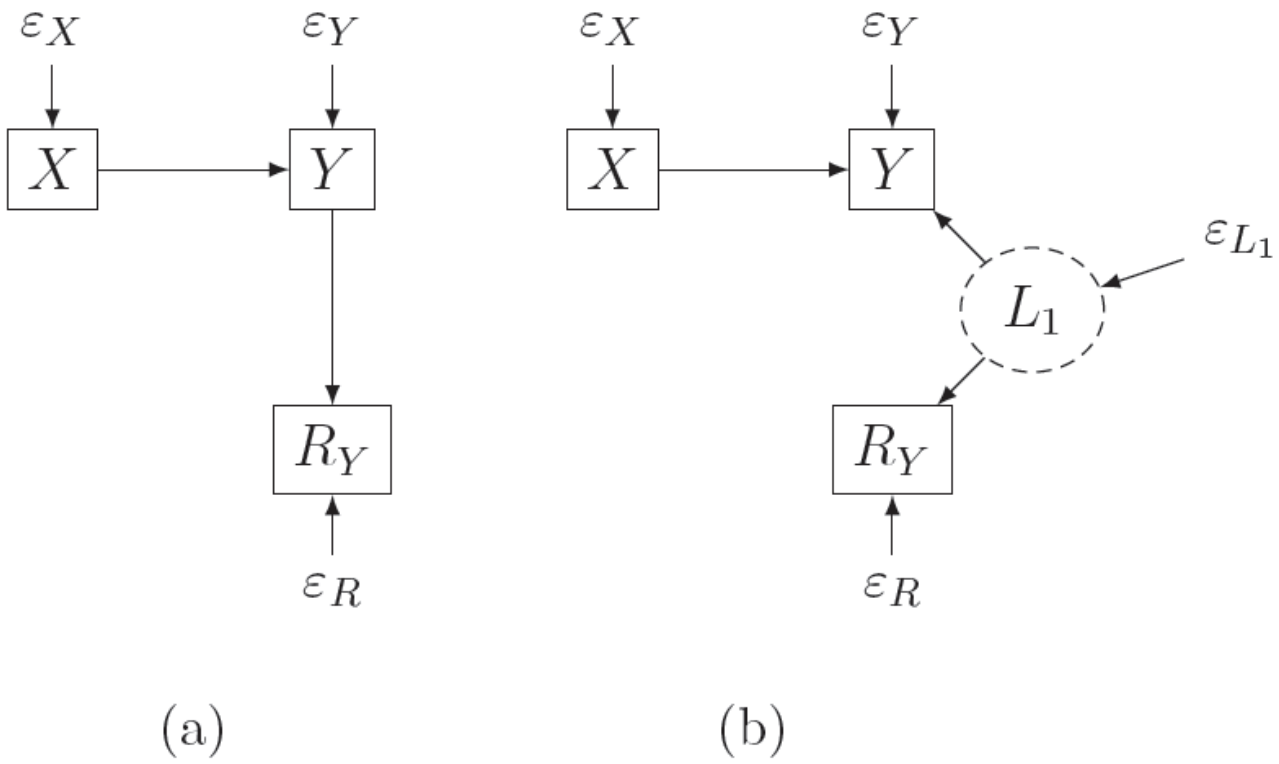

ε_X
ε_Y
X
Y
R_Y
ε_R
ε_X
ε_Y
X
Y
A
ε_A
R_Y
ε_R

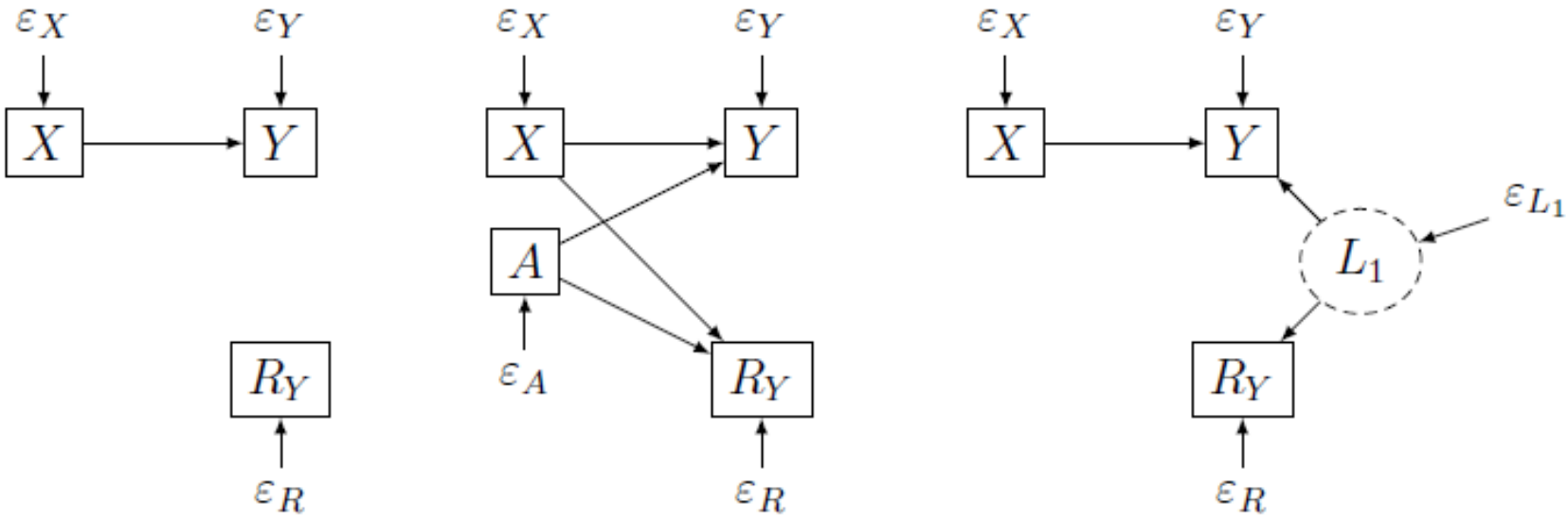
(a) (b)

