EP16: Missing Values in Clinical Research: Multiple Imputation

3. Analysis & Pooling

Nicole Erler

Department of Biostatistics, Erasmus Medical Center

n.erler@erasmusmc.nl

Erasmus MC
University Medical Center Rotterdam

Analysis Step

Multiple imputed datasets:

X_1	X_2	X_3	X_4
1.4	9.2	1.8	2.0
0.5	12.4	2.3	0.1
-0.5	10.7	2.6	-1.6
⋮	⋮	⋮	⋮

X_1	X_2	X_3	X_4
1.4	13.3	1.8	2.0
0.5	12.4	2.1	0.6
-0.5	10.2	2.6	-1.7
⋮	⋮	⋮	⋮

X_1	X_2	X_3	X_4
1.4	10.0	1.8	2.0
0.5	12.4	2.2	-1.4
-0.5	8.6	2.6	-1.0
⋮	⋮	⋮	⋮

Analysis Step

Analysis model of interest, e.g.,

$$x_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_4 + \varepsilon$$

Analysis Step

Analysis model of interest, e.g.,

$$x_1 = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_4 + \varepsilon$$

Multiple sets of results:

	est.	se
β_0	-0.15	0.22
β_1	0.16	0.02
β_2	-0.59	0.03
β_3	0.28	0.03

	est.	se
β_0	0.19	0.16
β_1	0.14	0.01
β_2	-0.59	0.03
β_3	0.20	0.03

	est.	se
β_0	0.04	0.22
β_1	0.14	0.01
β_2	-0.58	0.03
β_3	0.28	0.03

Pooling

Why pooling?

Recall from Section 1:
We need to represent missing values by a **number of imputations**.
➡ *m* imputed datasets

Pooling

Why pooling?

Recall from Section 1:
We need to represent missing values by a **number of imputations**.
➡ m imputed datasets

From the different imputed datasets we get **different sets of parameter estimates**, each of them with a standard error, representing the uncertainty about the estimate.

Pooling

Why pooling?

Recall from Section 1:
We need to represent missing values by a **number of imputations**.
➡ m imputed datasets

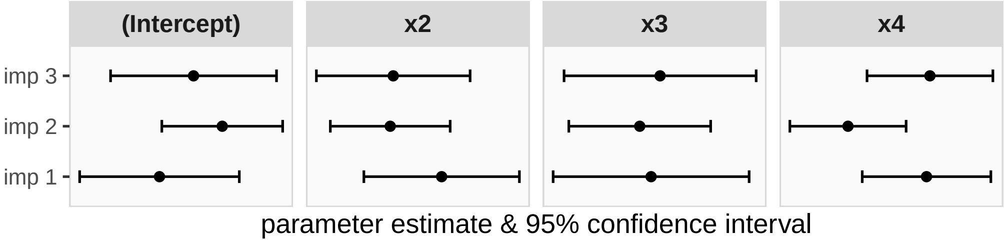
From the different imputed datasets we get **different sets of parameter estimates**, each of them with a standard error, representing the uncertainty about the estimate.

We want to **summarize** the results and describe **how (much) the results vary** between the imputed datasets.

Pooling

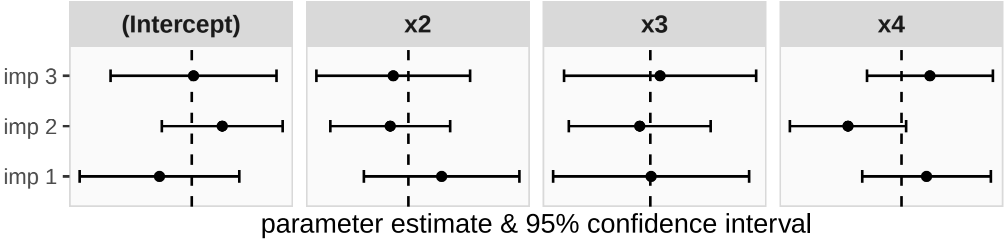
In the results from multiply imputed data there are **two types of variation/uncertainty**:

- **within** imputation (represented by the confidence intervals)
- **between** imputation (horizontal shift between results)



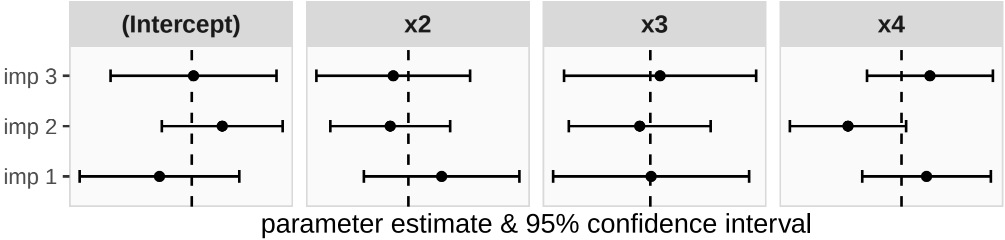
Pooling

To summarize the results, we can take the mean of the results from the separate analyses. This is the **pooled point estimate**.



