Estimating causal effects under untestable assumptions with non-ignorable missing data

UK Causal Inference meeting 2017

Manuel Gomes

April 7, 2017

THETA
Team for Health Economics
Policy and Technology Assessment

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

theta.lshtm.ac.uk

- Causal inference with non-ignorable missing data

- Selection models

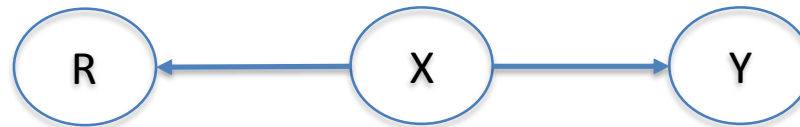- The role of the exclusion restriction

- The REFLUX study

- Discussion

Non-ignorable missing data

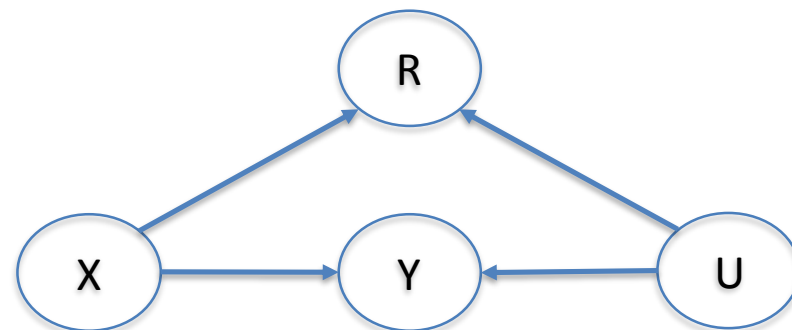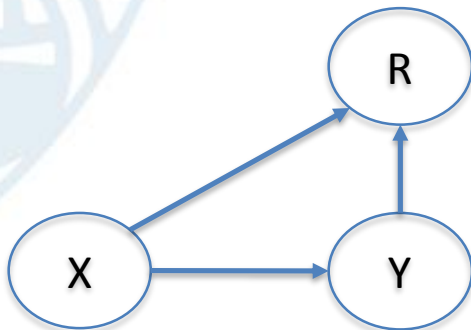Suppose we would like to estimate the causal effect of Y on X, but Y is partially observed.

$$f(Y \mid X; \boldsymbol{\beta}) \qquad R = \begin{cases} 1 \ if \ Y \ is \ observed \\ 0 \ if \ Y \ is \ missing \end{cases}$$

Ignorable missing data (Missing at random - MAR)

Non-ignorable missing data (Missing not at random - MNAR)

Joint model of Y and $\mathbf{R}$ is required with non-ignorable missing data

$$f(Y, R \mid X; \boldsymbol{\theta})$$

<u>Why selection models?</u>

   - commonly used in social and health sciences

   - familiarity with the Heckman selection model

   - follows naturally from the substantive and missing data models:

$$f(Y, R \mid X; \boldsymbol{\theta}) = f(Y \mid X; \boldsymbol{\beta}) \, P(R \mid Y, X; \boldsymbol{\gamma})$$

Heckman model (2-step)

Y_{1i} - outcome; Y_{2i} - latent variable ($R_i = 1$ if $Y_{2i} > 0$; 0 otherwise)

$$Y_{1i} = \beta X_i + \varepsilon_i$$
$$Y_{2i} = \gamma Z_i + v_i$$

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon \\ & 1 \end{pmatrix} \right)$$

$$E(Y_{1i}|X_i, Y_{2i} > 0) = \beta X_i + \rho\sigma_\varepsilon \lambda_i \qquad \lambda_i = \frac{\phi(\gamma Z_i)}{\Phi(\gamma Z_i)}$$

Step 1: Regress Y_{2i} on Z_i (probit) in the full sample to estimate $\hat{\gamma}$ and construct $\hat{\lambda}_i$