MNAR

$$P(R \mid Y_{obs}, Y_{mis}) \neq P(R \mid Y_{obs})$$

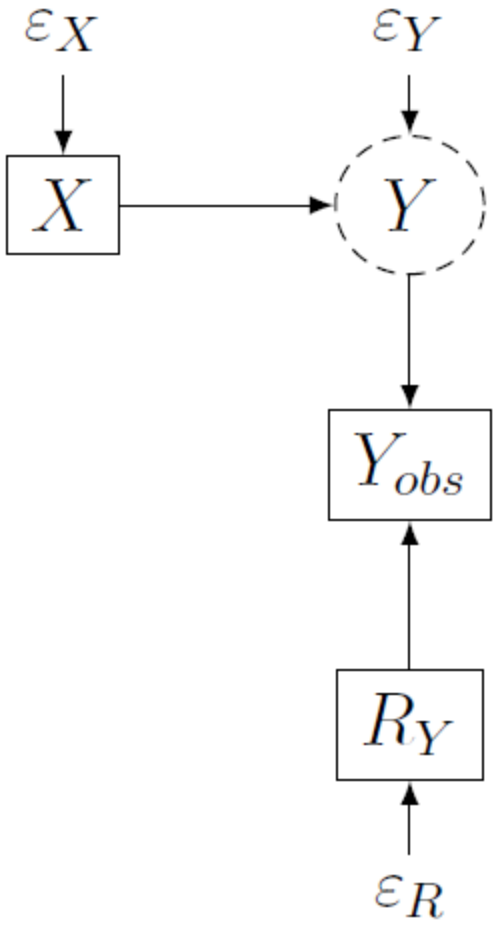
MNAR

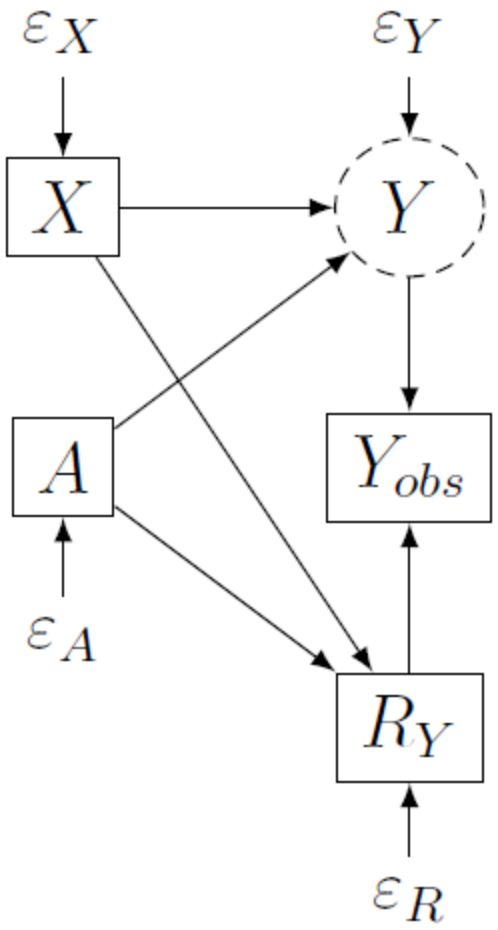

ε_X
ε_Y
X
Y
R_Y
ε_R
(a)
ε_X
ε_Y
X
Y
ε_{L_1}
L_1
R_Y
ε_R
(b)

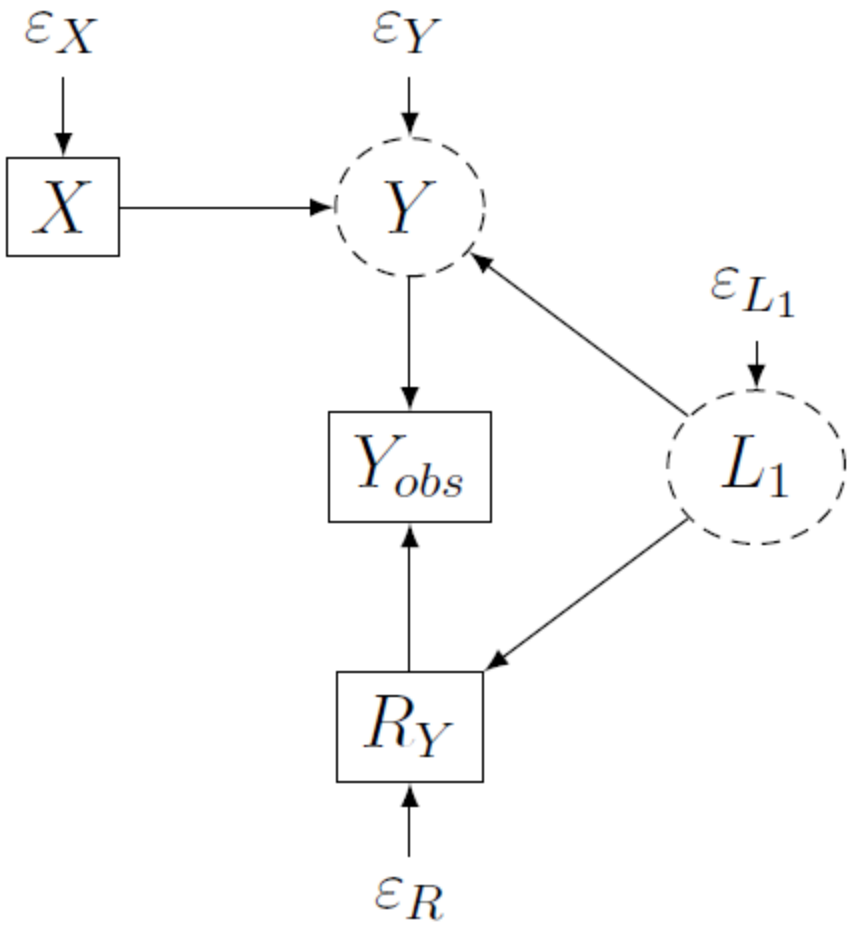


(a) MCAR (b) MAR (c) MNAR

Thoemmes & Rose, 2012; Mohan & Pearl, 2013


ε_X
ε_Y
X
Y
Y_{obs}
R_Y
ε_R

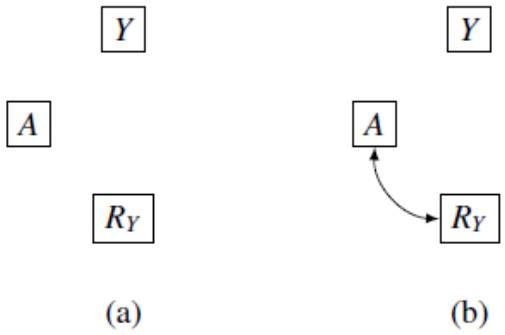

ε_X
ε_Y
X
Y
A
Y_{obs}
ε_A
R_Y
ε_R


ε_X
ε_Y
X
Y
ε_{L_1}
Y_{obs}
L_1
R_Y
ε_R

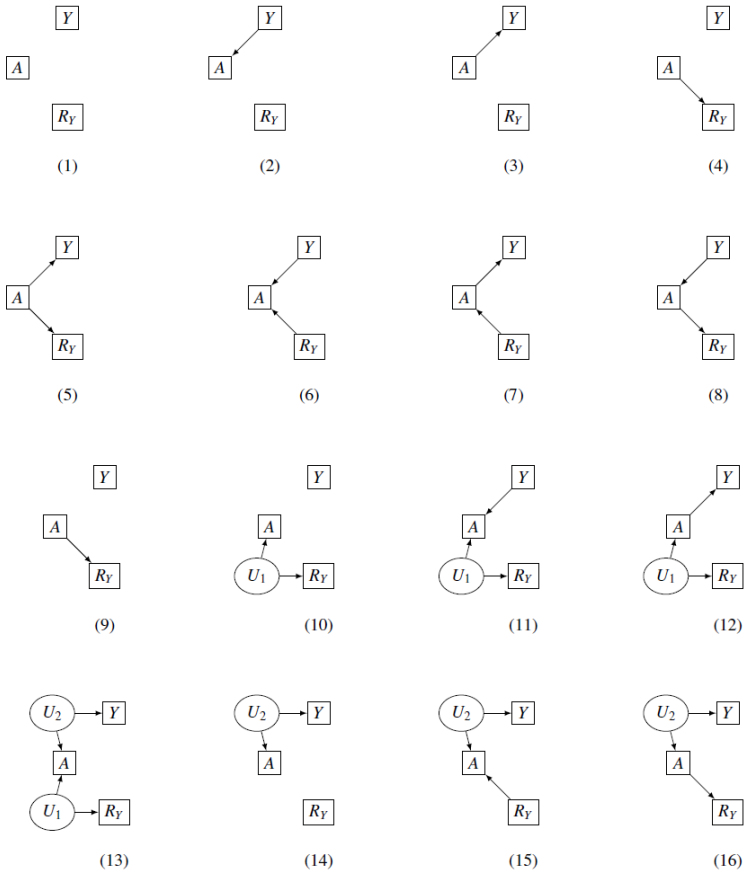
Current approaches

- Inclusive approach (use as many as possible)
- Data-driven approach
  - vanBuuren (correlated with Y or Ry at .1 or larger)
  - Enders (correlated with Y at .4 or larger)

