Handling Missing Data in Research: A Practical Guide

Multiple Imputation

Waylon Howard

Webinar, November 19, 2024

Multiple Imputation

A three-step process

1. Imputation Phase[1]
2. Analysis Phase
3. Pooling Phase

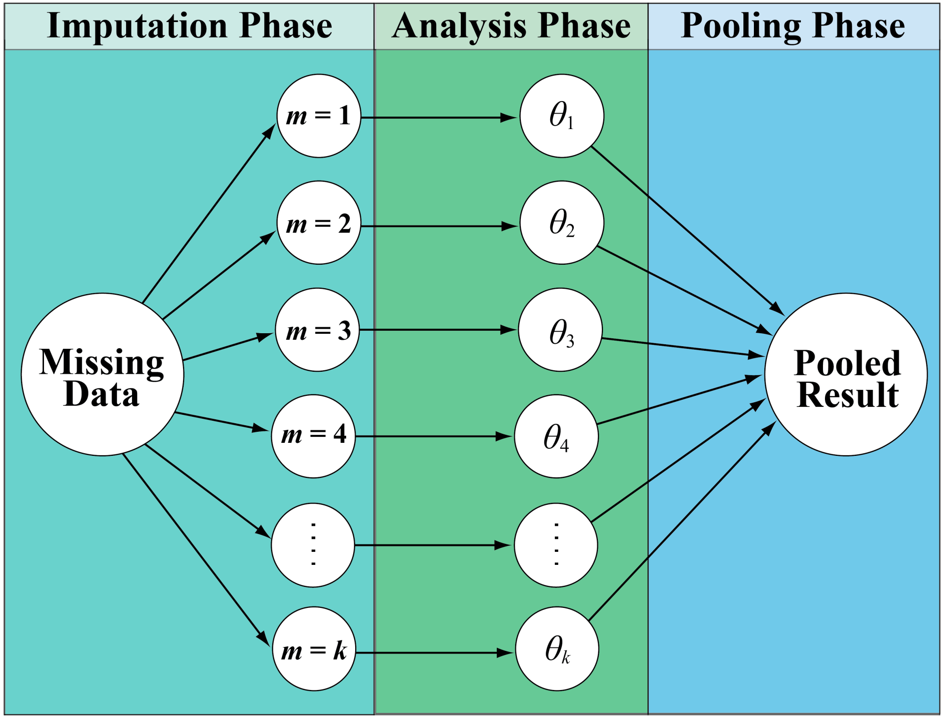
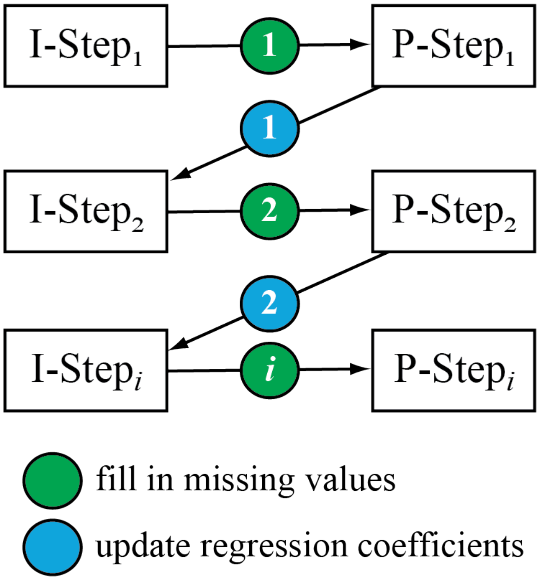


Illustration adapted from Enders, C. K. (2022).

[1] A wide range of algorithms exists to tackle various types of data.