Step 2: Estimate parameters of interest ($\hat{\beta}$) in the observed sample from

$$Y_{1i} = \beta X_i + \beta_\lambda \hat{\lambda}_i + \varepsilon_i$$

Consistent variance: $V(\varepsilon_i|X_i, Y_{2i} > 0) = \sigma_\varepsilon^2(1 - \rho^2(\lambda_i^2 + \gamma Z_i \lambda_i))$

Joint full maximum likelihood

Substantive model

$$Y_i = \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

Missing data model

$$\text{logit}(P(R_i = 1)) = \gamma Z_i + \alpha Y_i$$

Equivalent to Heckman selection specification (derivation in Appendix)

- Models are often jointly estimated via EM or MCMC techniques.

Log likelihood:

Marginal probability that $R_i = 0$

Joint probability of Y_i and $R_i = 1$

$$log\, L\,(\beta, \gamma, \alpha, \sigma_\varepsilon) = \sum_{R=0} \log[1 - \Phi(\gamma Z_i)] + \sum_{R=1} \left[-log\sigma_\varepsilon + \right.$$

Multiple imputation using Heckman

Imputed values drawn from:

$$Y_i^{miss} \sim N(\beta X_i + \beta_\lambda \lambda_i', \sigma_\varepsilon^2), \qquad \lambda_i' = \frac{-\phi(\gamma Z_i)}{1 - \Phi(\gamma Z_i)}$$

λ_i' - inverse Mills ratio derived from conditional expectation of $R_i = 0$

Algorithm

1) Run Heckman's first step (probit) and compute $\hat{\lambda}_i' = \frac{-\phi(\hat{\gamma} Z_i)}{1 - \Phi(\hat{\gamma} Z_i)}$

2) Fit Heckman's second step (OLS) to estimate $\hat{\beta}, \hat{\beta}_\lambda, \hat{\sigma}_\varepsilon$

3) Compute Bayesian posterior draws for $\beta^*, \beta_\lambda^*, \sigma_\varepsilon^*$ in the standard way

4) Draw ε from $N(0, \sigma_\varepsilon^{2*})$

5) For each Y_i^{miss}, impute Y_i^* from $Y_i^* = \beta^* X_i + \beta_\lambda^* \hat{\lambda}_i' + \varepsilon^*$

6) Repeat steps 1) to 5) M times to obtain M imputed datasets

7) Apply substantive model to each imputed dataset and the resulting estimates can combined using Rubin's rules as usual.

Key assumptions in selection models

- Parametric assumptions
  - Distributional assumptions for (ε_i, v_i), link function, etc.
  - Been largely addressed (semi-parametric, non-parametric approaches)

- Exclusion restrictions
  - Variables that predict R but are (conditionally) independent of Y
  - Heckman's performance shown to depend on valid exclusion restrictions when handling sample selection
  - Unclear to what extent Heckman relies on this in MNAR settings
  - Importance of these variables to other selection models received little attention

AIM: - Critically assess the role of exclusion restrictions in selection models across typical MNAR settings.

