MRC | Biostatistics Unit

Double Robust Methods for Missing Data

Shaun Seaman
MRC Biostatistics Unit, Cambridge University
and Stijn Vansteelandt, Ghent University

Incomplete data

Missing data are common in epidemiology and social science.

In surveys, participants do not answer all survey questions, e.g. those about sexual behaviour or income.

In cohort studies and clinical trials, which may last for months/years, some individuals may drop out during study — lose interest in study, too unwell to attend, move abroad, . . .

Consequences:

- estimates are less accurate than if we had all the data ($\Rightarrow$ efficiency loss)
- (should) worry that the individuals on whom we have collected data may not be representative of population even if original sample was ($\Rightarrow$ bias)

Inverse probability weighting (IPW) and imputation

- **Complete case analysis**: Use only individuals with complete data ('complete cases').
  E.g. Estimate mean height in population. Sample 50 males and 50 females for study. $1/2$ males and $1/3$ of females participate. Mean height in sample overestimates population height.
- **IPW**: Weight complete cases to represent the whole sample.
  E.g. Assign weights $\frac{1}{1/2} = 2$ to participating males and $\frac{1}{1/3} = 3$ to females. Each participating male represents himself and one non-participating male.
- **Imputation**: Replace missing values by plausible values.
  Various flavours: mean imputation, regression imputation, multiple imputation, maximum likelihood, Bayesian modelling.

Estimating a population mean

Want to estimate $\beta = E(Y)$ from sample of n individuals.

Y is missing for some individuals.

$R = 1$ if Y observed; $R = 0$ if Y missing.

Observe variables X on everyone in sample.

Complete case estimator:

$$\frac{\sum_{1=1}^{n} R_i Y_i}{\sum_{1=1}^{n} R_i}$$

Unbiased if $R \perp\!\!\!\perp Y$. Does not use X.

Weaker assumption is $R \perp\!\!\!\perp Y \mid X$, i.e. 'missing at random' (MAR).
Equivalently, $P(R = 1 \mid X, Y) = P(R = 1 \mid X)$.

IPW and regression imputation

Inverse probability weighting (IPW)

Specify model $\pi(X;\alpha)$ for $P(R=1 \mid X)$.
Estimate α by fitting model to all individuals.

$$\hat{\beta}_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\alpha})}$$

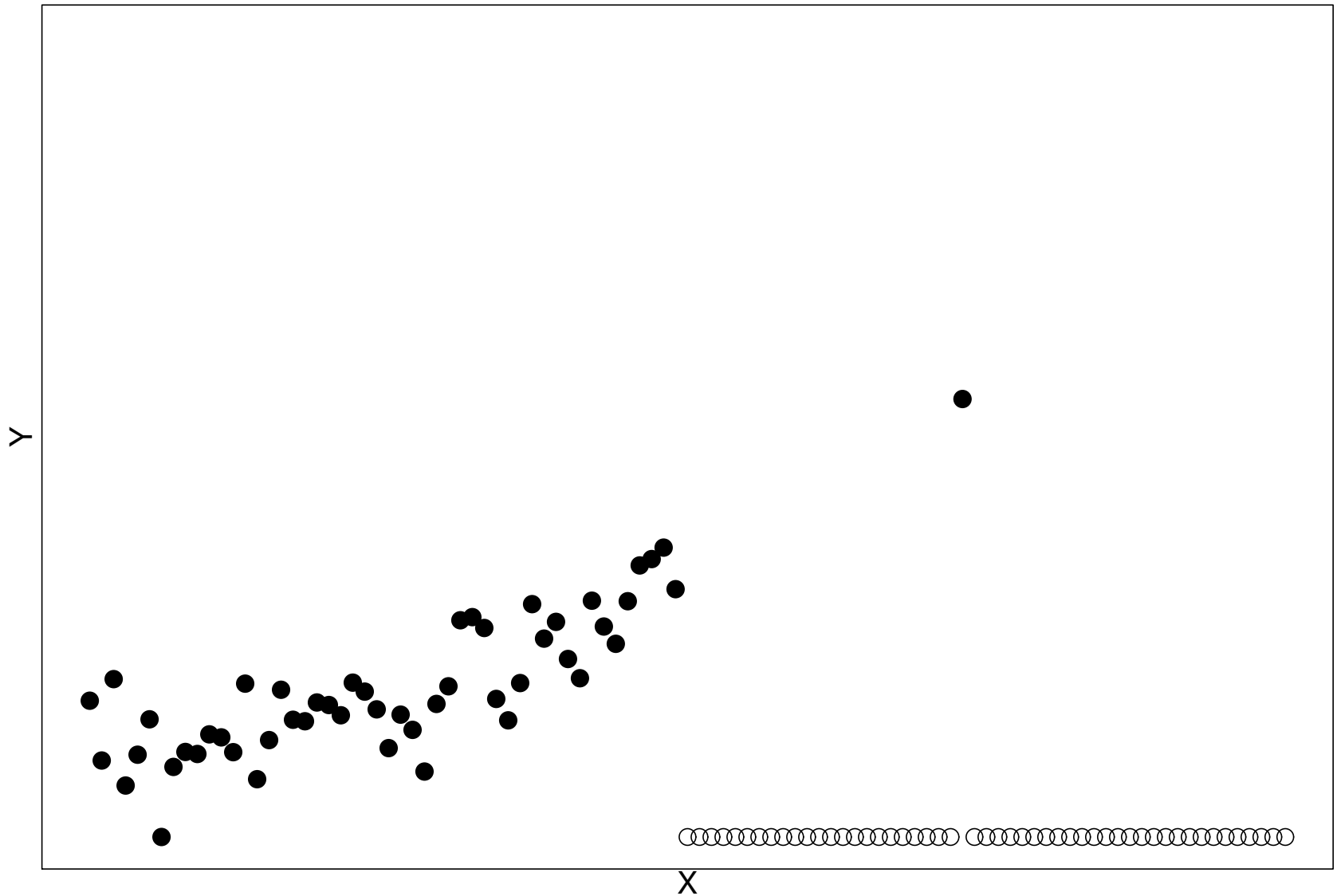
Regression imputation (RI)