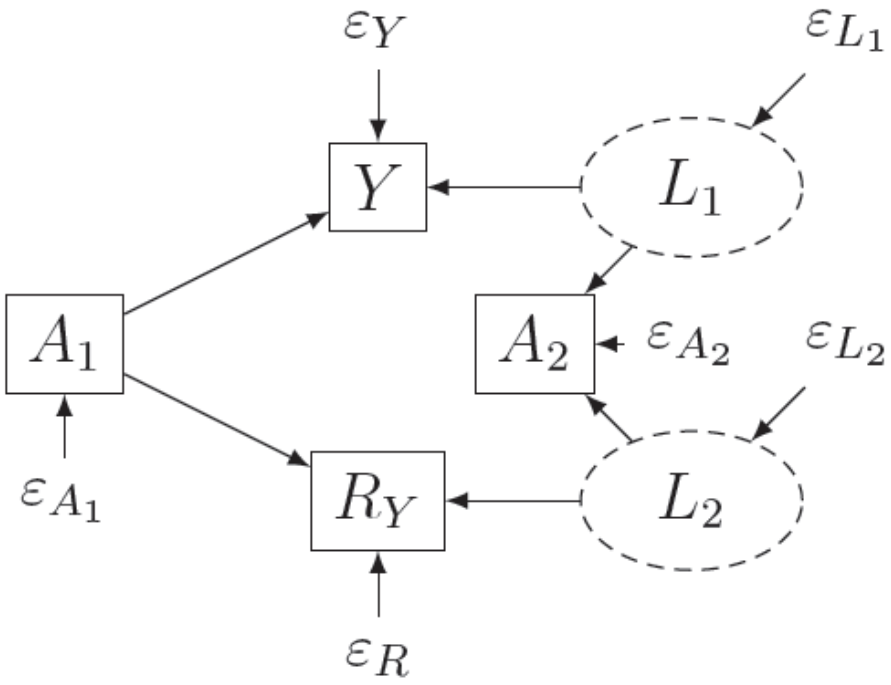
Classic depiction of auxiliary variables



Structural depiction of auxiliary variables



Bias-inducing variables

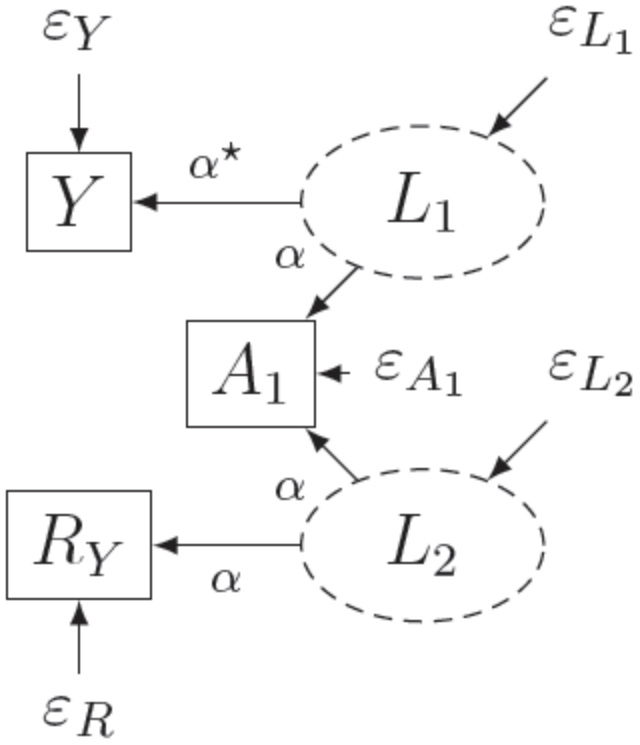


Bias-inducing variables

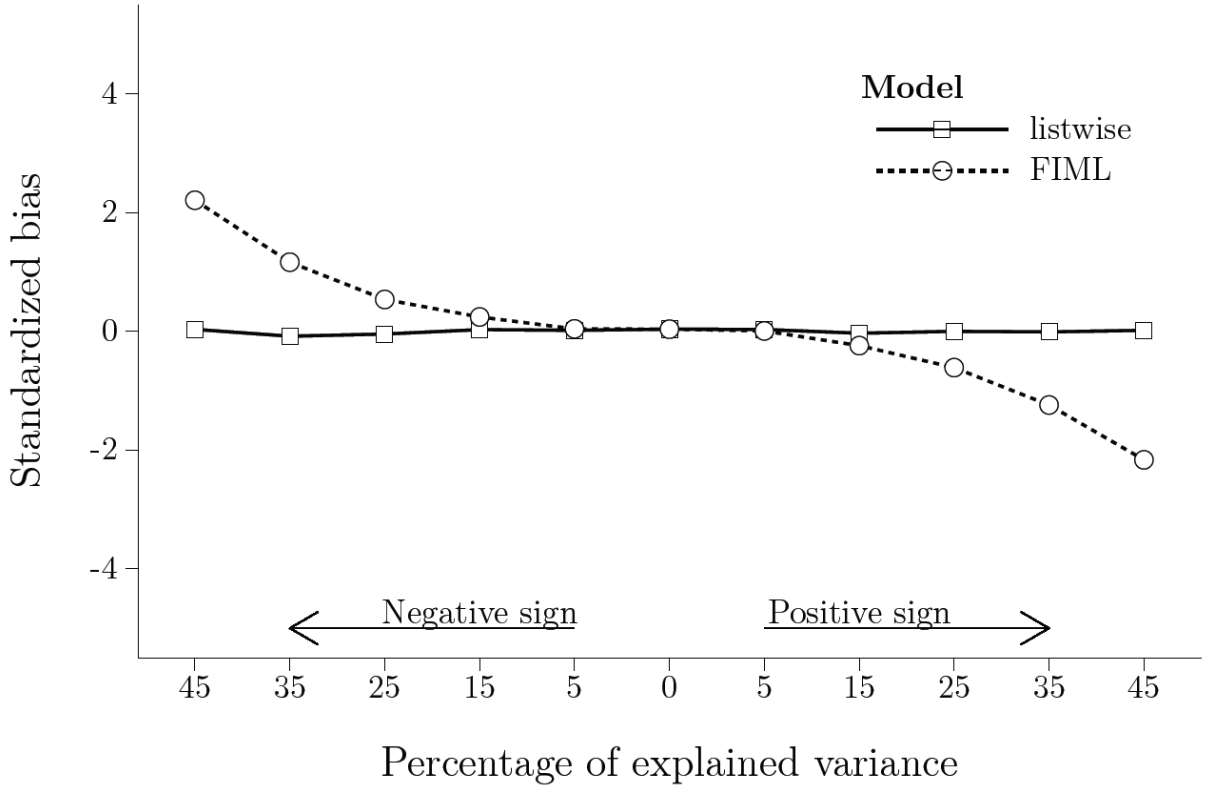
Results of illustrative example of bias-inducing auxiliary variable.