Design – MNAR settings

Simulation study
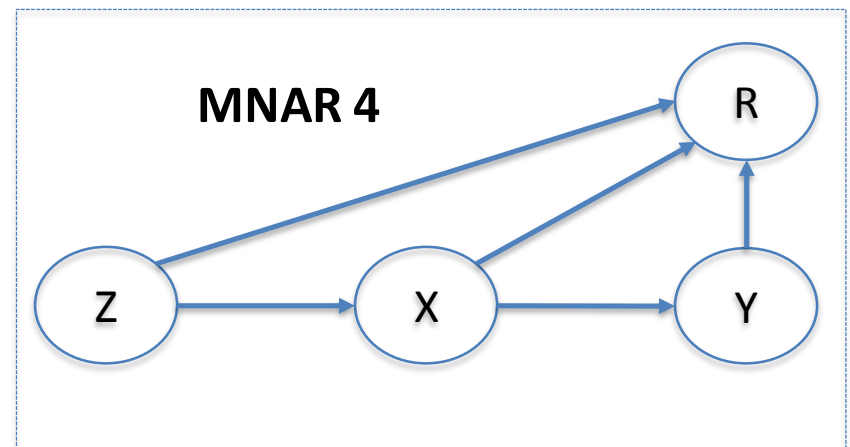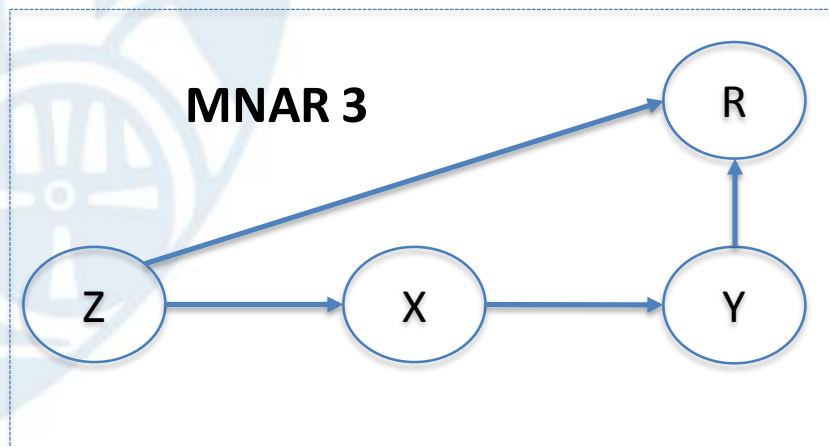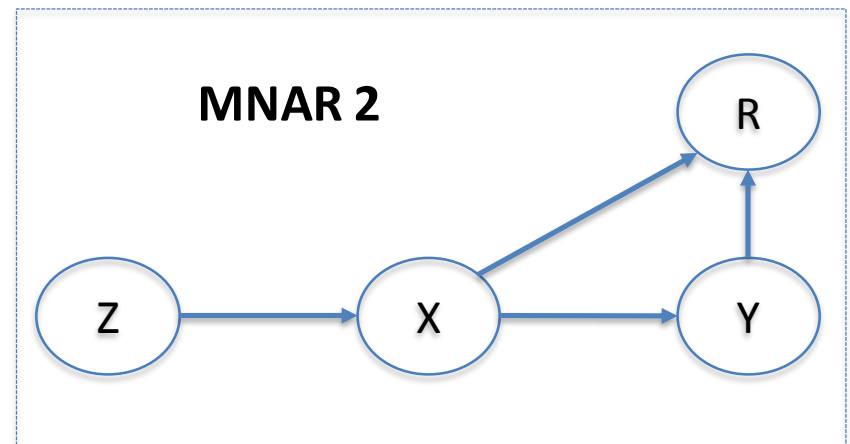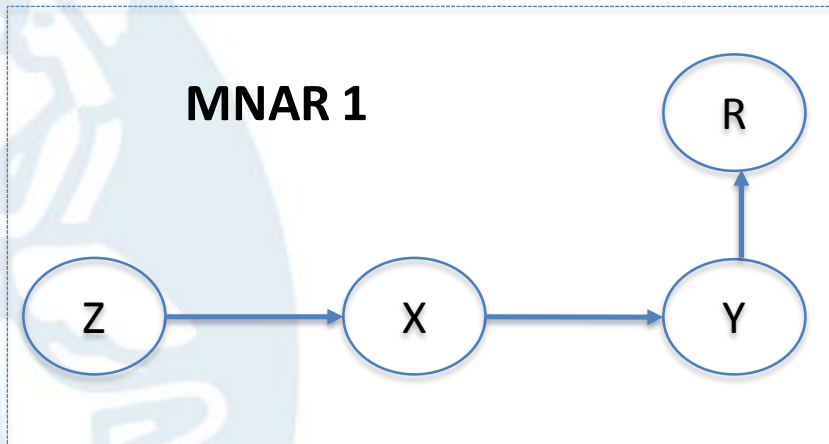
Scenarios (within each mechanism)

