Imputation when data cannot be pooled

Nicola Orsini

Department of Global Public Health
Karolinska Institutet

2024 UK Stata Conference

September 12-13, 2024

Outline

- Context and problem
- Key idea underlying the method
- `mi impute from`
- Example of 100% missing confounder
- Simulation results
- Final remarks

What is the context?

- Collaborative efforts such as pooling or consortia projects are commonly undertaken to address complex research questions, enhance precision, and improve the generalizability of findings
- Individual data is often not pooled but harmonized and analyzed at individual sites (i.e., distributed data networks) due to regulatory constraints and the need for timely results
- Systematic (100%) missing data is likely to occur
- `mi impute` cannot be used without any observed data

What is the idea?

- The variable systematically missing in one study site can be of any type (quantitative, qualitative) and any shape
- One or more study sites within the network have data to estimate an imputation model
- Files containing the estimated regression coefficients and their associated precision from the imputation model are shared across the network
- Imputations are generated by inverting the predicted conditional cumulative probabilities

How is it implemented in Stata?

It is a user-written imputation method involving two commands:

- `mi_impute_from_get` receives list of files (.txt, .xlsx) containing estimated regression coefficients and returns formatted matrices. If multiple files are specified, it combines regression coefficients using an inverse-variance weighted least squares model.

- `mi impute from` receives the formatted regression coefficients, takes a random draw from their posterior, and generates multiple imputations

Both commands require the specification of the imputation model using the option `imodel()`.

What type of imputation models?

- `qreg` for modelling conditional quantiles of a quantitative variable.

  If p predictors, then $99(p+1)$ regression coefficients

- `mlogit` for modelling conditional probabilities of a categorical variable.

  If p predictors and k levels, then $k(p+1)$ regression coefficients

- `logit` is used for modelling the conditional probability of a binary variable.

  If p predictors, then $p+1$ regression coefficients

How does it work conditional quantile imputation?

Consider a continuous variable z_i completely missing in Study 1.

- In another study site, saying Study 2, estimate p-quantile regression model for the continuous variable z_i conditionally on predictors $\mathbf{w}_i$

$$Q_{z_i|\mathbf{w}_i}(p) = \mathbf{w}_i\gamma(p) \quad p \in \{0.01, 0.02, \ldots, 0.99\}$$

- Back to Study 1, draw a random value U_i from a random continuous uniform distribution $\mathcal{U}(0,1)$ for the i-th individual
- Extract the floor $f = \lfloor U_i\% \rfloor$ and modulus mod $= U_i\% - \lfloor U\% \rfloor$
- The m-th imputation $z_i^{(m)}$ for the i-th individual is the weighted average of the f and $f+1$ conditional predicted quantiles

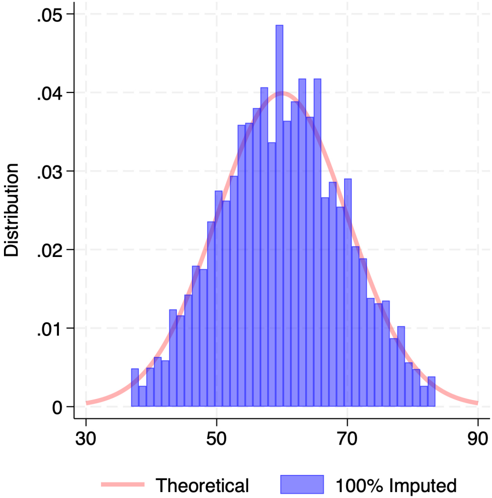
$$z_i^{(m)} = (1 - \text{mod})\hat{Q}_{z_i|\mathbf{w}_i}(f) + \text{mod}\,\hat{Q}_{z_i|\mathbf{w}_i}(f+1)$$

Thiesmeier R, Bottai M, Orsini N. (2024). Systematically missing data in distributed data networks: multiple imputation when data cannot be pooled. *Journal of Statistical Computation and Simulation*. In Press.