Specify model $m(X;\gamma)$ for $E(Y \mid X)$.
Estimate γ by fitting model to individuals with $R=1$.

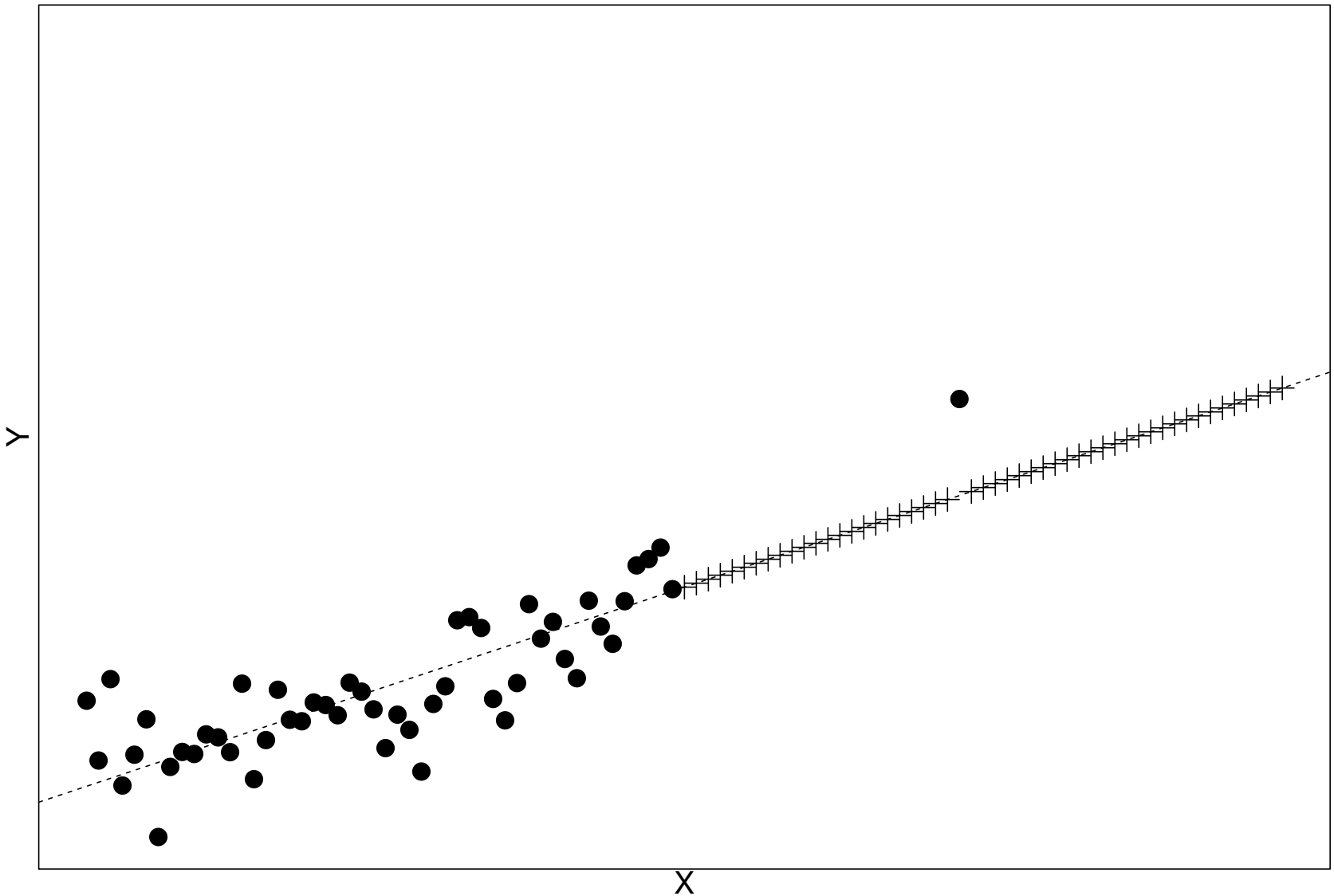
$$\hat{\beta}_{\mathrm{RI}} = \frac{1}{n}\sum_{i=1}^{n} m(X_i;\hat{\gamma}) \quad = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\{R_i Y_i + (1-R_i)m(X_i;\hat{\gamma})\}$$