- <u>Scenario A</u>: 20% missing, $cor(Y, R), = 0.1$ (weak MNAR), $cor(Z, X) = 0.7$
- <u>Scenario B</u>: 20% missing, $cor(Y, R), = 0.4$ (strong MNAR), $cor(Z, X) = 0.7$
- <u>Scenario C</u>: 50% missing, $cor(Y, R), = 0.4$ (strong MNAR), $cor(Z, X) = 0.7$
- <u>Scenario D</u>: 50% missing, $cor(Y, R), = 0.4$ (strong MNAR), $cor(Z, X) = 0.3$

Methods

- MI (assuming MAR)
- 2-step Heckman approach
- MI based on the 2-step Heckman procedure
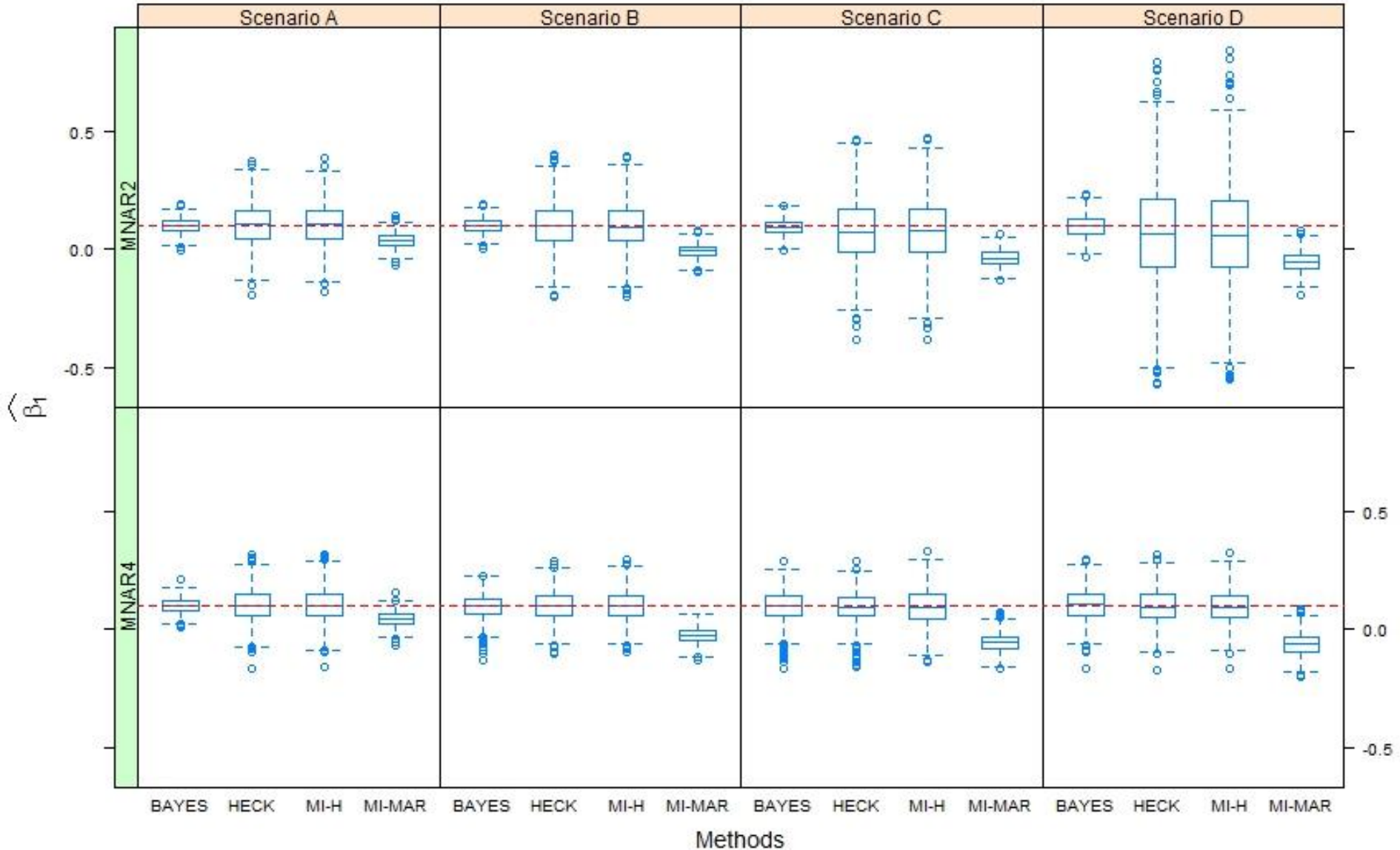- Joint Bayesian selection model