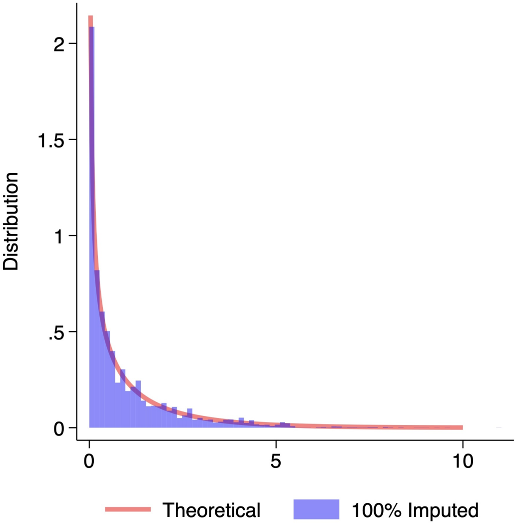
Mata function for conditional quantile imputation

```
mata:
real colvector mi_impute_cmd_from_xb_qreg(real matrix X, real rowvector b)
{
            p = cols(X)
            u = runiform(rows(X),1, 0.01, .99)*100
            pvec = J(1, 99, NULL)
            for (i=1; i<=99; i++) pvec[i] = &(b[1, (i*p)-(p-1)::(i*p)])
            yi = J(rows(X), 1, .)
            for (i=1; i<=rows(X); i++) {
                f = floor(u[i])
                mod = mod(u[i],1)
                yi[i] = (1-mod)*(X[i,]* *pvec[f]')+mod*(X[i,]* *pvec[f+1]')
            }
return(yi)
}
```

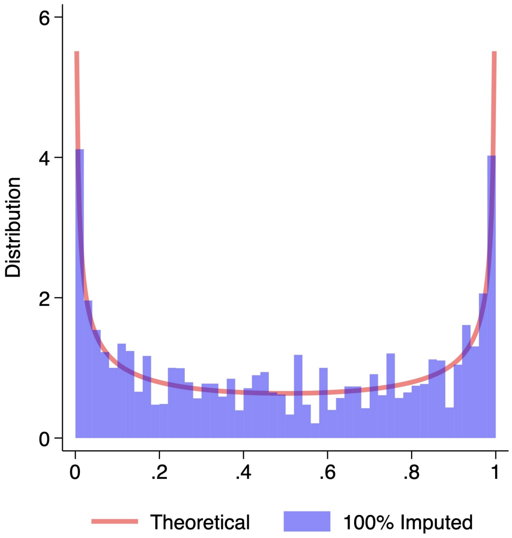
Impute a Normal distribution



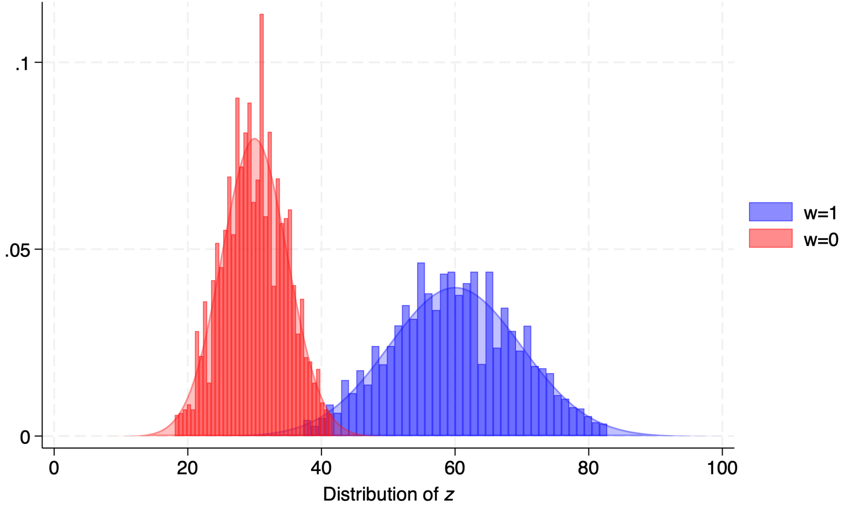
Impute a χ^2 distribution with 1 degree of freedom