	M (SD)	Raw bias in means	Bias reduction compared to listwise
Complete data	.03 (1.00)		
Listwise	.19 (.98)	.16	
FIML with A_1	.06 (.96)	.03	81%
FIML with A_2	.30 (.98)	.27	−69%
FIML with both	.14 (.98)	.11	31%

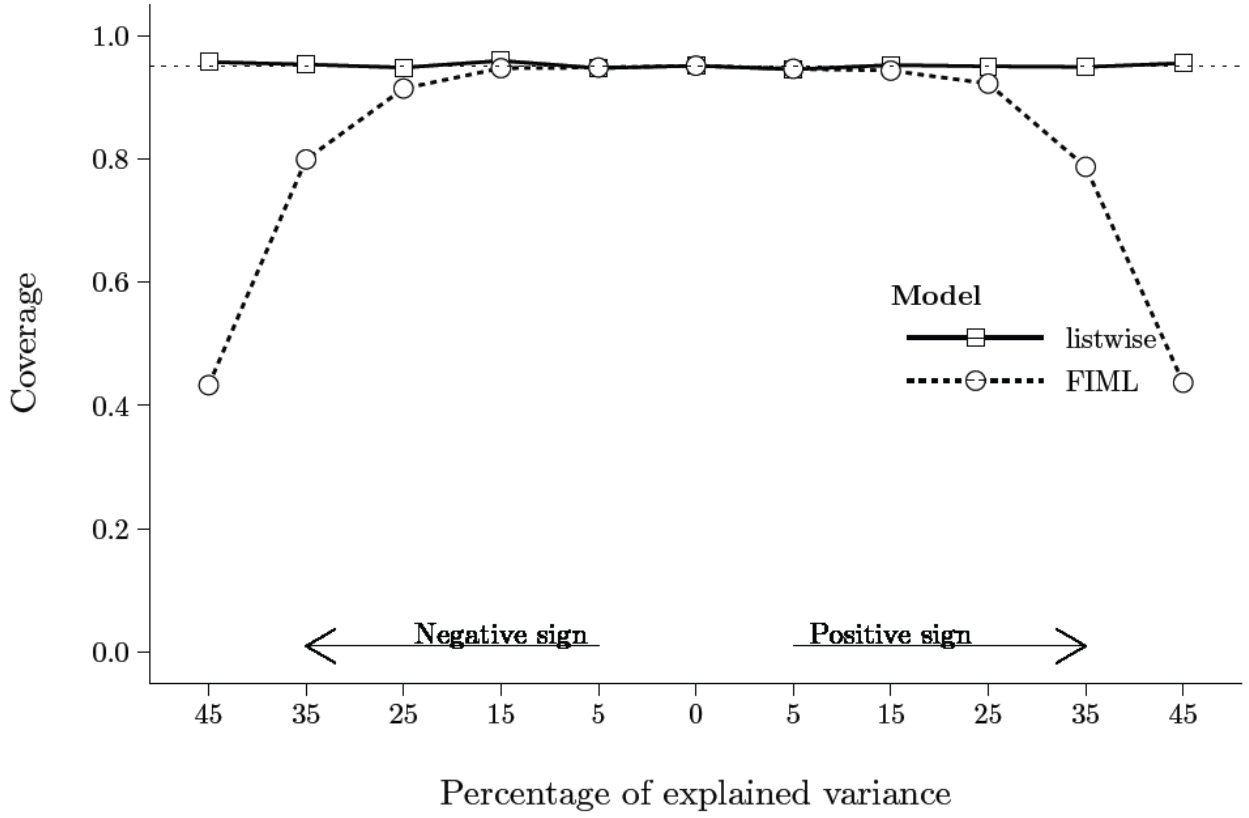
Simulation study 1.1