Imputation Phase

algorithm concept



I-Step$_1$ → 1 → P-Step$_1$

P-Step$_1$ → 1 → I-Step$_2$

I-Step$_2$ → 2 → P-Step$_2$

P-Step$_2$ → 2 → I-Step$_i$

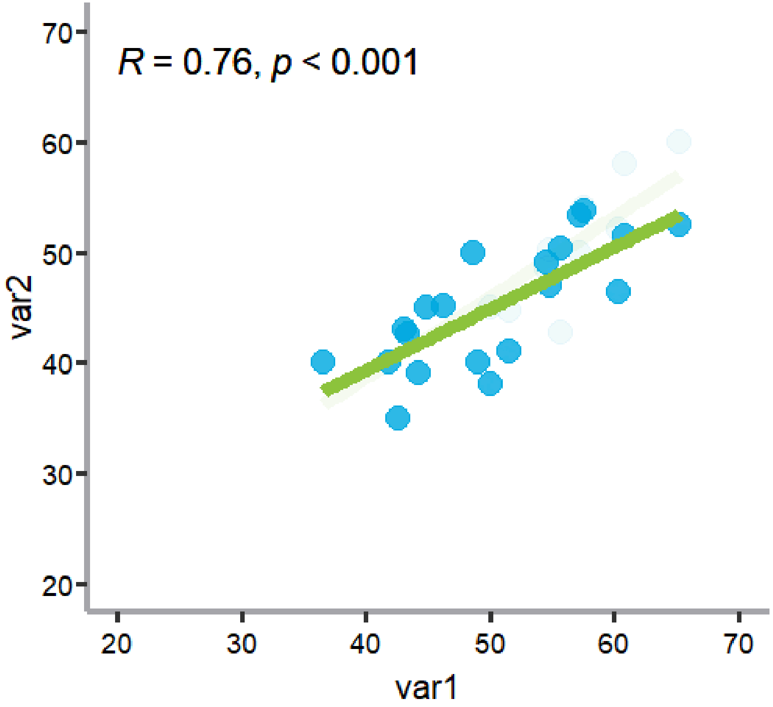
I-Step$_i$ → i → P-Step$_i$

🟢 fill in missing values

🔵 update regression coefficients

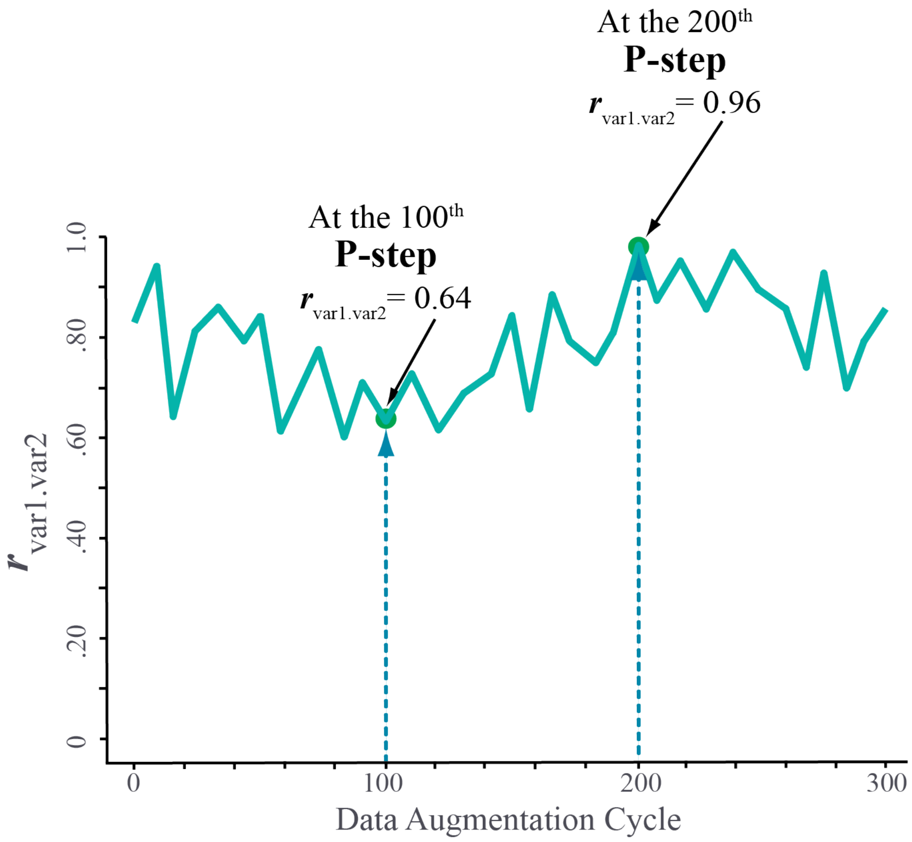
possible regression lines

$$\mathrm{var2}_i^* = \left[\hat{\beta}_0 + \hat{\beta}_1(\mathrm{var1}_i)\right] + z_i$$



Time-series plot

algorithm convergence



Analysis Phase

Run an analysis on each imputed data set separately.

