IDS 702: Module 5.2

Imputation methods I

Dr. Olanrewaju Michael Akande

Strategies for handling missing data

- Item nonresponse:
  - use complete/available cases analyses
  - single imputation methods
  - multiple imputation
  - model-based methods
- Unit nonresponse:
  - weighting adjustments
  - model-based methods (identifiability issues!).
- We will only focus on item nonresponse.
- If you are interested in building models for both unit and item nonresponse, here is a paper on some of the research I have done on the topic: https://arxiv.org/pdf/1907.06145.pdf

IDS 702

Complete/available cases analyses

What can happen when using available case analyses with different types of missing data?

- MCAR: unbiased when disregarding missing data; variance increase (losing partially complete data)
- MAR: biased (depending on the strength of MAR and amount of missing data) when missing data mechanism is not modeled; variance increase (losing partially complete data).
- NMAR: generally biased!

IDS 702

Single imputation methods

- Marginal/conditional mean imputation
- Nearest neighbor imputation:
  - hot deck imputation
  - cold deck imputation
- Use observation from one of the previous time periods (for panel data)
  - LOCF -- last observation carried forward
  - BOCF -- baseline observation carried forward

IDS 702

Mean imputation

Plug in the variable mean for missing values.

- Point estimates of means OK under MCAR
- Variances and covariances underestimated.
- Distributional characteristics altered.
- Regression coefficients inaccurate.

Similar problems for plug-in conditional means.

Nearest neighbor imputation

Plug in donors' observed values.

- Hot deck: for each non-respondent, find a respondent who "looks like" the non-respondent in the same dataset
- Cold deck: find potential donors in an external but similar dataset. For example, respondents from a 2016 election poll survey might serve as potential donors for non-respondents in the 2018 version of the same survey.
- Common metrics: Statistical distance, adjustment cells, propensity scores.

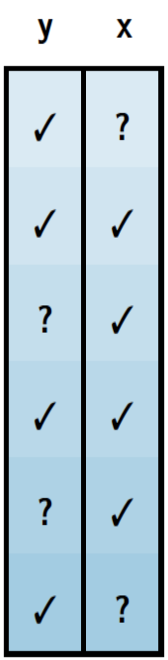
IDS 702

Nearest neighbor imputation

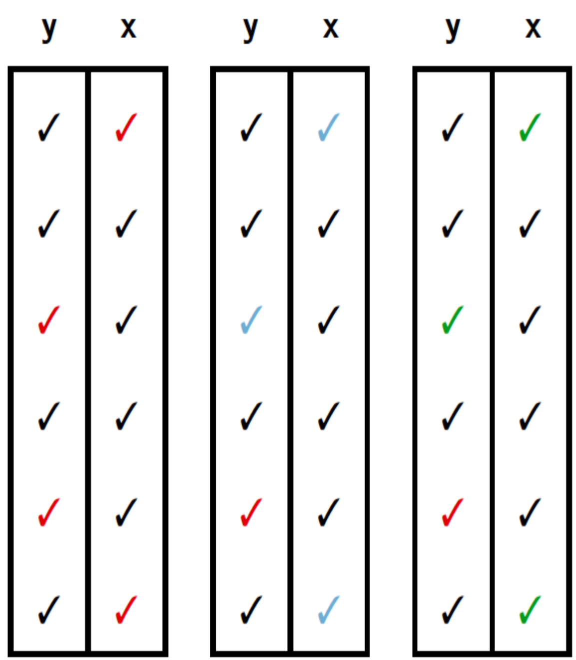
- Point estimates of means OK under MAR.
- Variances and covariances underestimated.
- Distributional characteristics OK.
- Regression coefficients OK under MAR.

Multiple imputation (MI)

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



(a) Observed Data



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

IDS 702

MI EXAMPLE

Suppose

- $Y =$ income (unit of measurement is $10,000)
- $X =$ level of education (0 = undergraduate, 1 = graduate)

Y	X
11.9	1
16.1	1
12.9	0
?	0
12.1	?
12.6	0
?	1

(a) Data

Y	X	Y	X	Y	X
11.9	1	11.9	1	11.9	1
16.1	1	16.1	1	16.1	1
12.9	0	12.9	0	12.9	0
11.8	0	12.8	0	13.0	0
12.1	1	12.1	0	12.1	1
12.6	0	12.6	0	12.6	0
11.2	1	13.6	1	11.7	1

(b) Multiply-imputed datasets

MI: INFERENCES FROM MULTIPLY-IMPUTED DATASETS

Rubin (1987)

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

$$q_j = q(d_j)$$

$$u_j = u(d_j)$$

IDS 702

MI EXAMPLE: INFERENCES FROM MULTIPLY-IMPUTED DATASETS

Suppose we are interested in estimating the mean income in our example. Then

- Q = μ_Y
- $q = \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$
- u = $\hat{\mathbb{V}}[\bar{y}] = \dfrac{s^2}{n}$
- In each imputed dataset d_j, calculate

$$q_j = \bar{y}_j \quad \text{and} \quad u_j = \frac{s_j^2}{n}$$

IDS 702

MI: QUANTITIES NEEDED FOR INFERENCE

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

IDS 702

MI: INFERENCES FROM MULTIPLY-IMPUTED DATASETS

- MI estimate of Q:

$$\bar{q}_m$$

- MI estimate of variance is:

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

- Use t-distribution inference for Q

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

Notice that the variance incorporates uncertainty both from within and between the m datasets.

IDS 702

MI EXAMPLE

Back to our income example,

Y	X
11.9	1
16.1	1
12.9	0
11.8	0
12.1	1
12.6	0
11.2	1

$q_1 = \bar{y} = 12.66$

$u_1 = \hat{\mathbb{V}}[\bar{y}] = 0.37$

Y	X
11.9	1
16.1	1
12.9	0
12.8	0
12.1	0
12.6	0
13.6	1

$q_2 = \bar{y} = 13.14$

$u_2 = \hat{\mathbb{V}}[\bar{y}] = 0.29$

Y	X
11.9	1
16.1	1
12.9	0
13.0	0
12.1	1
12.6	0
11.7	1

$q_3 = \bar{y} = 12.90$

$u_3 = \hat{\mathbb{V}}[\bar{y}] = 0.32$

By the way, $\bar{y} = 12.64$ from the "true complete dataset".

IDS 702

MI EXAMPLE

- MI estimate of Q:

$$\bar{q}_m = \sum_{j=1}^{m} \frac{q_j}{m} = \frac{12.66 + 13.14 + 12.90}{3} = 12.90$$

- Between variance

$$b_m = \sum_{j=1}^{m} \frac{(q_j - \bar{q}_m)^2}{m-1} = 0.06$$

- Within variance

$$\bar{u}_m = \sum_{j=1}^{m} \frac{u_j}{m} = \frac{0.37 + 0.29 + 0.32}{3} = 0.33$$

- MI estimate of variance is:

$$T_m = (1 + 1/m)b_m + \bar{u}_m = (1 + 1/3)0.06 + 0.33 = 0.41$$

Where should the imputations come from? We will answer that soon!

What's next?

Move on to the readings for the next module!