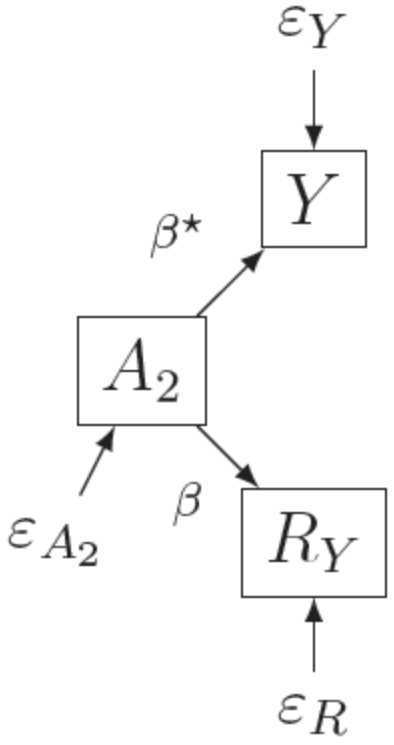
Simulation study 1.1



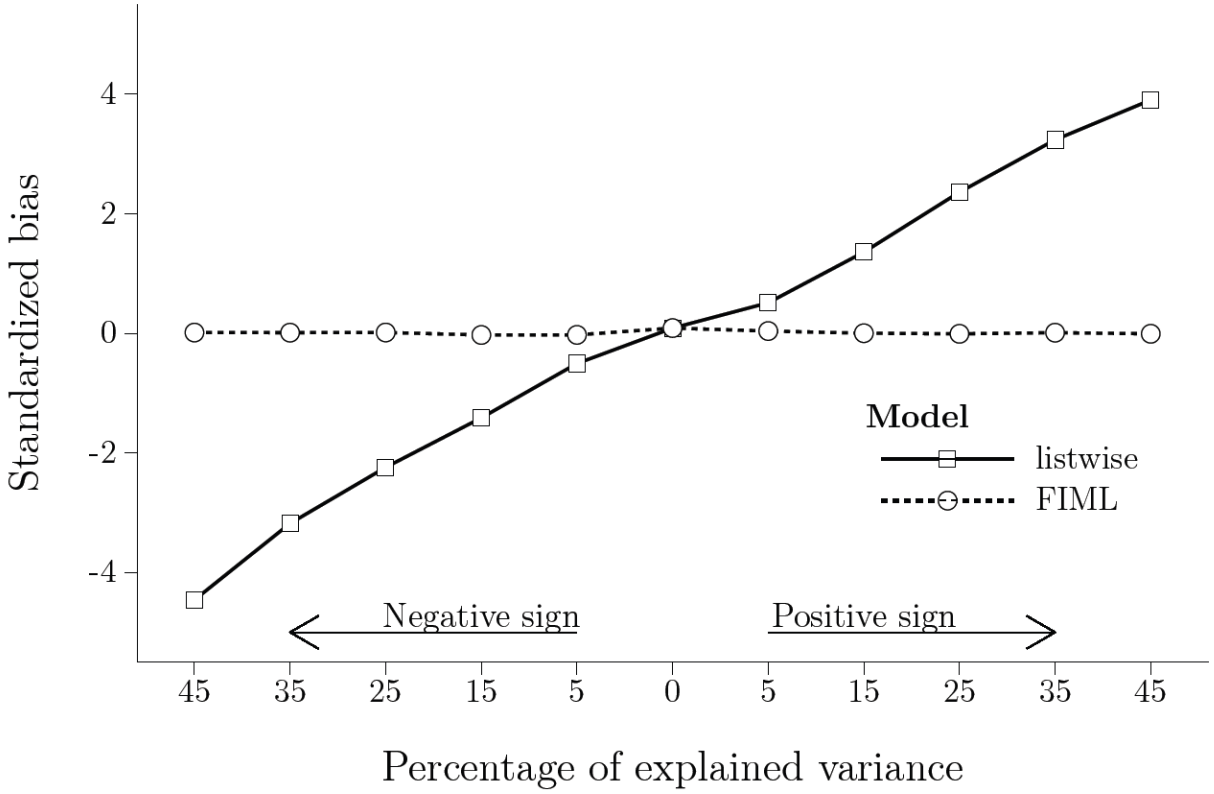
Simulation study 1.1



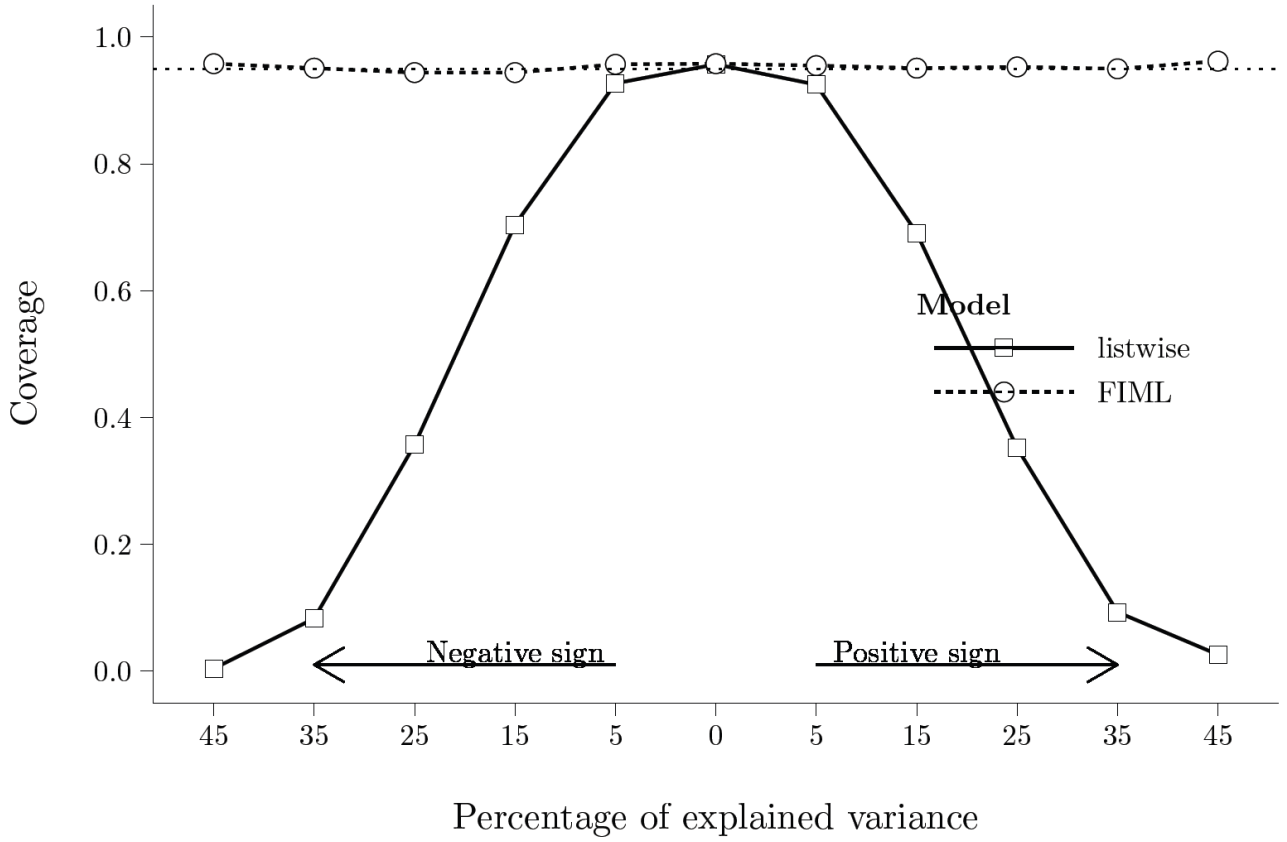
Simulation study 1.2



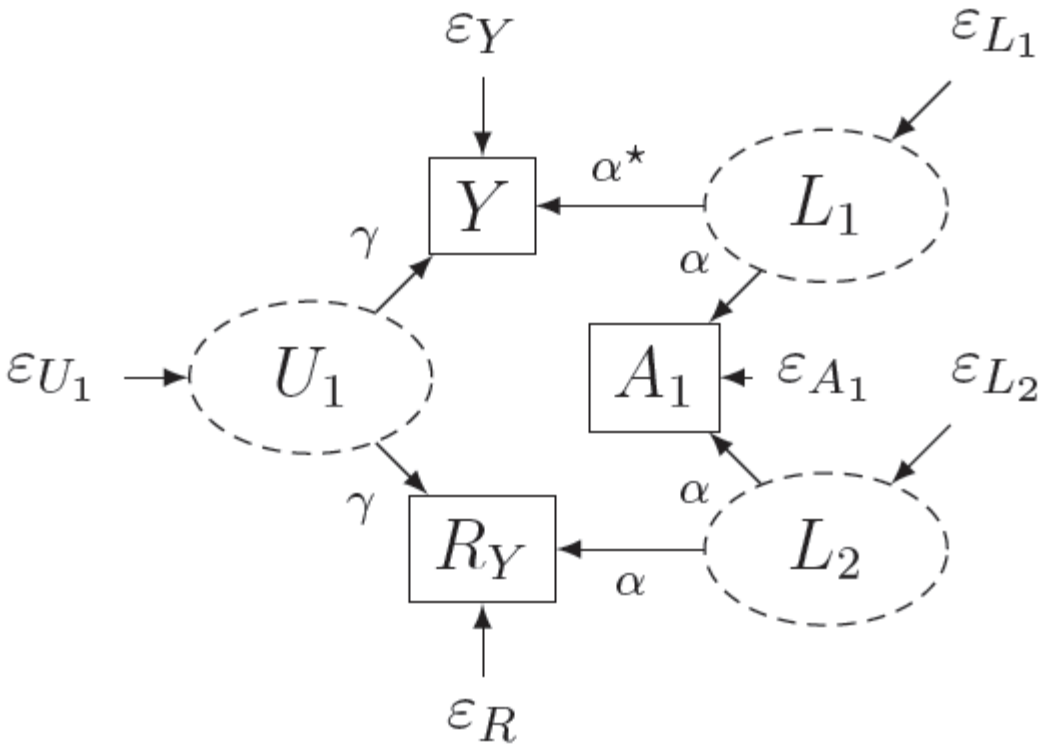
Simulation study 1.2



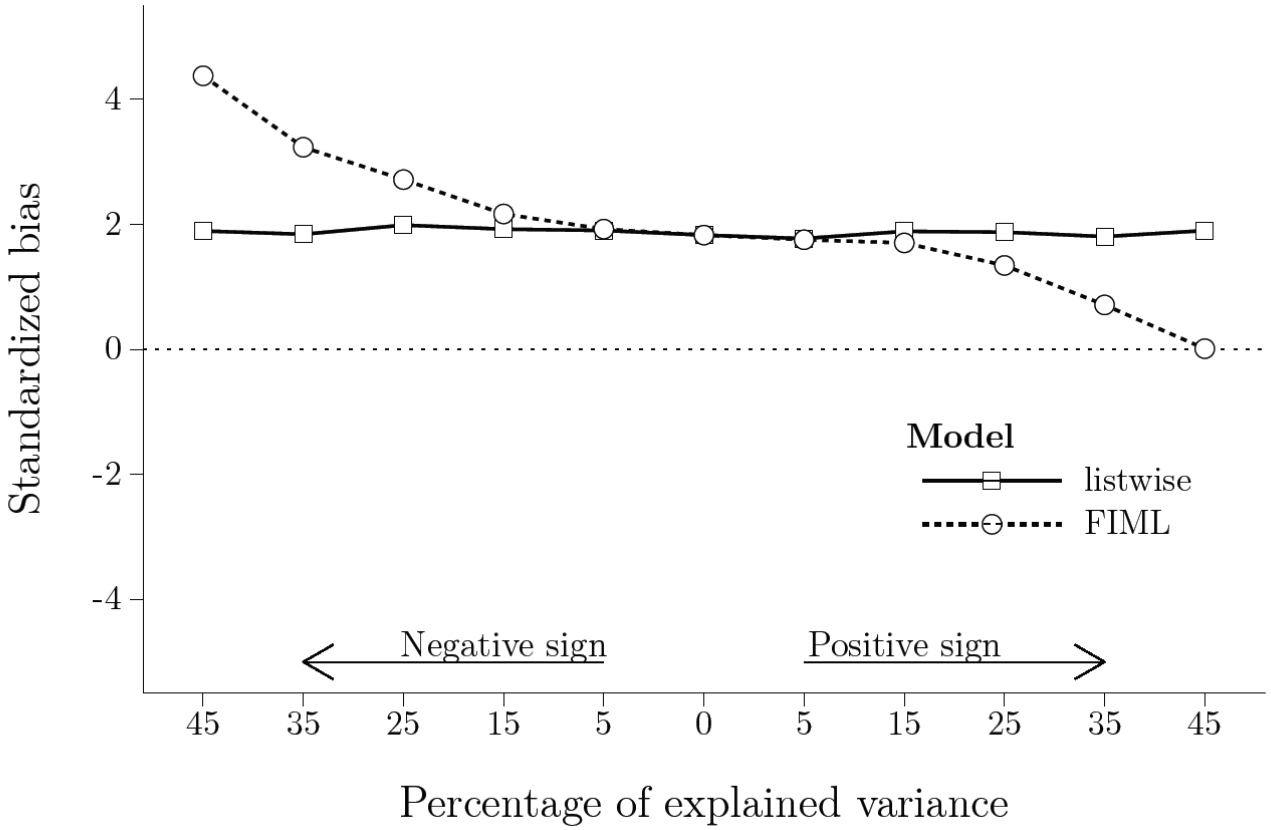