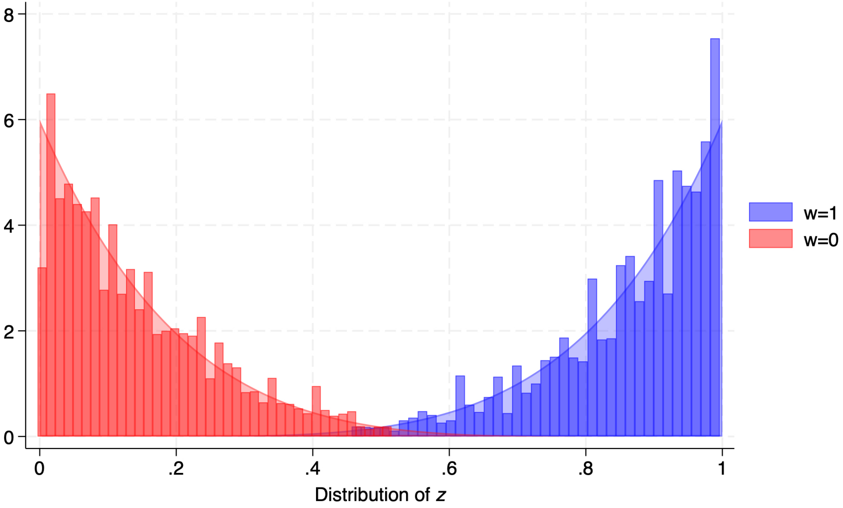
Impute a Beta distribution



Impute a Normal distribution conditionally on a binary predictor



Impute a Beta distribution conditionally on a binary predictor



Syntax

```
use study_1, clear

mi set wide

mi register imputed z

mi_impute_from_get , b(e_b) v(e_v) imodel(qreg) ///
colnames(w _cons)

mat i_b = r(get_ib)
mat i_v = r(get_iV)

mi impute from z , add(1) b(i_b) v(i_v) imodel(qreg)
```

100% missing confounder in one study

	Study 1 N=3,766	**Study 2** N=2,382	**Study 3** N=4,182	**Study 4** N=2,260	**Study 5** N=1,401
Exposure (%)	15	17	23	26	30
Outcome (%)	25	33	33	29	28
Crude OR	2.0(1.5-2.5)	1.8(1.5-2.1)	1.8(1.5-2.2)	1.6(1.2-2.0)	1.5(1.3-1.7)
C Adjusted	1.6(1.2-2.0)	1.4(1.2-1.7)	1.5(1.2-1.9)	1.2(0.9-1.6)	1.2(1.1-1.4)
Z & C Adjusted	**NA**	1.2(1.0-1.4)	1.3(1.1-1.6)	1.1(0.8-1.5)	1.1(0.9-1.2)

Mechanisms underlying confounding effects

- Common causes of exposure and outcome

  $C \sim \text{Bern}(0.4)$
  $Z \sim \chi^2(1)$

- Exposure

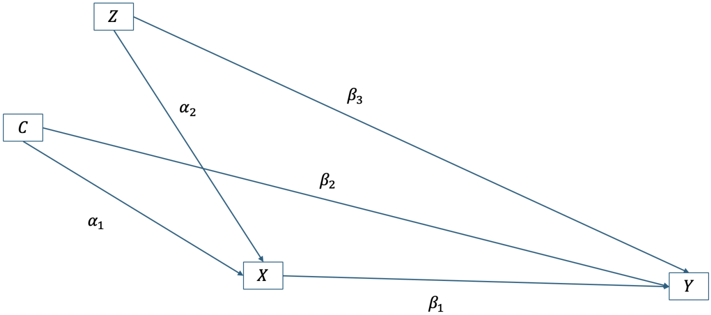
  $X \sim \text{Bern}(\text{invlogit}(\alpha_0 + \alpha_1 C + \alpha_2 Z))$

- Outcome

  $Y \sim \text{Bern}(\text{invlogit}(\beta_0 + \beta_1 X + \beta_2 C + \beta_3 Z))$

Target of statistical inference is β_1 representing the C and Z adjusted conditional effect of the treatment X on the outcome Y.

Type of confounding



Confounders C and Z are strongly increasing the probability of being exposed ($\alpha_1 > 0$, $\alpha_2 > 0$) as well as the outcome probability ($\beta_2 > 0$, $\beta_3 > 0$).

The conditional effect of the exposure is a small increment in the outcome probability $\beta_1 = \ln(1.2) = 0.18$

One way to test `mi impute from`

- Generate Study 1 from the confounding mechanism, estimate $\hat{\beta}_1$, and then set confounder Z to missing

- Generate Study 2 from the same confounding mechanism, estimate the conditional quantile imputation model

- Open Study 1, generate 10 multiple imputations using `mi impute from`, estimate $\bar{\beta}_1$ using `mi estimate`

  If `mi impute from` works well, we can expect that the sampling distribution of $\hat{\beta}_1$ based on fully observed data and the sampling distribution of $\bar{\beta}_1$ based on fully externally multiple imputed data should be bell-shaped and centered about the parameter β_1 set in the simulation.

