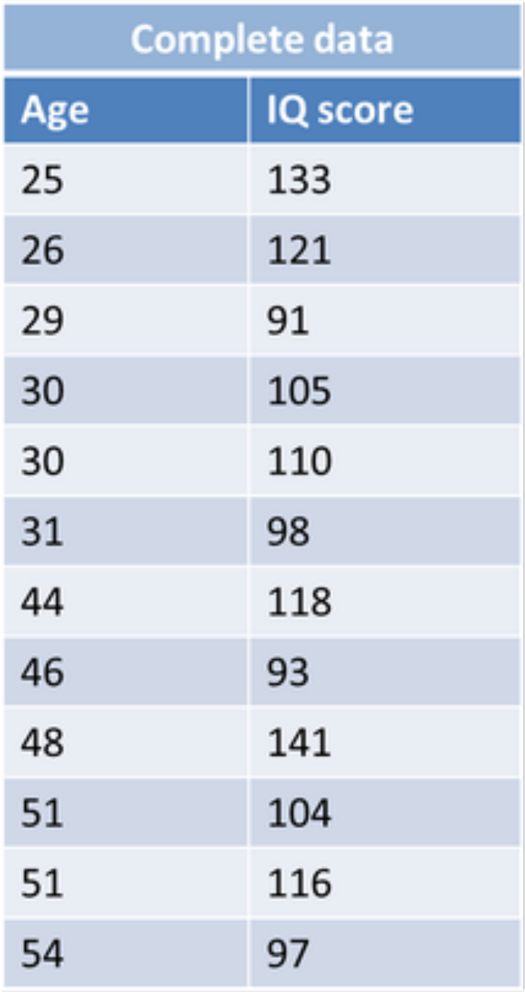
Lecture 6

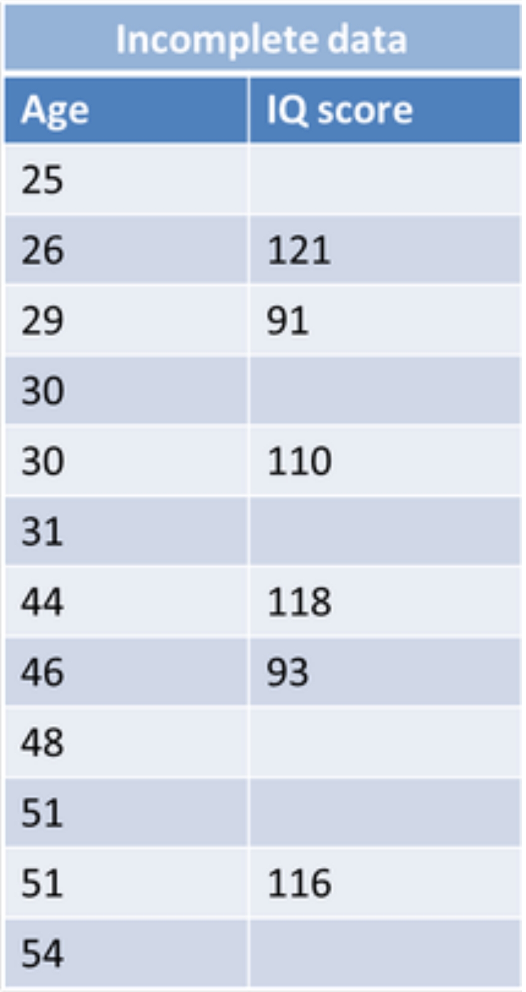
Missing Data and Imputation

Instructor: Hamed Hasheminia

Missing Data



Complete data	
Age	**IQ score**
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97



Incomplete data	
Age	**IQ score**
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Ignoring or inappropriately handling missing data may lead to…

- Biased estimates
- Incorrect standard errors
- Incorrect inferences/results

Missing Data Mechanisms

- Missing Completely at Random (MCAR)
  - Dropping unobserved observations is fine!
- Missing at Random (MAR): Missingness depends on observed data
  - E.g., women more likely to respond to men
  - Can use weighting or ***imputation*** approaches to deal with the missingness
  - Usually people try to solve this type of missing data issue
- Not missing at random (NMAR): Missingness depends on unobserved values
  - Probability of reporting psychiatric treatment depends on whether or not they have received it
  - Probability of someone reporting their income depends on what their income is
  - No easy way to deal with that!