RI is more efficient than IPW, especially when X is very predictive of Y or R.
But when X is very predictive of R, risk of extrapolation.

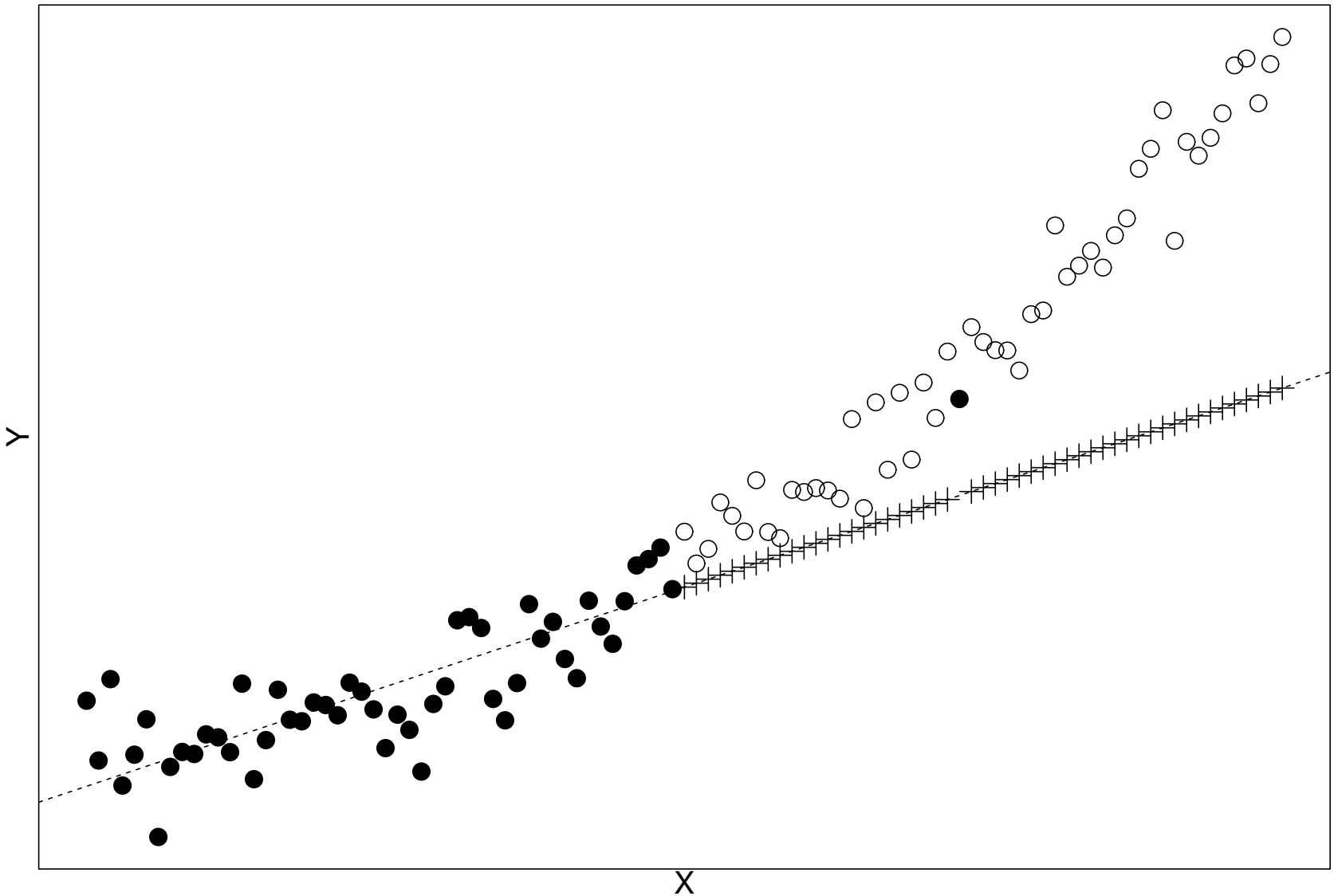
Extrapolation



Extrapolation



Extrapolation



Extrapolation

Extrapolation occurs when little overlap between the regions of X-space where complete and incomplete cases lie.

Easy to spot here, but might not be if X were higher-dimensional.

Could use more flexible imputation model $m(X;\gamma)$, but

- this would greatly increase SE of $\hat{\beta}_{\mathrm{RI}}$
- can be hard to ensure that a flexible model is flexible enough outside the region of X-space where complete-cases lie.

Fit of $\pi(X;\alpha)$ is (arguably) easier to check, and estimated SE of $\hat{\beta}_{\mathrm{IPW}}$ (a weighted mean) reflects true uncertainty.

Also might be happier modelling $P(R = 1 \mid X)$ than $E(Y \mid X)$.

But $\hat{\beta}_{\mathrm{IPW}}$ is inefficient.

Augmented IPW estimator

IPW can be made more efficient by augmentation.

$$\begin{aligned}\hat{\beta}_{\mathrm{DR}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i;\hat{\alpha})} + \frac{1}{n}\sum_{i=1}^{n}\left\{1-\frac{R_i}{\pi(X_i;\hat{\alpha})}\right\} m(X_i;\hat{\gamma}) \\ &= \frac{1}{n}\sum_{i=1}^{n} m(X_i;\hat{\gamma}) - \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\alpha})}\{m(X;\hat{\gamma}) - Y_i\}\end{aligned}$$

Less efficient than $\hat{\beta}_{\mathrm{RI}}$, but not much less if X very predictive of Y.

$\hat{\beta}_{\mathrm{DR}}$ is double robust: consistent if $\pi(X;\alpha)$ or $m(X;\gamma)$ is correct.

If $\pi(X;\alpha)$ correct,