Parameter of interest
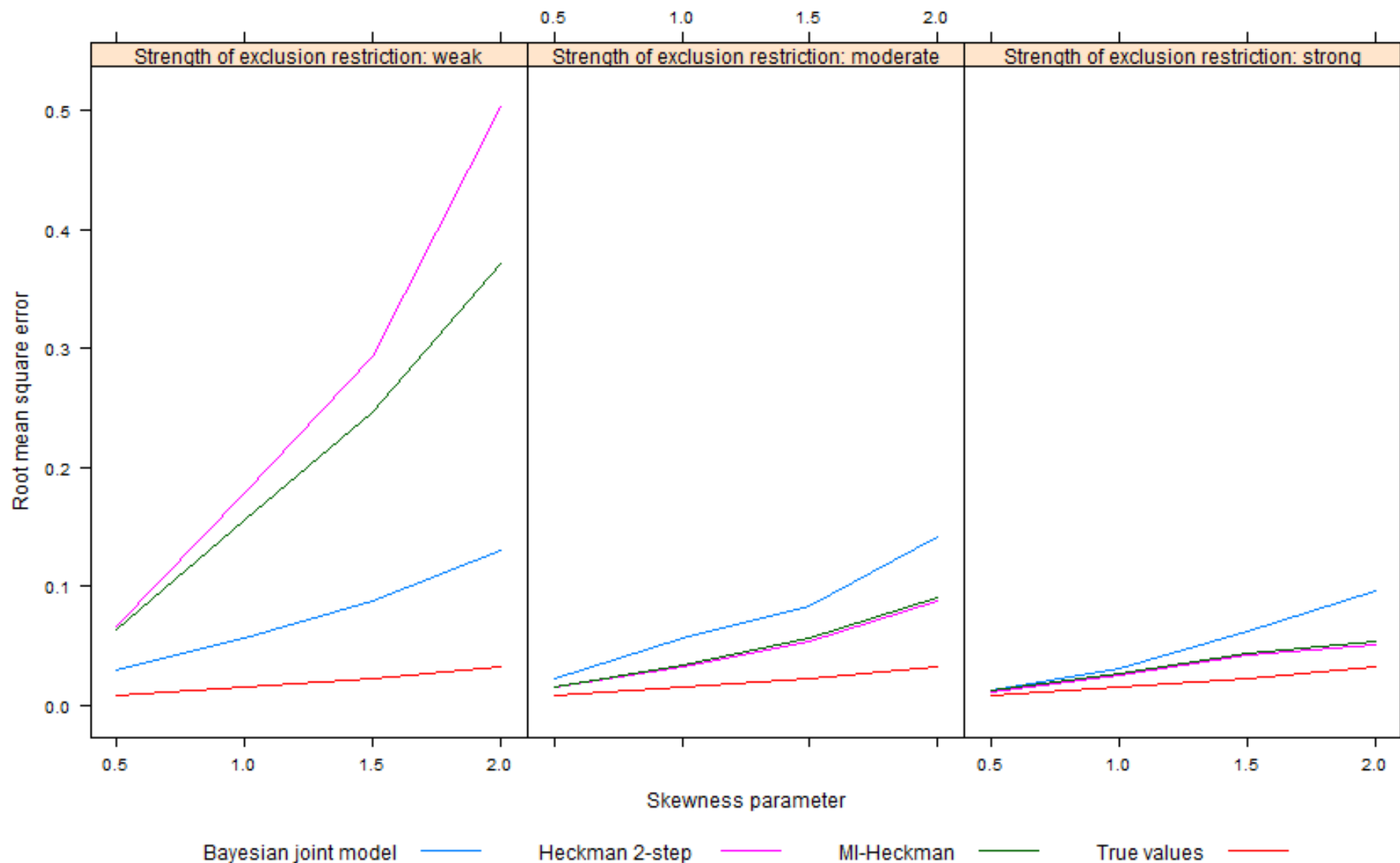
$$Y_i = \beta_0 + \boldsymbol{\beta_1} X_i + \varepsilon_i$$

