Incomplete Data Analysis

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR Plano de Recuperação e Resiliência

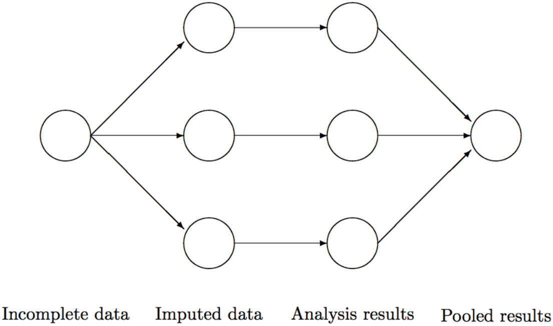
GOVERNO DE PORTUGAL

Financiado pela União Europeia NextGenerationEU

U.PORTO

FC FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

Multiple imputation



↪ In summary:

1. **Imputation**: impute multiple times.
2. **Analysis**: analyse each of the datasets.
3. **Pooling**: combine results, taking into account additional uncertainty.

Multiple imputation

Step 1

- ↪ Create a number ($M > 1$) of copies of the incomplete dataset, and use an appropriate procedure to impute the missing values in each of these copies.
- ↪ The imputed datasets are composed of a fixed portion – the observed data– and a varying portion – the imputed values. Since we do not know the true values that are missing it seems reasonable that the imputed values used in each copy should in general differ from each other.
- ↪ The choice of *M* is discussed later in the next set of slides.

Multiple imputation

Step 2

- We have created M imputed datasets that are now complete. How do we analyse them?
- For now, we will assume that our focus is on estimating a single (univariate) parameter, which we denote by θ.
- For instance, θ can be the mean or median of a variable, the proportion of individuals in a particular categorical (level) of a factor variable, a coefficient in a regression model, etc.
- For each imputed dataset, perform the analysis of interest (e.g., estimating the mean or fitting the regression model) that would have been performed in the absence of missing values. In the MI literature, the model of interest is sometimes referred to as the substantive model.
- Store the parameter estimate and its variance (the squared standard error). The estimate of θ obtained from the mth ($m = 1, \ldots, M$) complete dataset is denoted by $\widehat{\theta}^{(m)}$ and its (estimated) variance (squared standard error) by $\widehat{U}^{(m)}$.

Multiple imputation

Step 3

- After step 2, we have the results from M analyses, that is, we have $\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(M)}$, and $\widehat{U}^{(1)}, \ldots, \widehat{U}^{(M)}$.

- How do we now combine them to come up with a final estimate and how to measure the uncertainty about such estimate?

- According to the so-called **Rubin's rules**, the multiple imputation estimate of θ, $\widehat{\theta}^{\mathrm{MI}}$, is the average of the M individual estimates, that is,

$$\widehat{\theta}^{\mathrm{MI}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\theta}^{(m)}.$$

Multiple imputation

Step 3

- ↪ To estimate the variance of $\widehat{\theta}^{\mathrm{MI}}$, we **do not** simply average the variances from each dataset. It is slightly more complicated, but not that much complicated!

- ↪ First, we calculate the **between-imputation** variance

$$B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}^{(m)} - \widehat{\theta}^{\mathrm{MI}} \right)^2 ,$$

- ↪ This is simply the usual unbiased sample variance formula applied to $\widehat{\theta}^{(1)}, \ldots, \widehat{\theta}^{(M)}$.

- ↪ B measures how much the estimates of θ vary across the imputed datasets.

- ↪ If there is very little missing data, the estimates from the different imputed datasets will be very similar and the between imputation variance will be small.

- ↪ The larger the amount of missing data, the larger the variability in the estimates between the imputed datasets, and the larger the between imputation variance will be.

- ↪ B thus captures uncertainty in $\widehat{\theta}^{\mathrm{MI}}$ due to missing data.

Multiple imputation

Step 3

↪ Second, we calculate the **within-imputation** variance

$$\bar{U} = \frac{1}{M} \sum_{m=1}^{M} \widehat{U}^{(m)},$$

where $\widehat{U}^{(m)}$ is the estimated variance of $\widehat{\theta}^{(m)}$. This is simply the average of the individual variance estimates.

↪ The within imputation variance $\bar{U}$ is measuring the uncertainty due to the fact that the sample is of finite size (i.e., we are not using the entire population). This is the usual source of uncertainty in parameter estimates.