Imputation	Mean	SE	r
1	43.23	1.03	0.51
2	42.40	1.15	0.23
3	44.97	1.03	0.77
4	48.10	1.69	0.85
5	45.35	1.07	0.81

imp1	imp2	imp3	imp4	imp5
40.00	40.00	40.00	40.00	40.00
40.00	40.00	40.00	40.00	40.00
35.00	35.00	35.00	35.00	35.00
43.00	43.00	43.00	43.00	43.00
42.60	42.60	42.60	42.60	42.60
39.00	39.00	39.00	39.00	39.00
45.00	45.00	45.00	45.00	45.00
45.20	45.20	45.20	45.20	45.20
50.00	50.00	50.00	50.00	50.00
40.00	40.00	40.00	40.00	40.00
41.55	44.70	45.08	50.74	48.94
44.79	40.05	48.33	47.74	44.61
42.45	39.58	45.25	54.46	47.49
36.53	36.14	47.40	61.35	47.93
47.00	37.99	46.01	47.35	48.00
46.01	46.91	45.48	55.91	46.05
44.40	50.41	52.85	54.95	49.09
46.19	35.84	47.53	54.59	52.68
40.68	54.29	50.16	59.05	50.38
55.18	42.21	51.55	55.98	52.03

Pooling Phase

Combine each analysis using Rubin's Rules.

Imputation	M	SE	r
1	43.23	1.03	0.51
2	42.40	1.15	0.23
3	44.97	1.03	0.77
4	48.10	1.69	0.85
5	45.35	1.07	0.81
Pooled	44.81	2.71	0.63

Fisher's z-transformation:

$r = 0.84$

Average estimates: $\hat{\theta} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}_t$

Mean: $\frac{(43.23+42.40+44.97+48.10+45.35)}{5} = 44.81$

Correlation: $\frac{(0.51+0.23+0.77+0.85+0.81)}{5} = 0.63$

Variance Within: $\frac{1}{m} \sum_{t=1}^{m} SE_t^2$

$$\frac{(1.03^2+1.15^2+1.03^2+1.69^2+1.07^2)}{5} = 1.49$$

Variance Between: $\frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}_t - \bar{\theta})^2$

$$\frac{\begin{array}{l}(43.23-44.81)^2+(42.40-44.81)^2+\\(44.97-44.81)^2+(48.10-44.81)^2+(45.35-44.81)^2\end{array}}{4} = 4.86$$

Pooled SE: $\sqrt{V_W + V_B + \frac{V_B}{m}}$

$$\sqrt{1.49 + 4.86 + \frac{4.86}{5}} = 2.71$$

Reporting

Step 1. Report rates of missing data for primary analysis variables.

> We analyzed the incomplete data and found missing values in all four analysis variables, with missing rates ranging from 2 to 28%. Of the 200 participants, we found 14,129 out of 119,400 total data points were missing (11.83%).

covariance coverage table

	1	2	3	4
1. var1	1.00			
2. var2	0.88	1.00		
3. var3	0.98	0.87	1.00	
4. var4	0.93	0.72	0.91	1.00

Extended Reporting. The covariance coverage table in the supplemental materials shows the proportion of cases with complete data for each individual variable (on the main diagonal) and for each variable pair (off-diagonal entries).

Reporting 📄 📄

Step 2. Imputation model assumptions.

> We found 5 missign data patterns. Typically, scores were missing because the data collection app randomly failed to save some responses. Little's Missing Completely at Random (MCAR) test indicated that missing data were not likely to derive from a MAR or MNAR mechanism, $\chi 2(12) = 6.45$, p = 0.892. Furthermore, mean comparisons revealed no significant differences among demographic variables between participants with complete data and those with incomplete data.

Extended Reporting. Provide sensitivity analyses for MI algorithms (special data structures = special algorithms) and/or MNAR models.

Reporting 📄 📄 📄

Step 3. Report methods and software.

> To address the issue of incomplete data, we employed the multiple imputation (MI) technique using the mice package in R. Specifically, we utilized the default mice() function using the norm algorithm to generate 100 imputed datasets under the MAR assumption. To ensure convergence and minimize autocorrelation between the imputed datasets, we set a conservative 60 iterations per imputation cycle. The imputation process involved 9 variables, including the four primary variables—var1, var2, var3, and var4—and five auxiliary variables: age, sex, race, education, and income.

Extended Reporting. Provide supplemental materials for details and code.

Considerations

- Sample size
- Rounding
- Number of imputations
- Interactions
- Categorical data
- Clustered data
- Large datasets

Myths 🤥

- Imputation is making up data
- Unfair to impute the DV
- Must have MAR to use MI/FIML
- Check imputation to make sure (sort of)
- Imputed values are meaningful
- No need with large samples

Any questions?