Results: MNAR 1

Scenario	Method	% bias	95% CI coverage	rMSE
MNAR1a % missing=0.2 Cor(Y,R)=0.1 Cor(Z,X)=0.7	Full data	1%	0.947	0.023
	MI (MAR)	5%	0.942	0.026
	Heckman 2-step	5%	0.998	0.155
	MI_Heckman	4%	0.998	0.089
	Joint Bayesian	1%	0.945	0.026
MNAR1d % missing=0.5 Cor(Y,R)=0.4 Cor(Z,X)=0.3	Full data	1%	0.947	0.023
	MI (MAR)	16%	0.924	0.041
	Heckman 2-step	9%	0.990	0.359
	MI_Heckman	5%	0.997	0.129
	Joint Bayesian	2%	0.934	0.040

Results: Exclusion restriction in joint selection models

	Method	% bias	95% CI coverage	rMSE
Weak exclusion restriction	Full data	0%	0.943	0.016
	Heckman 2-step	2%	0.986	0.178
	MI-Heckman	1%	0.969	0.156
	Bayesian_noZ	32%	0.740	0.040
	Bayesian model	49%	0.564	0.056
Strong exclusion restriction	Full data	0%	0.943	0.016
	Heckman 2-step	4%	0.964	0.032
	MI-Heckman	5%	0.983	0.034
	Bayesian_noZ	55%	0.400	0.061
	Bayesian model	50%	0.537	0.057

- Work adds to previous evidence on Heckman's 2-step
  - Reliance on strong exclusion restrictions
  - Robust to departures from Normality (Moments estimator)

- We propose MI based on a selection model
  - Makes no additional assumptions
  - Performs similarly to the underlying selection model (Heckman's), but slightly more precise (estimates based on whole sample)

- Joint selection models perform well but
  - Rely heavily on distributional assumptions
  - Exclusion restrictions need to be included

- Longitudinal study looking at causal effect of surgery on 5-year quality-adjusted life-years on patients with reflux disease.

- 52% missing QALYs
  - Can't exclude MNAR
  - Patients in worse health may feel the surgery is not working and may be less likely to complete questionnaires/answer the phone

- MNAR setting
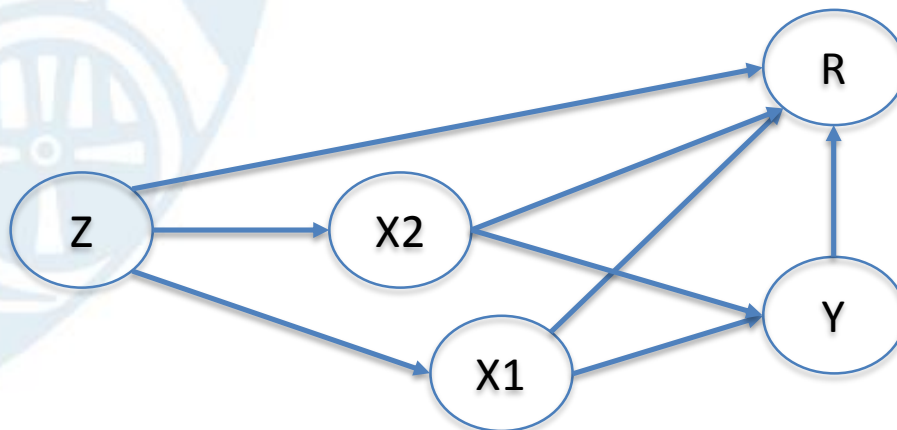
Z – general views about medicine