$$E\left[\left\{1-\frac{R}{\pi(X;\alpha)}\right\} m(X;\gamma) \mid X\right] = \left\{1-\frac{E(R \mid X)}{\pi(X;\alpha)}\right\} m(X;\gamma) = 0$$

If $m(X;\gamma)$ correctly specified,

$$E\left[\frac{R}{\pi(X;\alpha)}\{m(X;\gamma) - Y\} \mid X, R\right] = \frac{R}{\pi(X;\alpha)}\{m(X;\gamma) - E(Y \mid X, R)\} = 0$$

Double robust estimators

Augmented IPW estimator is **one** example of DR estimator of β.

There are also DR estimators of parameters in models. E.g.

- $g\{E(Y \mid X)\} = \beta^\top X$ (generalised linear model with link function g), with Y and/or part of vector X partially observed.
- $g\{E(Y_t \mid X_t)\} = \beta^\top X_t$, where $Y = (Y_1, \ldots, Y_T)$ are repeated outcomes and X_t is vector of time-dependent covariates (generalised estimating equations), with Y_t's partially observed due to dropout.
- Hazard ratios in Cox proportional hazards model, with a partially observed covariate.
- Area under received operating curve, with partially observed outcome or predictor.

Problems with 'standard' DR estimator

Return to estimating $\beta = E(Y)$.

Can use any consistent estimators of (α, γ) in $\pi(X; \alpha)$ and $m(X; \gamma)$

When $\pi(X; \alpha)$ and $m(X; \gamma)$ both correctly specified, efficiency of $\hat{\beta}_{\mathrm{DR}}$ does not depend on the choice.

Early work used maximum likelihood estimators ('standard' DR estimator).

It was noticed later that standard DR estimator *can* be:

- very inefficient (even more than IPW) when imputation model misspecified;
- very biased when both models misspecified.

Also, $\hat{\beta}_{\mathrm{DR}}$ may lie outside observed range of Y.
May even lie outside parameter space of β.