Pooling

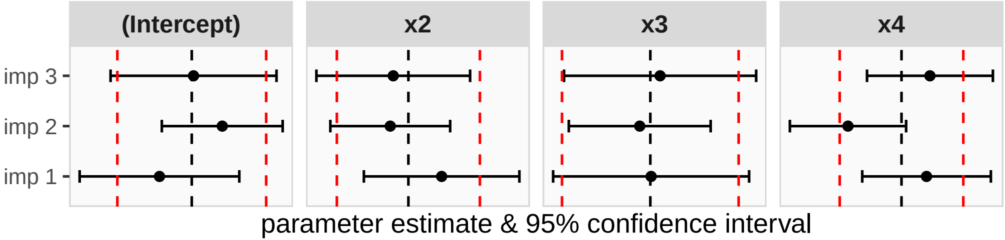
To summarize the results, we can take the mean of the results from the separate analyses. This is the **pooled point estimate**.



But does the same work for the standard error (or bounds of the CIs)?

Pooling

To summarize the results, we can take the mean of the results from the separate analyses. This is the **pooled point estimate**.



But does the same work for the standard error (or bounds of the CIs)?

The averaged CI's (marked in red) seem to underestimate the total variation (within + between).

Rubin's Rules

The most commonly used method to pool results from analyses of multiply imputed data was introduced by Rubin (1987), hence **Rubin's Rules**.

Notation:
m: number of imputed datasets
Q_ℓ: quantity of interest (e.g., regr. parameter β) from ℓ-th imputation
U_ℓ: variance of Q_ℓ (e.g., $var(\beta) = se(\beta)^2$)

Pooled parameter estimate:

$$\bar{Q} = \frac{1}{m}\sum_{\ell=1}^{m}\hat{Q}_\ell$$

Rubin's Rules

The **variance** of the pooled parameter estimate is calculated from the **within and between imputation variance**.

Average within imputation variance:

$$\bar{U} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{U}_\ell$$

Between imputation variance:

$$B = \frac{1}{m-1} \sum_{\ell=1}^{m} \left(\hat{Q}_\ell - \bar{Q}\right)^T \left(\hat{Q}_\ell - \bar{Q}\right)$$

Total variance:

$$T = \bar{U} + B + B/m$$

Rubin's Rules

Confidence intervals for pooled estimates can be obtained using the **pooled standard error** $\sqrt{T}$ and a **reference** t **distribution** with degrees of freedom

$$\nu = (m-1)\left(1 + r_m^{-1}\right)^2,$$

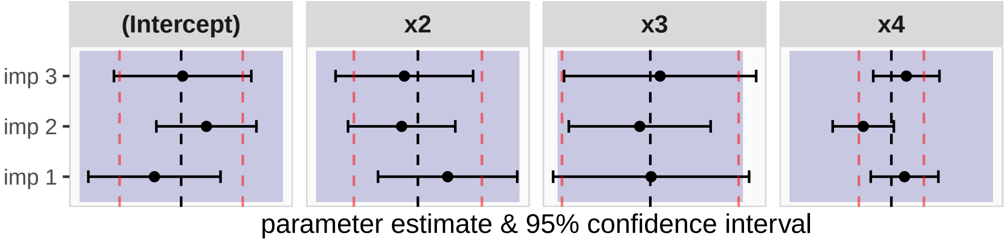
where $r_m = \frac{(B+B/m)}{\bar{U}}$ is the relative increase in variance that is due to the missing values.

The $(1-\alpha)$ **100% confidence interval** is then

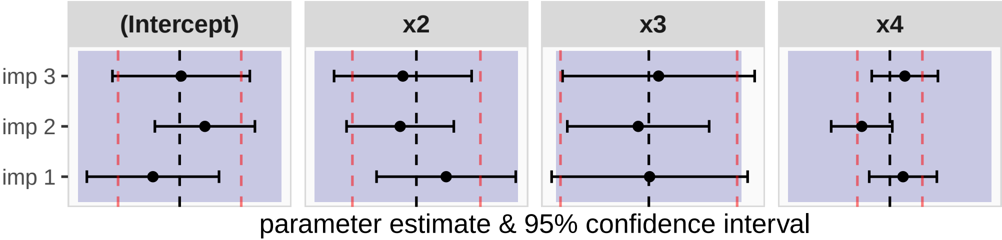
$$\bar{Q} \pm t_\nu(\alpha/2)\sqrt{T},$$

where t_ν is the $\alpha/2$ quantile of the t distribution with ν degrees of freedom.

Rubin’s Rules



Rubin's Rules



The corresponding **p-value** is the probability

$$Pr\left\{F_{1,\nu} > \left(Q_0 - \bar{Q}\right)^2 / T\right\},$$

where $F_{1,\nu}$ is a random variable that has an F distribution with 1 and ν degrees of freedom, and Q_0 is the null hypothesis value (typically zero).

References

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley. https://books.google.nl/books?id=0KruAAAAMAAJ