X1 – key prognostic factors (e.g. age)

X2 – predictors of R but not included in model (e.g. education)

Results: case study

	Complete cases (N=231)	MI (MAR) (N=453)	Heckman 2-step (N=231)	MI-Heckman (N=453)	Joint model (N=453)
Surgery	0.361 (0.101)***	0.410 (0.101)***	0.434 (0.098)***	0.442 (0.095)***	0.443 (0.095)***
Male	-0.153 (0.097)	-0.192 (0.099)	-0.212 (0.098)*	-0.211 (0.094)*	-0.197 (0.094)*
Age	-0.003 (0.004)	-0.006 (0.004)	-0.007 (0.004)	-0.006 (0.004)	-0.006 (0.004)
Baseline EQ-5D	2.066 (0.227)***	2.151 (0.223)***	2.171 (0.216)***	2.112 (0.197)***	2.174 (0.198)***
REFLUX score	-0.004 (0.003)	-0.007 (0.004)	-0.007 (0.004)	-0.007 (0.003)*	-0.007 (0.003)*
BMI	-0.026 (0.013)*	-0.010 (0.012)	-0.006 (0.012)	-0.010 (0.011)	-0.008 (0.011)
(...)					

- It is not about the right answer (!)
  - Untestable assumptions

- Where do you want to put your assumptions?
  - Valid exclusion restrictions are rare in health settings
  - 'True' distribution or model specification are unknown

- MNAR best handled via sensitivity analysis
  - should empower decision makers to frame transparent, readily interpretable assumptions
  - allow defensible inferences/recommendations

Let Y_{1i} and Y_{2i} be defined as before (omitting the individual index i)

$$Y_1 = \beta_1 X_1 + e_1, \qquad e_1 \sim N(0, \sigma_1^2)$$

$$Y_2 = \beta_2 X_2 + e_2, \qquad e_2 \sim N(0, 1), \qquad \text{cor}(e_1, e_2) = \rho$$

Suppose there is a random variable $U \sim N(0, 1 - \rho^2)$ such that

$$e_2 = \lambda e_1 + U$$

Where $\lambda = \rho \sigma_1$, and U is independent of e_1. Substituting for e_2 in Y_2

$$Y_2 = \lambda Y_1 + \gamma X^* + U,$$

Where

$$\gamma X^* = X(\beta_{2a} - \lambda \beta_{1a}) - \beta_{1b} X_1^* + \beta_{2b} X_2^*$$

for

$$X_1 = (X, X_1^*), \ X_2 = (X, X_2^*), \ \beta_1^T = (\beta_{1a}^T, \beta_{1b}^T), \ \beta_2^T = (\beta_{2a}^T, \beta_{2b}^T)$$

Define

$$\lambda^* = \frac{\lambda}{\sqrt{1-\rho^2}}, \gamma^* = \frac{\gamma}{\sqrt{1-\rho^2}}, U^* = \frac{U}{\sqrt{1-\rho^2}}$$

Then $U^* \sim N(0, 1)$ and

$$P(Y_2 > 0 | X_1, X_2, Y_1) = P(\lambda Y_1 + \gamma X^* \geq -U)$$
$$= P(\lambda^* Y_1 + \gamma^* X^* \geq -U^*)$$
$$= \Phi(\lambda^* Y_1 + \gamma^* X^*)$$

A standard probit selection model.