Simulation study 1.2



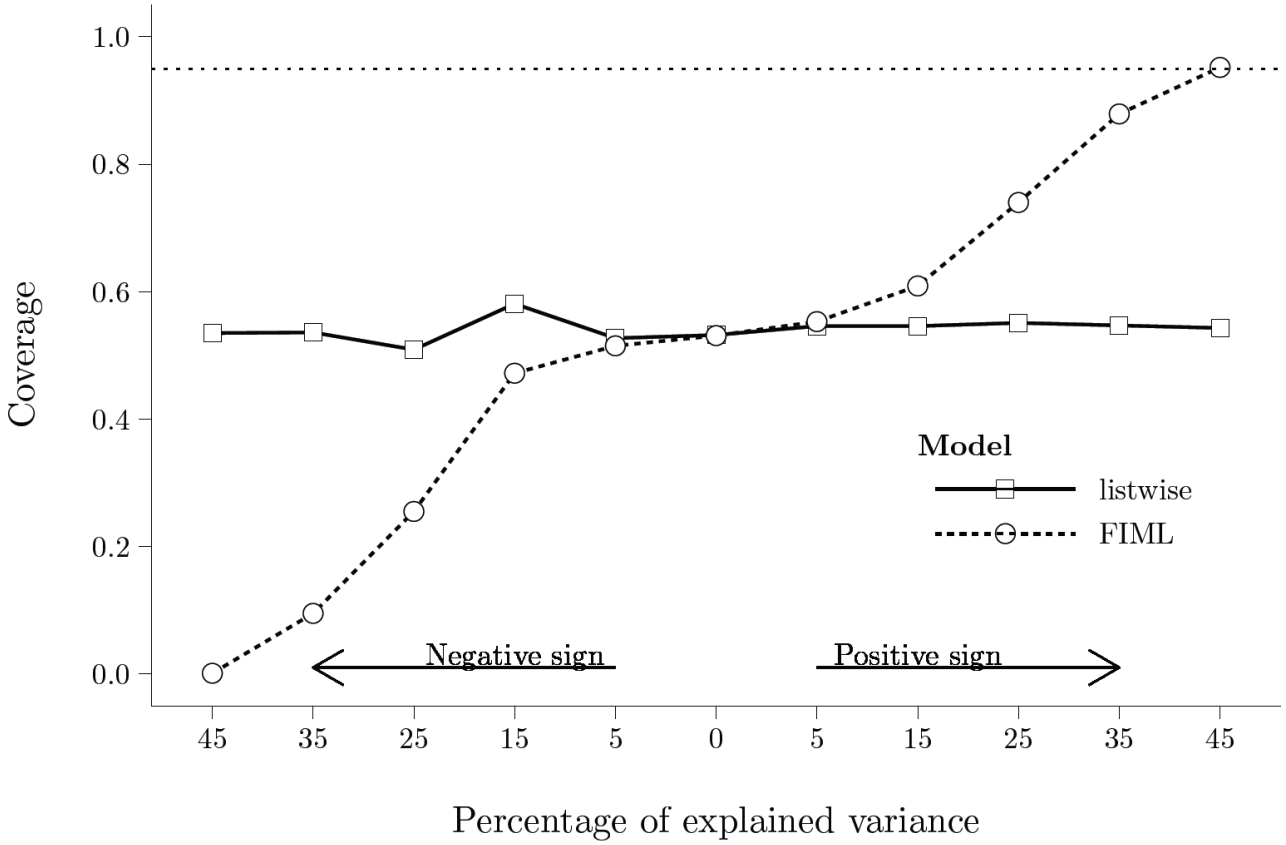
Simulation study 2.1



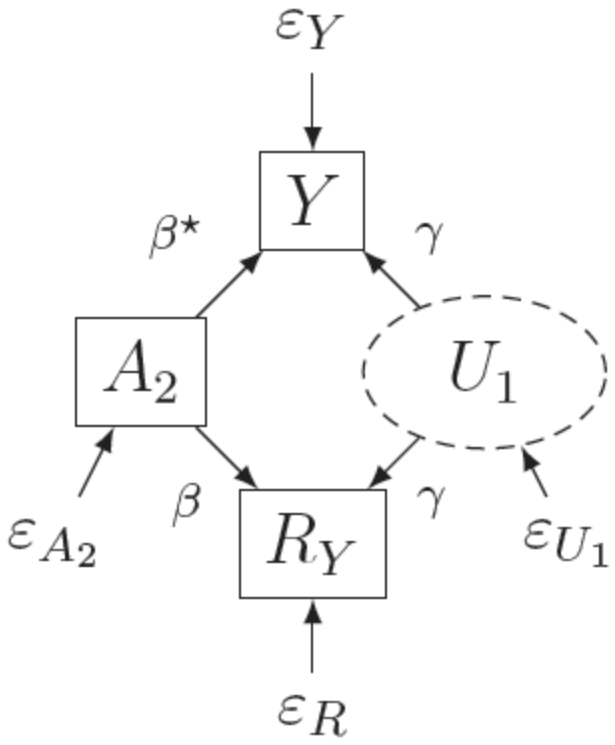
Simulation study 2.1



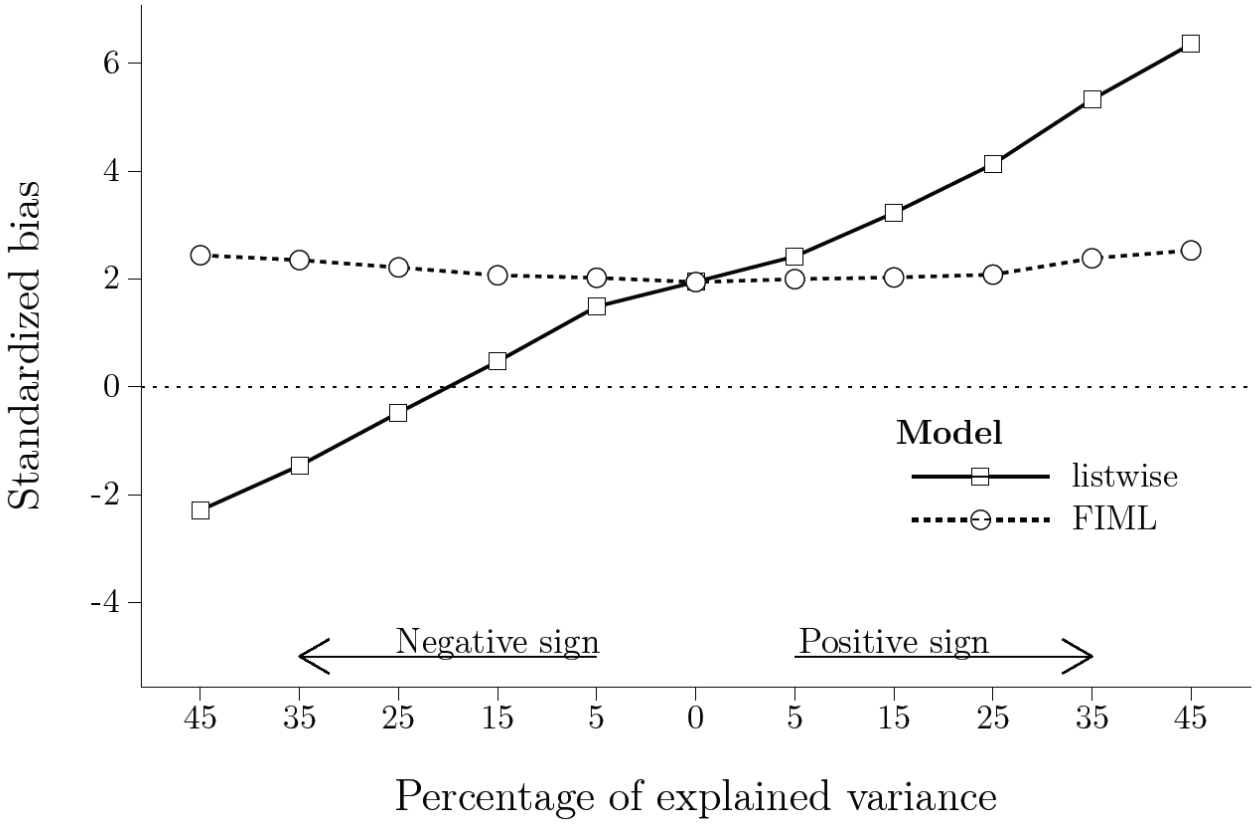
Simulation study 2.1



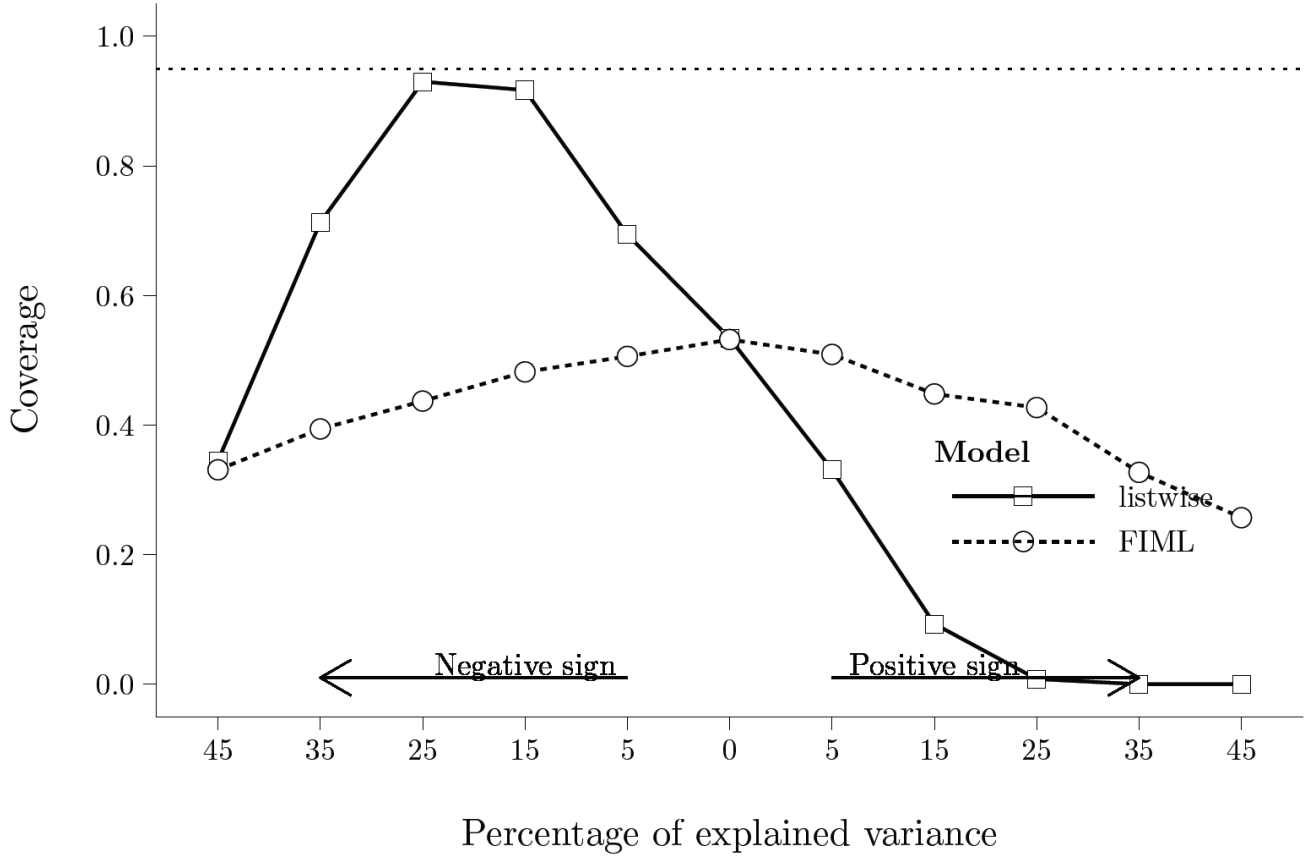
Simulation study 2.2



Simulation study 2.2



Simulation study 2.2



Results

- In cases of MCAR / MAR, bias can be induced through auxiliary variables that induce dependencies

- In cases of MNAR, bias can be reduced or induced through auxiliary variables that exhibit various structural relatioships

