UMC Utrecht

Imputation of Missing Data

Thomas Debray, PhD
Assistant Professor
T.Debray@umcutrecht.nl

Jeroen Hoogland, GradStat
Junior Researcher
J.Hoogland-2@umcutrecht.nl

Recap

- Why should we worry about missing data?
  - Ignoring missing data (complete/available case analysis) <u>looses power</u> (always) and <u>introduces bias</u> (in non-MCAR situations)
  - Ad hoc methods such as (subgroup)mean cause <u>bias</u> and may affect both <u>power</u> and <u>type I error</u>
  - The missing indicator method gives <u>biased</u> estimates unless the study design specifically suits the method

More advanced methods are needed!

Recap

- What are the main types of missing data?

- *Missing completely at random*—There are no systematic differences between the missing values and the observed values. For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer

- *Missing at random*—Any systematic difference between the missing values and the observed values can be explained by differences in observed data. For example, missing blood pressure measurements may be lower than measured blood pressures but only because younger people may be more likely to have missing blood pressure measurements

- *Missing not at random*—Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values. For example, people with high blood pressure may be more likely to miss clinic appointments because they have headaches

BMJ 2009; 338 doi: https://doi.org/10.1136/bmj.b2393

Methods to handle missing data

Complete case analysis (CC)

Available case analysis (AC)

Missing indicator method

Overall mean/median imputation

Subgroup mean/median imputation

Single (multivariable) regression based imputation

Multiple regression based imputation

Univariate missing data

Consider we have one variable with missing observations for some participants.

	vacc	age	sex	cvd	pulm	DM	contact	hosp
1	1	66	0	NA	1	0	27	0
2	1	73	1	1	0	0	4	0
3	1	75	1	NA	0	0	8	0
4	1	76	1	1	0	0	7	0
5	1	77	1	1	0	0	7	0
6	1	78	1	1	0	0	5	0
7	1	80	1	NA	0	0	9	0
8	1	81	1	NA	0	0	17	0
9	1	66	0	0	0	0	10	0
10	1	67	0	NA	0	0	13	0
11	1	69	0	0	0	0	5	0
12	1	70	0	0	0	0	5	0
13	1	66	0	1	0	0	13	0
14	0	67	0	1	0	0	35	0
15	1	67	0	NA	0	0	5	0

Single imputation by regression

- Develop a prediction model to predict the missing values based on the observed data
  - Include the outcome (!) of the analysis model
  - Include all variables of the analysis model
  - Include yet (!) unknown predictors of the missing value

- This prediction model is also called the imputation model

- Imputation model is used to estimate for each subject with missing cvd its actual value, given all other predictor (including outcome) values

Single imputation by regression

- In our example, we can use logistic regression:

$$\Pr(\mathrm{cvd} = 1) = \mathrm{logit}^{-1}(a + b_1\mathrm{vacc}_i + b_2\mathrm{age}_i + b_3\mathrm{sex}_i + b_4\mathrm{pulm}_i + b_5\mathrm{DM}_i + b_6\log(\mathrm{contact}_i) + b_7\mathrm{hosp}_i)$$

In R, we have:

```
impmodel = glm(cvd ~ vacc + age + sex + pulm + DM + log(contact) + hosp,
family=binomial(), data=data)
```