Improved double robust estimators

Many proposals have been made. Here is one example:

Targeted maximum likelihood

Estimate (α, γ) so

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(X_i;\hat{\alpha})}\{m(X;\hat{\gamma}) - Y_i\} = 0$$

Advantages:

- $\hat{\beta}_{\mathrm{DR}}$ is now a RI estimator, so $\hat{\beta}_{\mathrm{DR}}$ within parameter space
- allows very flexible models $m(X;\gamma)$
- less sensitive to misspecification in simulation studies

Steps:

1. Fit flexible preliminary model $m^{(0)}(X;\gamma^{(0)})$ and model $\pi(X;\alpha)$
2. Fit model $m(X;\gamma)$ with offset $m^{(0)}(X;\hat{\gamma}^{(0)})$ and covariate $\frac{R}{\pi(X;\hat{\alpha})}$

Data-adaptive Methods

$\hat{\beta}_{\mathrm{DR}}$ is less efficient than $\hat{\beta}_{\mathrm{RI}}$ when $m(X;\gamma)$ correctly specified.

Why not just use RI with flexible, data-adaptive estimator for $E(Y \mid X)$ and forget about DR estimator?

Flexible model means less concern about misspecification.

- kernel smoothing
- penalised likelihood (e.g. Lasso)
- ensemble learners (e.g. Superlearner)

DR estimators are very useful when data-adaptive methods used.

Consider $\hat{\beta}_{\mathrm{RI}} = \frac{1}{n}\sum_{i=1}^{n} m(X_i;\hat{\gamma})$ with data-adaptive estimator $\hat{\gamma}$.

$\hat{\gamma}$ typically has complicated finite-sample distribution and non-uniform convergence of this to a normal distribution.

$\hat{\beta}_{\mathrm{RI}}$ inherits these properties of $\hat{\gamma}$, so that uniformly valid confidence intervals with nominal coverage are hard to obtain.

Inference after model selection

Consider simple case of $X = (X_1, X_2)$ and

$$m(X;\gamma) = \begin{cases} \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 & \text{if } \gamma_2 \text{ significant at 5\% level} \\ \gamma_0 + \gamma_1 X_1 & \text{if } \gamma_2 \text{ not significant} \end{cases}$$

For fixed γ: For large enough n the probability of selecting X_2 is ≈ 0 or ≈ 1, and so

$$\sqrt{n}(\hat{\beta}_{\mathrm{RI}} - \beta) \sim \text{Normal}(0, \sigma^2) \qquad \text{approximately.}$$

Confidence intervals based on normal approximation have good coverage when n is large enough.

But how big n needs to be **depends on** γ.
No matter how large n is, can find γ such that normal approximation is poor.

Example

$R = 1$ for half the sample ($n' = \frac{n}{2}$) and $R = 0$ for other half.

For individuals with $R = 1$: $X_1 \sim \text{Normal}(0, 1)$

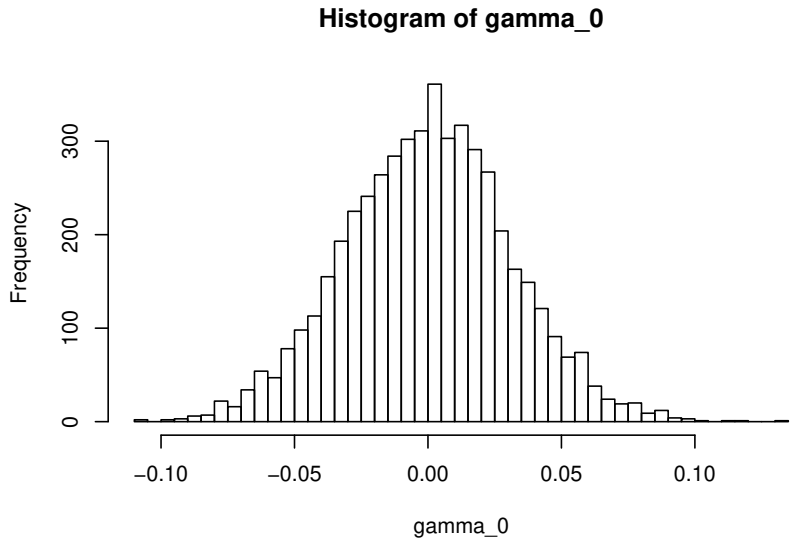
For individuals with $R = 0$: $X_1 \sim \text{Normal}(2, 1)$