Multiple imputation

Step 3

- ↪ It is tempting to conclude that the **total variance** V^{MI} is equal to the sum of $\bar{U}$ and B, but that would be incorrect.
- ↪ We need to incorporate the fact that $\widehat{\theta}^{\text{MI}}$ itself is estimated using finite M, and thus only approximates $\widehat{\theta}^{\text{MI}}_{\infty}$, the estimator that would have been obtained for an infinitely large number of imputations $M = \infty$.
- ↪ Rubin (1987, eq. 3.3.5) shows that the contribution to the variance of this factor is systematic and equal to B_{∞}/M. Since B approximates B_{∞} (estimated between imputation variance for infinitely many imputations), we may write:

$$
\begin{aligned}
V^{\text{MI}} &= \bar{U} + B + \frac{B}{M} \\
&= \bar{U} + \left(1 + \frac{1}{M}\right) B,
\end{aligned}
$$

for the total variance of $\widehat{\theta}^{\text{MI}}$.

Multiple imputation

Step 3

- ↪ The inclusion of the term B/M is critical to make multiple imputation work at low values of M.

- ↪ Not including it would result in p-values that are too low or confidence intervals that are too short.

Multiple imputation

Step 3

↪ In summary, the total variance V^{MI} stems from three sources:

1. $\bar{U}$, the variance caused by the fact that we are taking a sample rather than observing the entire population. This is the conventional measure of variability.
2. B, the extra variance caused by the fact that there are missing values in the sample.
3. B/M, the extra simulation variance caused by the fact that $\widehat{\theta}^{\mathrm{MI}}$ itself is estimated for finite M.

↪ Note that if there were no missing values then B would be equal to zero and the estimated total variance V^{MI} would be $\bar{U}$.

Multiple imputation

Rubin's rules–toy example

- A confidence interval for θ can be constructed based on V^{MI} and $\widehat{\theta}^{\text{MI}}$.

- Specifically, the $(1-\alpha)100\%$ confidence interval is then

$$\widehat{\theta}^{\text{MI}} \pm t_\nu\left(\frac{\alpha}{2}\right)\sqrt{V^{\text{MI}}},$$

  with $t_\nu\left(\frac{\alpha}{2}\right)$ is the $\alpha/2$ quantile of the t distribution with $\nu = (M-1)(1+1/r_M)^2$, where $r_M = (1+1/M)B/\bar{U}$ is the relative increase in variance due to missing values.

- Notice that r_M does not depend on the sample size of the observed data. This can lead to situations where the degrees of freedom are larger than those for the complete case analysis, which is inappropriate.

- To avoid this problem, Barnard and Rubin (1999) proposed an improvement to calculate the degrees of freedom. This improved version is implemented in the `mice` package.

Multiple imputation

Rubin's rules–toy example

- Suppose we take a survey of five people, measuring their height and weight. Only three of them disclosure their weight; the other two don't give it just because of random chance. The data are:

Height (inches)	Weight (pounds)
65	130
68	140
70	150
72	NA
75	NA

- The aim of the analysis (step 2) is to regress the weight on the height, that is, our statistical model of interest is

$$\text{weight} = \beta_0 + \beta_1 \text{height} + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \sigma^2).$$

Multiple imputation

Rubin's rules–toy example

↪ Suppose that five plausible values for each missing weight have been generated (represented below in blue) to create five complete datasets.

Height	Weight–1	Weight–2	Weight–3	Weight–4	Weight–5
65	130	130	130	130	130
68	140	140	140	140	140
70	150	150	150	150	150
72	157	166	155	157	156
75	171	169	167	171	168
Estimated slope ($\widehat{\beta}_1$)	4.12	4.26	3.71	4.12	3.83
$\widehat{U} = \widehat{\text{var}}(\widehat{\beta}_1)$	(0.025)	(0.346)	(0.024)	(0.025)	(0.018)

Multiple imputation

Rubin's rules–toy example

- The final estimate for the slope is

$$\widehat{\beta}_1^{\mathrm{MI}} = \frac{1}{5}(4.12 + 4.26 + 3.71 + 4.12 + 3.83) = 4.008$$

- The within imputation variance is

$$\bar{U} = \frac{1}{5}(0.025 + 0.346 + 0.024 + 0.025 + 0.018) = 0.0876$$

- The between imputation variance is

$$\begin{aligned} B = &\frac{1}{4}\{(4.12 - 4.008)^2 + (4.26 - 4.008)^2 + (3.71 - 4.008)^2 + (4.12 - 4.008)^2 + (3.83 - 4.008)^2\} \\ = &\, 0.05227 \end{aligned}$$