Improper ways of handing missing Data!

- Ignoring it
    - Ignore it; just run models without doing anything about missingness
    - You only rely on defaults of the software
- Complete Case (listwise deletion)
    - Restricts the analysis to only observed data
        - Assumes missingness is MCAR
        - Lowers the power of your models
        - Generally leads to biased results

Improper ways of handing missing Data!

- Single Imputation (Fill in ("impute") each missing value
  - Mean/Median/fill backward/fill forward
  - Regression Prediction ("Conditional Mean Imputation")
    - Impute mean within categories of observed covariates (gender, race, etc.)
    - Fit regression model among observed cases, use to predict response for individuals with missing values

Slightly better ways of handing missing Data!

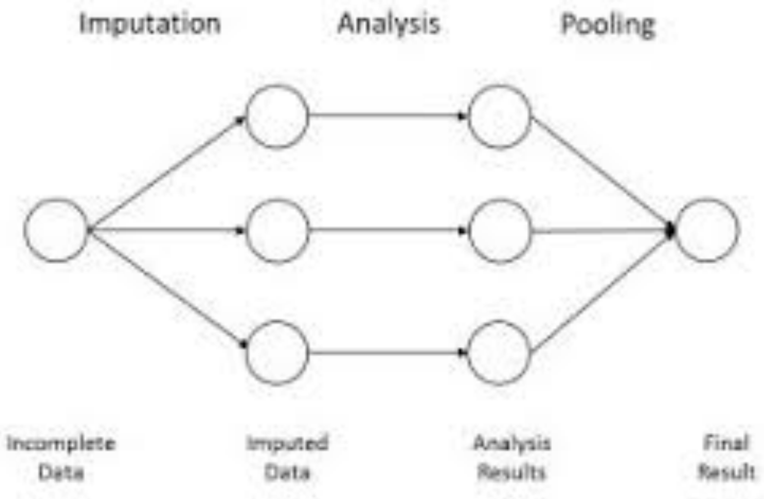
- "Hot-deck"
  - for an individual with missing data, find individuals with the same observed values on other variables, randomly pick one of their values as the one to use for imputation
- Regression Prediction plus error ("Stochastic regression imputation")
  - Same as regression imputation – you just add random error to it.

Summary of single imputation approaches

- among those discussed, the best are regression prediction plus error or hot-deck (based on categorical versions of all the variables observed)
- Can be reasonable if you have few missing data (< 5%)
- BUT .. Results in overly precise estimates
  - Simply treats all of the values as observed values
  - So does not take into account the uncertainty in the imputations.
- What should we do instead?

Appropriate ways of handling missing values

- Weighting
- Multiple Imputation
- Maximum Likelihood (FIML) * Above the scope of this class



- OUR GOAL IS NOT TO GET CORRECT PREDICTIONS OF MISSING VALUES; GOALS IS TO OBTAIN ACCURATE PARAMETER ESTIMATES FOR RELATIONSHIPS OF INTEREST

Nonresponse Weighting

- Often used to deal with Attrition
- Generate model predicting response given observed covariates
- Weight Respondent by their inverse probability of response
  - Weights the respondents up to represent the full sample
  - Same idea as survey sampling weight
- Use Analysis methods that allow for weights
- Works well for simple missing data patterns (e.g. Attrition)

Nonresponse Weighting

- A simple example …
- Imagine 100 males and 100 females in sample
- But only 80 males and 75 females respond
- Male respondents will get weight of 100/80 = 1.25
- Female respondents will get weights of 100/75 = 1.33

Multiple Imputation

- Same idea as single imputation, but fills in each missing value multiple times
    - Like repeating the stochastic mean imputation multiple times
- Creates multiple (e.g., 100) “complete” data sets
- Analyses then run separately on each dataset and result ***combined*** across datasets

Summary

- We have three types of missing data MCAR, MAR, NMAR
- For MCAR – we can simply drop missing data
- For MAR we can use single imputation or multiple imputation methods. It is highly recommended that you use multiple imputation methods – unfortunately, standard libraries in python do not handle this properly.
- For NMAR, there is no known way to handle missing data.

Acknowledgment

- Most of today's lecture is extracted from Lecture notes of Dr. Elizabeth A. Stuart