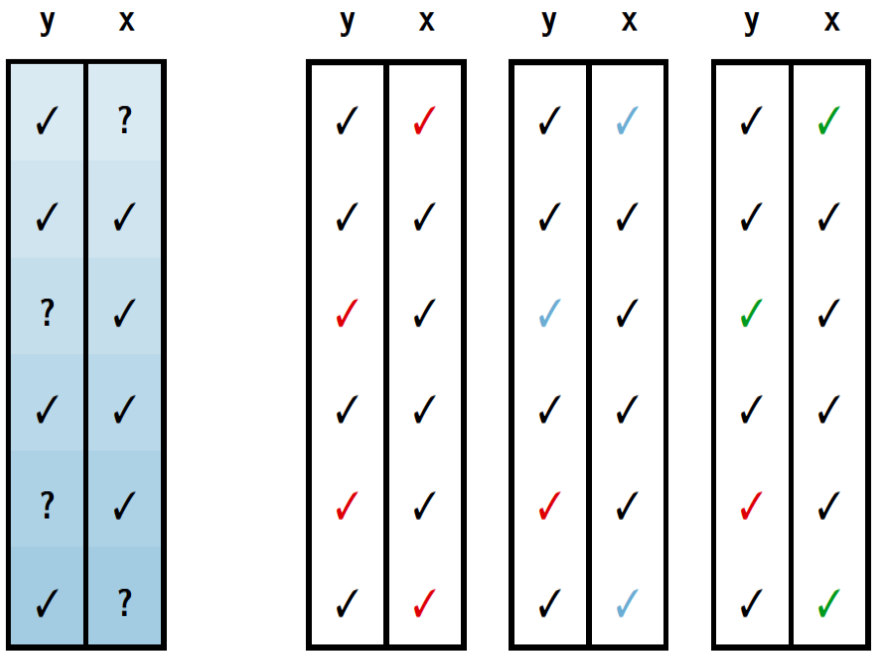
IDS 702: Module 5.3

Imputation methods II

Dr. Olanrewaju Michael Akande

MI RECAP

- Fill in dataset m times with imputations.
- Analyze repeated datasets separately, then combine the estimates from each one.
- Imputations drawn from probability models for missing data.



(a) Observed Data

(b) MI Datasets $(1, \ldots, m)$

MI RECAP

Rubin (1987)

- Population estimand: Q
- Sample estimate: q
- Variance of q: u
- In each imputed dataset d_j, where $j = 1, \ldots, m$, calculate

$$q_j = q(d_j)$$

$$u_j = u(d_j).$$

MI RECAP

- MI estimate of Q:

$$\bar{q}_m = \sum_{i=1}^{m} \frac{q_i}{m}.$$

- MI estimate of variance is:

$$T_m = (1 + 1/m)b_m + \bar{u}_m.$$

where

$$b_m = \sum_{i=1}^{m} \frac{(q_i - \bar{q}_m)^2}{m-1}; \quad \bar{u}_m = \sum_{i=1}^{m} \frac{u_i}{m}.$$

- Use t-distribution inference for Q

$$\bar{q}_m \pm t_{1-\alpha/2}\sqrt{T_m}.$$

MI: PROS AND CONS

- Advantages
  - Straightforward estimation of uncertainty
  - Flexible modeling of missing data
- Disadvantages (??)
  - Extra data sets to manage
  - Explicitly model-based

Resources for learning more

- Little and Rubin (2002), *Statistical Analysis with Missing Data*, Wiley
- Schafer (1997), *Analysis of Incomplete Multivariate Data*, CRC Press
- Reiter and Raghunathan (2007), "The multiple adaptations of multiple imputation," *JASA*.

Where should the imputations come from?

MI: WHERE SHOULD THE IMPUTATIONS COME FROM?

So where should we get reasonable replacements for the missing values from? There are two general approaches:

- Sequential modeling
  - Estimate a sequence of conditional models (think separate regressions for each variable!);
  - Impute from each model.
- Joint modeling
  - Choose a multivariate model for all the data (we will not cover joint multivariate models in this class; we will in STA602);
  - Estimate the model;
  - Impute from the joint model.

MI: SEQUENTIAL REGRESSION MODELS

Suppose the data includes three variables Y_1, Y_2, Y_3.

- Step 1: fill in missing values by simulating values from regressions based on complete cases;
- Step 2: regress $Y_1|Y_2, Y_3$ using completed data;
- Step 3: impute new values of Y_1 from this model;
- Step 4: repeat for $Y_2|Y_1, Y_3$ and $Y_3|Y_1, Y_2$ (repeat for all variables with missing data);
- Step 5: cycle through Steps 1 to Step 4 many times;
    - Usually the default number is 5, but there is not theory underpinning this default.

Final dataset is one imputed dataset. Repeat entire process m times to get m multiply-imputed datasets.

Existing software for sequential modeling

Free software packages

- MICE for R and Stata (so many conditional models to pick from, for example, predictive mean matching, random forest, linear regression, logistic regression, and so on);
- statsmodels MICE in python (only uses predictive mean matching);
- MI for R;
- IVEWARE for SAS.

In sequential modeling, one can specify many types of conditional models and include constraints on values.

Existing software for joint modeling

- Multivariate normal data
  - R: NORM, Amelia II;
  - SAS: proc MI;
  - Stata: .hlight[MI command.hlight[.
- Mixtures of multivariate normal distributions
  - R: EditImpCont (also does editing of faulty values).
- Multinomial data:
  - R: CAT (log-linear model), NPBayesImpute (latent class model).

Existing software for joint modeling

- Nested Multinomial data:
  - R: NestedCategBayesImpute (also generates synthetic data).
    update coming soon to allow for editing of faulty values
- Mixed data:
  - R: MIX (general location model).
- Many other joint models, but often without open source software.

Comparing sequential to joint modeling

Advantages

- Often easier to specify reasonable conditionals than a joint model.
- Complex samplers not often needed.
- Can use machine learning methods for conditionals.

Disadvantages

- Labor intensive to specify models.
- Incoherent conditionals can cause odd behaviors (e.g., order matters).
- Theoretical properties difficult to assess.

What if imputation and analysis model do not match?

- Imputation model more general than analysis model: conservative inferences.
- Imputation model less general than analysis model: invalid inferences.
- For sequential modeling, include all variables related to outcome and missing data (Schafer 1997).
- Include design information in models (Reiter *et al*. 2006, *Survey Methodology*).

Evaluating the fit of imputation models

- Plots of imputed and observed values (Abayomi *et al*, 2008, *JRSS-C*)
  - Imputed values that don't look like the observed values could *maybe* imply poor imputation models;
  - Useful as a sensibility check
- Model-specific diagnostics (Gelman *et al*, 2005, *Biometrics*)
  - Take a look at residual plots with marked observed and imputed values;
  - Look for obvious abnormalities.

Remarks

- Ignoring missing data is risky.
- Single imputation procedures at best underestimate uncertainty and at worst fail to capture multivariate relationships.
- Multiple imputation recommended (or other model-based methods).
- We discussed MI for MAR data. When data are NMAR, analysis can be much harder.
- In those scenarios, get missing data experts on your team.

Remarks

- Incorporate all sources of uncertainty in imputations, including uncertainty in parameter estimates.
- Want models that accurately describe the distribution of missing values.
- Important to keep in mind that imputation model are only used for cases with missing data.
  - Suppose you have 30% missing values;
  - Also, suppose your imputation model is "80% good" ("20% bad");
  - Then, completed data are only "6% bad"!

What's next?

Move on to the readings for the next module!