- Thus, the final estimate of the variance is

$$V^{\mathrm{MI}} = 0.0876 + \left(1 + \frac{1}{5}\right) \times 0.05227 = 0.150324$$

Multiple imputation

Rubin's rules – multivariate case

- ↪ Extensions to the case where the parameter of interest is a *p*-component vector, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, are straightforward.
- ↪ For the estimate of the parameter vector we have

$$\widehat{\boldsymbol{\theta}}^{\mathrm{MI}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\boldsymbol{\theta}}^{(m)}.$$

- ↪ In the multivariate context, the within-imputation covariance matrix is the average of the *M* covariance matrices, namely

$$\bar{\mathbf{U}} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbf{U}}^{(m)},$$

where $\widehat{\mathbf{U}}^{(m)}$ is the covariance matrix from the completed dataset *m*.

Multiple imputation

Rubin's rules – multivariate case

- The between-imputation covariance matrix is as follows

$$\mathbf{B} = \frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{\text{MI}}\right) \left(\widehat{\boldsymbol{\theta}}^{(m)} - \widehat{\boldsymbol{\theta}}^{\text{MI}}\right)^{T},$$

  where $\widehat{\boldsymbol{\theta}}^{(m)}$ contains the parameter estimates from the mth imputed dataset, and $\widehat{\boldsymbol{\theta}}^{\text{MI}}$ is the vector of pooled point estimates (i.e., the arithmetic average of the $\widehat{\boldsymbol{\theta}}^{(m)}$ vectors).
- The diagonal elements of **B** contain the between imputation variance estimate for individual parameters, and the off-diagonal elements quantify the extent to which the between imputation fluctuation in one parameter is related to the between imputation fluctuation in another parameter.
- Considered as a whole, the between imputation covariance matrix represents the additional sampling fluctuation that results from the missing data.
- Finally, the total covariance matrix combined the within and between imputation covariance matrices as follows

$$\begin{aligned} \mathbf{V}^{\text{MI}} =& \bar{\mathbf{U}} + \mathbf{B} + \frac{1}{M}\mathbf{B}, \\ =& \bar{\mathbf{U}} + \left(I + \frac{1}{M}\right)\mathbf{B}, \end{aligned}$$

  where I is the identity matrix.

Multiple imputation

Some remarks on the 3 steps

- ↪ As we shall see, the only complex part of multiple imputation is step one: formulate a good imputation model.
- ↪ The specification of an appropriate imputation model is the key issue, since if this is misspecified, there is the potential for bias.
- ↪ The second step, producing the final estimate, is straightforward as it treats each imputed dataset as if it were a real dataset, we just have to do it M times.
- ↪ As Schafer (1997) says, multiple imputation works by "*solving an incomplete-data problem by repeatedly solving the complete-data version*".
- ↪ The third step involves simple arithmetic and typically we do not need to implement Rubin's rules manually as they are coded into most multiple imputation packages.

Multiple imputation

How many imputations?

- ↪ A natural question arising in the context of multiple imputation is how many copies of the dataset, which we have denoted by *M*, we should use.
- ↪ The choice of *M* **does not** affect the validity of our estimates and inferences.
- ↪ However, it **does** affect their statistical efficiency and reproducibility.

Multiple imputation

How many imputations?

- $\hookrightarrow$ Rubin originally suggested that unless the fraction of missing data was large, $M = 3$ or $M = 5$ would typically suffice. This advice was based on the statistical efficiency of the multiple imputation point estimate.

- $\hookrightarrow$ Recall that the total variance estimate, based on Rubin's rules, for the multiple imputation point estimate $\widehat{\theta}^{\mathrm{MI}}$, is given by

$$\mathrm{V}^{\mathrm{MI}} = \bar{U} + \left(1 + \frac{1}{M}\right) B.$$

- $\hookrightarrow$ The most efficient estimator is obtained with $M = \infty$, for which

$$\mathrm{V}^{\mathrm{MI}} = \bar{U} + B.$$

Multiple imputation

How many imputations?