Results

- Correlational properties of variables can only be a weak guide as to which variables will reduce or induce bias
- A helpful and a hurtful auxiliary variable are statistically indistinguishable
- Theoretical assumptions about structure have to be invoked

Results

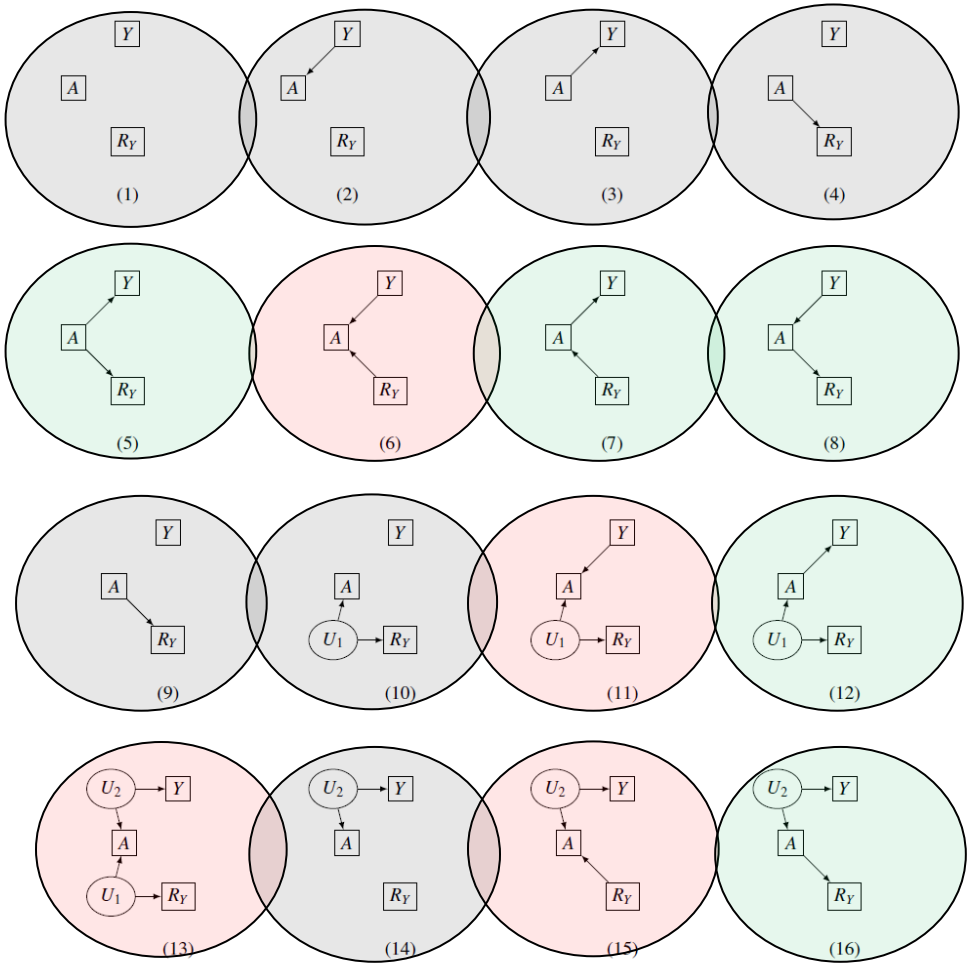
- If one assumes that MAR holds (given a set of analysis and auxiliary variables), then one should only include variables that are believed to reduce the dependencies between missingness and variables with missing data

Results

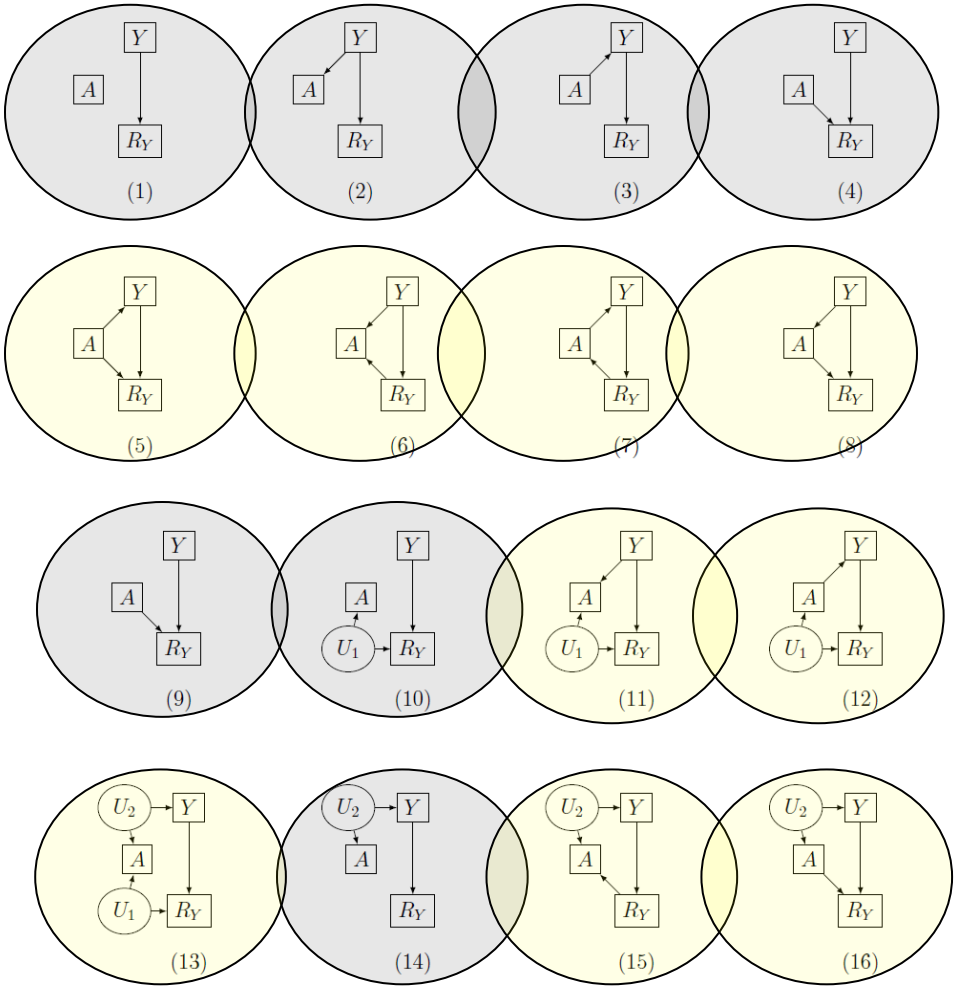
- If one assumes that MAR may not hold, any variable can induce or reduce bias – unless we know sign and magnitude of relationship

- Inclusive approach as a last resort, given our ignorance?

Structural depiction of auxiliary variables



Structural depiction of auxiliary variables



Discussion

- What is a **theoretically** defensible strategy for the selection of auxiliary variables?
  - Correlation is not a sufficient statistic to determine inclusion
  - Assumptions need to be made – in some cases about qualitative relationships between variables, in other cases sign and magnitude of relationships are needed
  - “Structural approach” can in theory select correct covariates, if correct structure is postulated
  - MAR is an assumption that should not be invoked without strong theoretical arguments

Discussion

- What is a **practically** defensible strategy for the selection of auxiliary variables?
  - We don't know
  - Maybe using variables strongly related to outcome
  - Maybe using all and hope that they cancel out
  - Maybe make weak assumptions and select based on structure

Cornell University

Want to find out?

felix.thoemmes@cornell