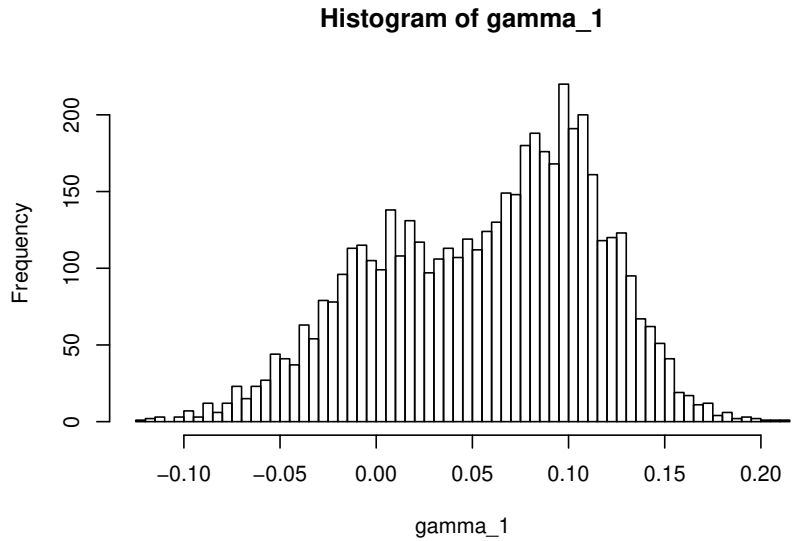
For all individuals:

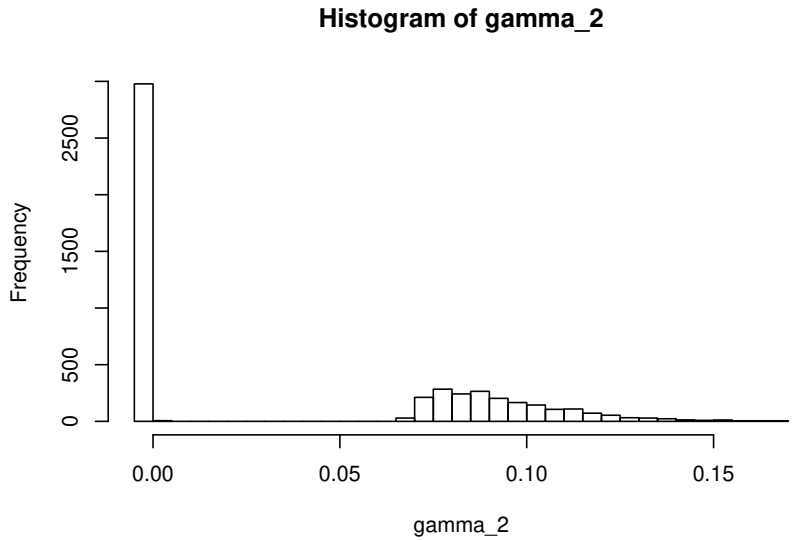
$$\begin{aligned} X_2 \mid X_1 &\sim \text{Normal}(X_1, 1) \\ Y &\sim \text{Normal}(\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2,\ 1) \\ \text{with } (\gamma_0, \gamma_1, \gamma_2) &= \left(0, \frac{1}{\sqrt{n'}}, \frac{2}{\sqrt{n'}}\right) \end{aligned}$$

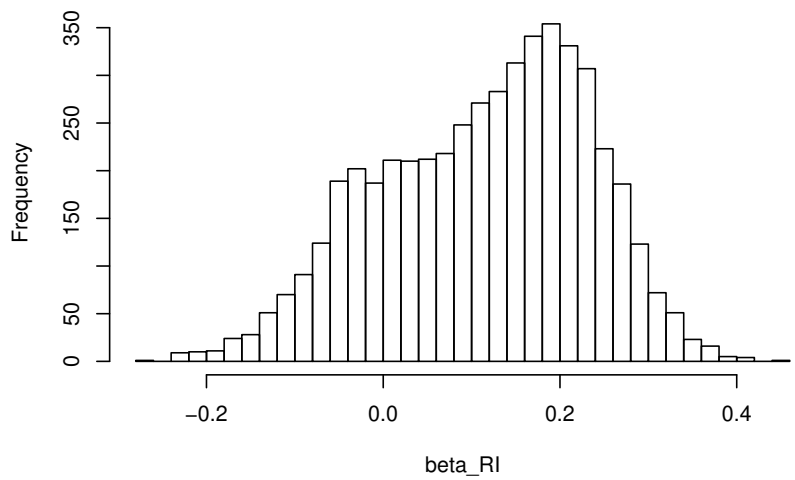
Note: Data are MAR (i.e. $R \perp\!\!\!\perp Y \mid X$).

