Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

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?

- 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.

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

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

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

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

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.