- ↪ The ratio of the variance of the finite M estimate to the $M = \infty$ estimate is

$$\frac{\bar{U} + (1 + \frac{1}{M})B}{\bar{U} + B} = 1 + \frac{1}{M}\frac{B}{\bar{U} + B}.$$

- ↪ When the amount of missing data is small, the term $\frac{B}{\bar{U}+B}$ is close to zero, and even a small value of M means that the variance is not much increased compared to using $M = \infty$.

- ↪ This is the rationale behind the advice that usually a small M is okay. We need to take this advice with some grains of salt, as computing power in the 70s or 80s was somewhat limited.

Multiple imputation

How many imputations?

↪ In the last 10 years or so, focus has also been placed on how the choice of M affects the reproducibility of the results.

↪ Multiple imputation involves generating random numbers.

↪ As a result, the point estimates, standard errors, confidence intervals, p-values, etc, all have some inherent Monte-Carlo noise.

↪ If someone was to re-run our code (with a different seed!), they would get slightly different results.

↪ We may want to pick up a M large enough so that our results are (almost!) reproducible, in the sense that if someone re-run our code, they would get results sufficiently close enough to ours.

Multiple imputation

How many imputations?

- ↪ The simple but computationally expensive approach is trial and error.
- ↪ For instance, we can run our whole multiple imputation procedure, say 3 times, and compare results.
- ↪ If the results differ by more than what we are comfortable with, we should increase M and try again until results are close enough.
- ↪ Further, imputing a dataset in practice often involves trial and error to adapt and refine the imputation model. Such initial explorations do not require large M.
- ↪ It is convenient to set, e.g. $M = 5$, during model building, and increase M only after being satisfied with the model for the 'final' round of imputation.

Multiple imputation

Proper MI (or parameter uncertainty in MI!)

- ↪ The validity of MI rests on how imputations are created and how that procedure relates to the model used to subsequently analyse the data.

- ↪ Remember that we have learned that stochastic regression imputation was a promising approach.

- ↪ So, in the MI context (and for simplicity let us think about a univariate pattern of missingness), if we run stochastic regression imputation M times (i.e., for each missing value we use M draws instead of one) in step 1, is this all we have to do? Well, not exactly...But why?

- ↪ Such approach would imply that the regression coefficients and the variance of the error term are known with certainty. Such approach is termed in the literature as **improper multiple imputation**.

Multiple imputation

Proper MI (or parameter uncertainty in MI!)

↪ In practice, the regression coefficients and the variance of the error term are seldom known and must be estimated.

↪ If we had drawn a different sample from the same population, then our estimates for the regression coefficients and for the variance of the error term would be different, perhaps slightly.

↪ The amount of extra variability is strongly related to the sample size, with smaller samples yielding more variable estimates.

Multiple imputation

Proper MI (or parameter uncertainty in MI!)

- The parameter uncertainty also needs to be included in the imputations.
- Therefore, to perform **proper multiple imputation**, we need to reflect the parameters' variability/uncertainty from one imputation to the next.
- As an aside, the variability of the imputed values in stochastic regression imputation is composed of variability of estimation plus noise.
- There are two main methods for taking into account the parameter uncertainty:
  - **Bayesian methods** draw the parameters directly from their posterior distributions. That is, for each copy m of the dataset, $m = 1, \ldots, M$, we would draw the parameters from the posterior distribution.
  - **Bootstrap methods**, in turn, resample the complete cases and re-estimate the parameters from the resampled data.

Multiple imputation

Proper MI (or parameter uncertainty in MI!)

↪ It is useful to consider the consequences of improper multiple imputation.

↪ In such approach, point estimates would still be valid. However, the total variance V^{MI} computed using Rubin's rules would be too small, because the between imputation variability would not include the uncertainty due to parameter estimation.

↪ As a result, the confidence intervals based on V^{MI} would be too narrow.

Multiple imputation

Choosing the imputation model

- ↪ To provide valid estimates and inferences, MI requires data to be MAR and imputation models need to be correctly specified.
- ↪ Of course, as George Box famously said: "*All models are wrong, but some are useful*". We should nevertheless ensure that our models are a good approximation to the reality.
- ↪ In what concerns the first step of MI, we should ensure that the imputation model preserves any effects we are interested in estimating/modelling (in the substantive model of step 2).
- ↪ If, for instance, the substantive model of interest includes interactions (between variables), then these should be preserved in the imputation model.
- ↪ Meng (1994) introduced the concept of congeniality to refer to the relation between the imputation model and the analysis model. The imputation model should be 'congenial' with the substantive model.