Repeated sampling distributions when $n = 2000$

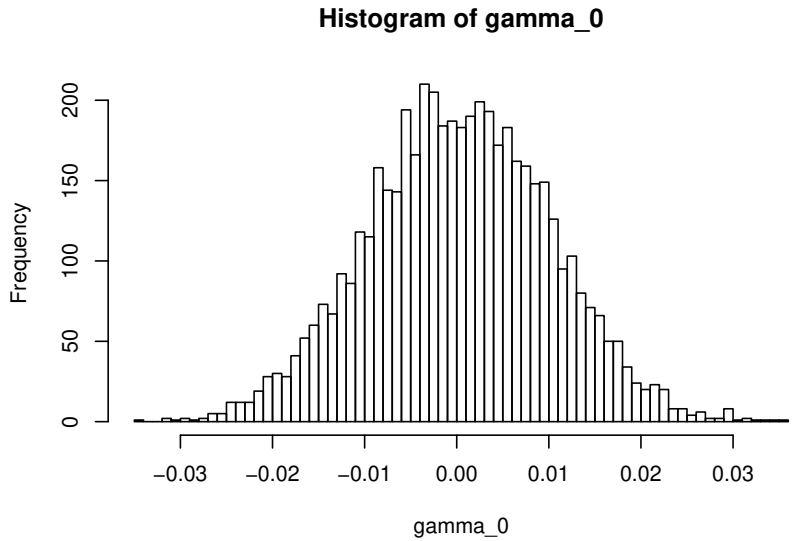


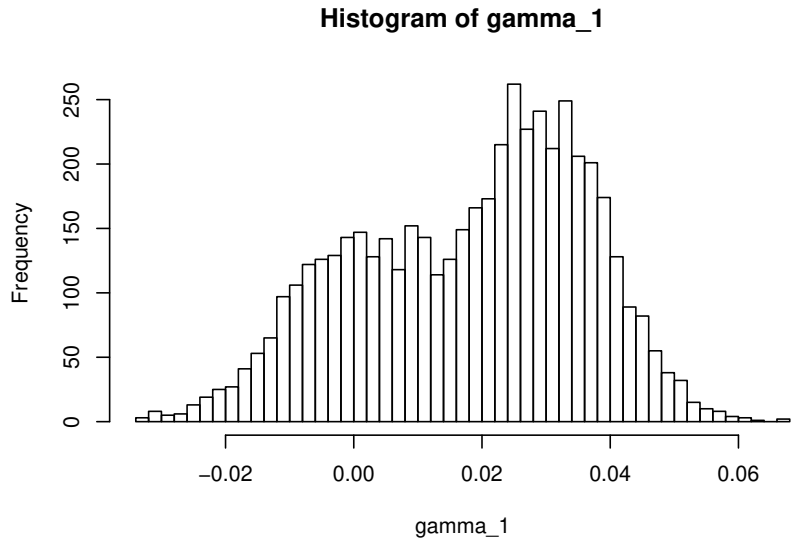


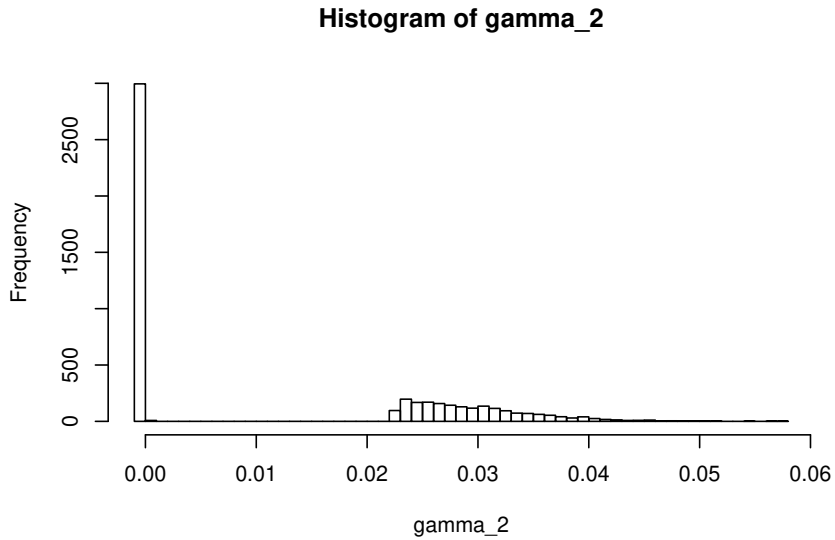


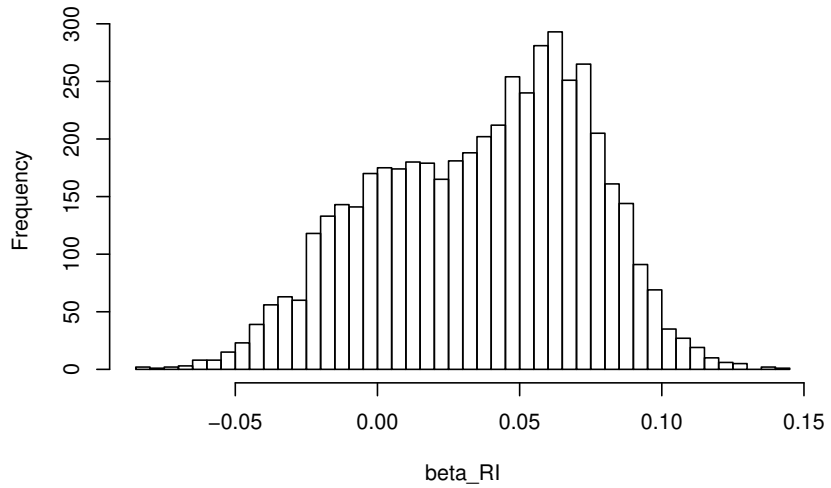


Repeated sampling distributions when $n = 20000$









Non-uniform convergence of $\hat{\gamma}$ causes non-uniform convergence of $\hat{\beta}_{\mathrm{RI}}$

$\hat{\gamma}$ does not converge **uniformly**.

By Taylor series expansion,

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_{\mathrm{RI}} - \beta) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(X_i; \hat{\gamma}) - \beta\} \\
&\quad = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(X_i; \gamma) - \beta\} + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial m}{\partial \gamma}(X_i; \gamma)\Big|_{\gamma=\gamma} \sqrt{n}(\hat{\gamma} - \gamma) \\
&\qquad + \text{higher-order terms}
\end{aligned}$$

$\hat{\beta}_{\mathrm{DR}}$ is 'immune' to non-uniform convergence of $\hat{\gamma}$

DR estimators are of the form $\hat{\beta}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n} q(X_i; \hat{\theta})$, where $\theta = (\alpha, \gamma)$.

$$\text{E.g.} \quad q(X_i; \hat{\theta}) = \frac{R_i Y_i}{\pi(X_i; \hat{\alpha})} + \left\{ 1 - \frac{R_i}{\pi(X_i; \hat{\alpha})} \right\} m(X_i; \hat{\gamma}).$$

$q(X; \theta)$ satisfies 'orthogonality' or 'immunisation' condition:

$$E\left\{ \frac{\partial q}{\partial \theta}(X; \theta)\Big|_{\theta=\theta} \right\} = 0.$$

So, if we use DR estimator, we can use data-adaptive method for $m(X; \gamma)$ (and for $\pi(X; \alpha)$) and still get uniformly valid inference.