Scenario 1: External imputation from identical confounding mechanism

Study 1 and Study 2 with sample size $n = 1,000$ come from the following mechanism

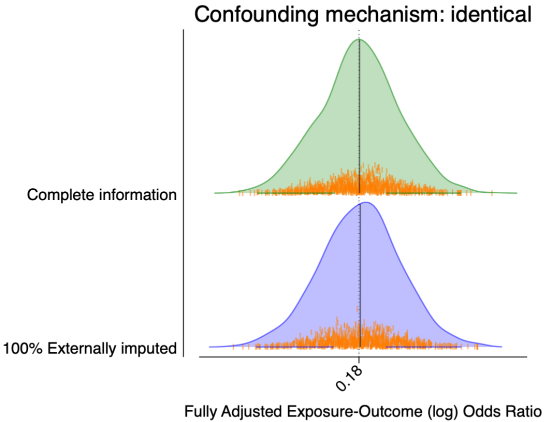
- Exposure

  $X \sim \text{Bern}(\text{invlogit}(\text{logit}(0.10) + \log(3)C + \log(1.3)Z))$

- Outcome

  $Y \sim \text{Bern}(\text{invlogit}(\text{logit}(0.20) + \log(1.2)X + \log(3)C + \log(1.3)Z))$

Comparison of simulated sampling distributions



The conditional effect of the exposure in Study 1 (under full data or 100% externally imputed) is centered about the parameter value $\beta_1 = 0.18$.

Scenario 2: External imputation from weaker confounding mechanism

Study 1 as before but Study 2 come from a weaker confounding mechanism.

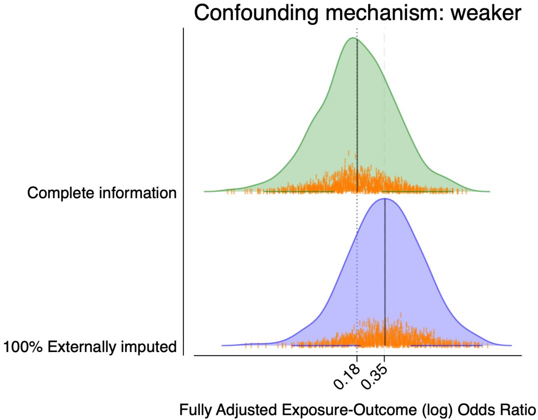
- Exposure

  $X \sim \text{Bern}(\text{invlogit}(\text{logit}(0.10) + \log(3)C + \log(1.1)Z))$

- Outcome

  $Y \sim \text{Bern}(\text{invlogit}(\text{logit}(0.20) + \log(1.2)X + \log(3)C + \log(1.1)Z))$

Comparison of simulated sampling distributions



The conditional effect of the exposure in Study 1 under 100% externally imputation is, on average, twice as much the parameter value $\beta_1 = 0.18$.

Scenario 3: External imputation from stronger confounding mechanism

Study 1 as before but Study 2 come from a stronger confounding mechanism.

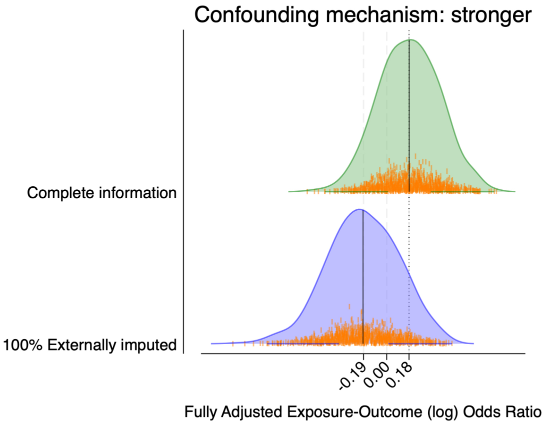
- Exposure

  $X \sim \text{Bern}(\text{invlogit}(\text{logit}(0.10) + \log(3)C + \log(1.6)Z))$

- Outcome

  $Y \sim \text{Bern}(\text{invlogit}(\text{logit}(0.20) + \log(1.2)X + \log(3)C + \log(1.6)Z))$