Closing remarks

- DR (and IPW) estimators need model for missingness. Relatively simple when one variable is partially observed or missingness is monotone. If not, plausible models difficult to find.
- Most work assumes data are MAR.
  Example of MNAR method:
  logit $P(R = 1 \mid X, Y) = \omega Y + a(X)$, which implies
  $f(Y \mid X, R = 0) \propto f(Y \mid X, R = 1) \exp(-\omega Y)$.
  DR estimator consistent if model for $a(x)$ or model for $f(Y \mid X, R = 1)$ is correctly specified.
- Much work on DR methods is in the literature on causal inference. Causal inference problems are often missing data problems.
  E.g. average effect (in population) of a treatment is $E(Y^{(1)} - Y^{(0)})$, where $Y^{(1)}$ and $Y^{(0)}$ are outcomes when treated and untreated respectively. Only observe one of $Y^{(1)}$ and $Y^{(0)}$.

Seaman and Vansteelandt (2018)

This presentation is based on:
Seaman and Vansteelandt (2018). 'Introduction to Double Robust Methods for Incomplete Data', Statistical Science.

This article also covers:

- semi-parametric theory underlying DR estimators
- DR estimators for more general missingness scenarios
- estimating SE of $\hat{\beta}_{\mathrm{DR}}$
- other improved DR estimators (including empirical likelihood)
- more about DR when data-adaptive methods used (including when number of covariates increases with *n*)
- connections between DR and 'design-consistent' sample survey estimators
- selection strategies for models $\pi(X;\alpha)$ and $m(X;\gamma)$
- review of software