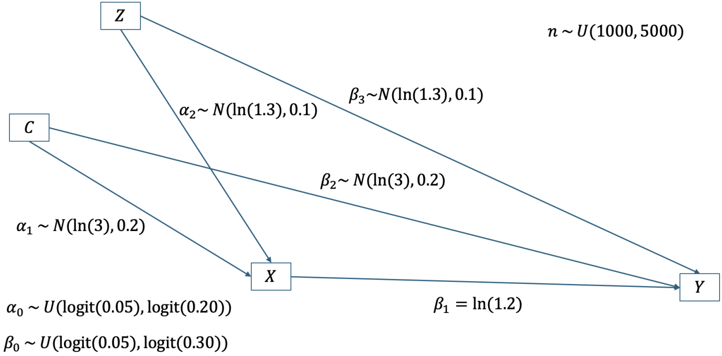
Comparison of simulated sampling distributions



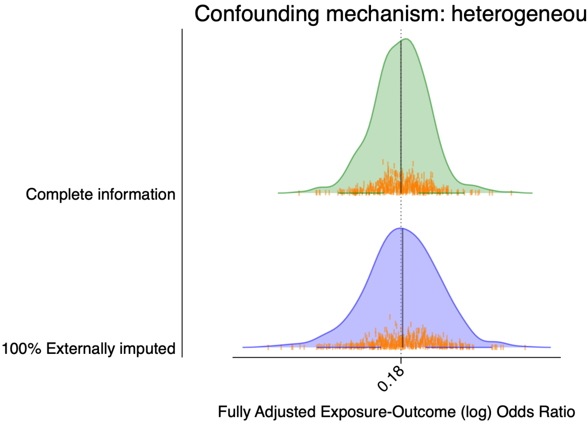
The conditional effect of the exposure in Study 1 under 100% externally imputation is, on average, at the opposite side of the parameter value $\beta_1 = 0.18$.

Scenario 4: External imputation from a heterogeneous mechanism

Study 1 as before but Study 2 come from a heterogeneous confounding mechanism.



Comparison of simulated sampling distributions



The conditional effect of the exposure in Study 1 under 100% externally imputation is, on average, centered about the parameter value $\beta_1 = 0.18$.

Back to our motivating example

```
use qreg_study_1_miss, clear
mi set wide
mi register imputed z

mi_impute_from_get , ///
b(e_b_s2 e_b_s3 e_b_s4 e_b_s5) ///
v(e_v_s2 e_v_s3 e_v_s4 e_v_s5) ///
    colnames(y x c _cons) imodel(qreg)

mat ib = r(get_ib)
mat iV = r(get_iV)

mi impute from z , add(10) b(ib) v(iV) imodel(qreg) ///
rseed(240912)

mi estimate, post eform: logistic y x c z
```

Table with the imputed estimate

	Study 1 N=3,766	**Study 2** N=2,382	**Study 3** N=4,182	**Study 4** N=2,260	**Study 5** N=1,401
Exposure (%)	15	17	23	26	30
Outcome (%)	25	33	33	29	28
Crude OR	2.0(1.5-2.5)	1.8(1.5-2.1)	1.8(1.5-2.2)	1.6(1.2-2.0)	1.5(1.3-1.7)
C Adjusted	1.6(1.2-2.0)	1.4(1.2-1.7)	1.5(1.2-1.9)	1.2(0.9-1.6)	1.2(1.1-1.4)
Z & C Adjusted	**1.3(1.0-1.7)**	1.2(1.0-1.4)	1.3(1.1-1.6)	1.1(0.8-1.5)	1.1(0.9-1.2)

$\hat{\beta}_1 = 1.333971$ based on complete data

$\bar{\beta}_1 = 1.304004$ based on external imputations from 4 heterogeneous studies using `mi impute from`

Next step in a collaborative effort would be the specification of a meta-analytical model to learn from multiple studies.

Final remarks

- `mi impute from` is based on the principle of inverting predicted conditional cumulative probabilities
- `mi impute from` can be used with both sporadic and systematic missing data
- `mi impute from` cannot be called by `mi impute chained`
- `mi impute from` using regression coefficients from imputation model estimated in data where completely different mechanisms are operating is likely to lead to the wrong inferential results
- This is an on-going joint work with Robert Thiesmeier & Matteo Bottai at Karolinska Institutet

References

- Thiesmeier R, Bottai M, Orsini N. (2024). Systematically missing data in distributed data networks: multiple imputation when data cannot be pooled. *Journal of Statistical Computation and Simulation*. In Press.

- Thiesmeier R, Bottai M, Orsini N. (2024). Imputation when data cannot be pooled. *Stata Journal*. On-going.