- The imputation model can only be fitted using participants with complete data

Single imputation by regression

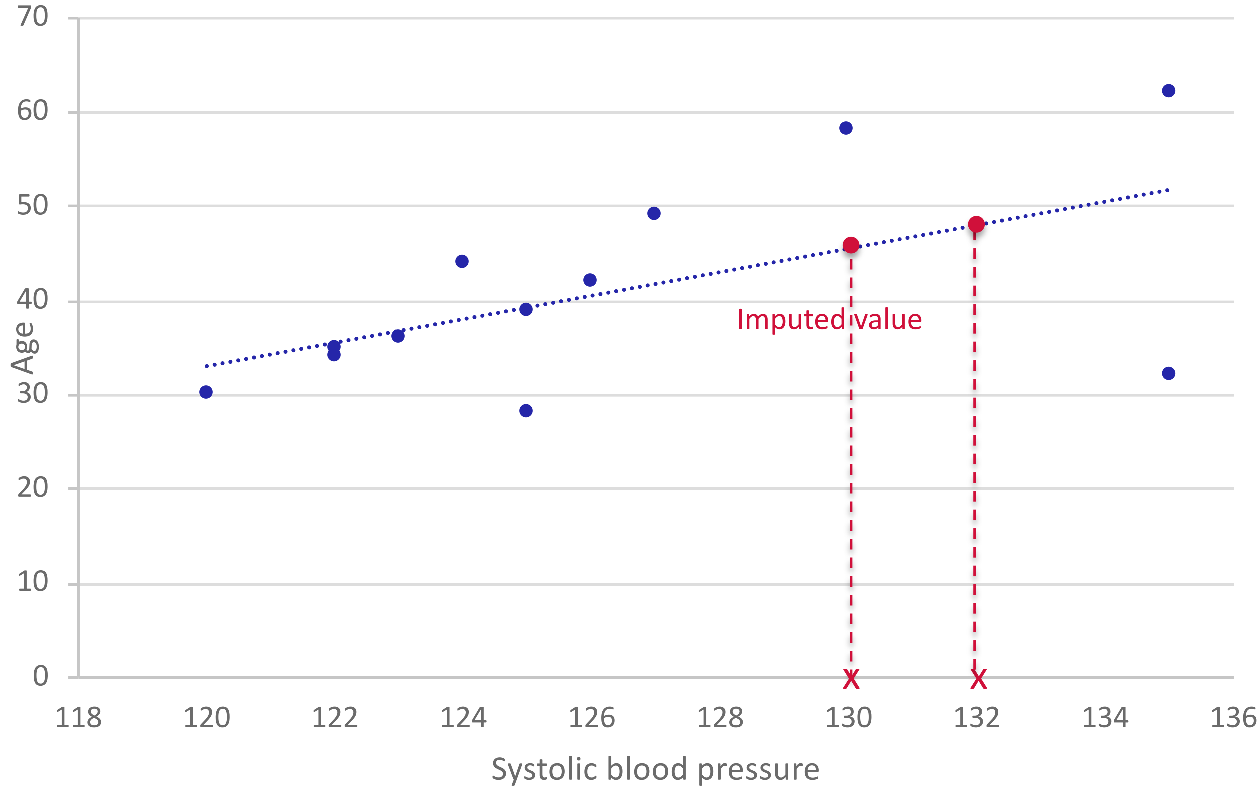
- Use predicted value directly as imputation

	vacc	age	sex	cvd	pulm	DM	contact	hosp
1	1	66	0	0.6688367	1	0	27	0
2	1	73	1	1.0000000	0	0	4	0
3	1	75	1	0.3919347	0	0	8	0
4	1	76	1	1.0000000	0	0	7	0
5	1	77	1	1.0000000	0	0	7	0
6	1	78	1	1.0000000	0	0	5	0
7	1	80	1	0.4549462	0	0	9	0
8	1	81	1	0.6397881	0	0	17	0
9	1	66	0	0.0000000	0	0	10	0
10	1	67	0	0.5355935	0	0	13	0
11	1	69	0	0.0000000	0	0	5	0
12	1	70	0	0.0000000	0	0	5	0
13	1	66	0	1.0000000	0	0	13	0
14	0	67	0	1.0000000	0	0	35	0
15	1	67	0	0.2778990	0	0	5	0

In R: `data$cvd[is.na(data$cvd)] <- predict(impmodel, newdata=data, type="response")[is.na(data$cvd)]`

Single imputation by regression

Example for a different missing data problem:



- Key benefit: uses information from all covariates and outcome
- Key problem: **lack of variation, multivariate missing data**

Single imputation by regression

Lack of variation

Each patient with same co-variates has same imputed value

#1 Uncertainty due to natural variation is ignored
(imputed values are conditional mean values)

→ Imputation will inflate the correlations in our data and thereby introduce bias

#2 Uncertainty of the estimated imputation model is ignored
(the imputation model coefficients are treated as true)

→ Estimated standard errors will be too low

Multivariate missing data

#3 Fitting of the imputation models cannot use rows with missing data! → return of the complete/available case analysis problem

Single imputation by regression

#1 Adding natural variation (binary case):

- Recall

  $$\pi_{\text{cvd}} = \Pr(\text{cvd} = 1) = \text{logit}^{-1}(a + b_1\text{vacc}_i + b_2\text{age}_i + b_3\text{sex}_i + b_4\text{pulm}_i + b_5\text{DM}_i + b_6\log(\text{contact}_i) + b_7\text{hosp}_i)$$

- Imputation is then given by a *random sample* from Bernoulli$\left(\pi_{\text{cvd}}\right)$

Single imputation by regression

<u>**#2** Adding natural variation (continuous case):</u>

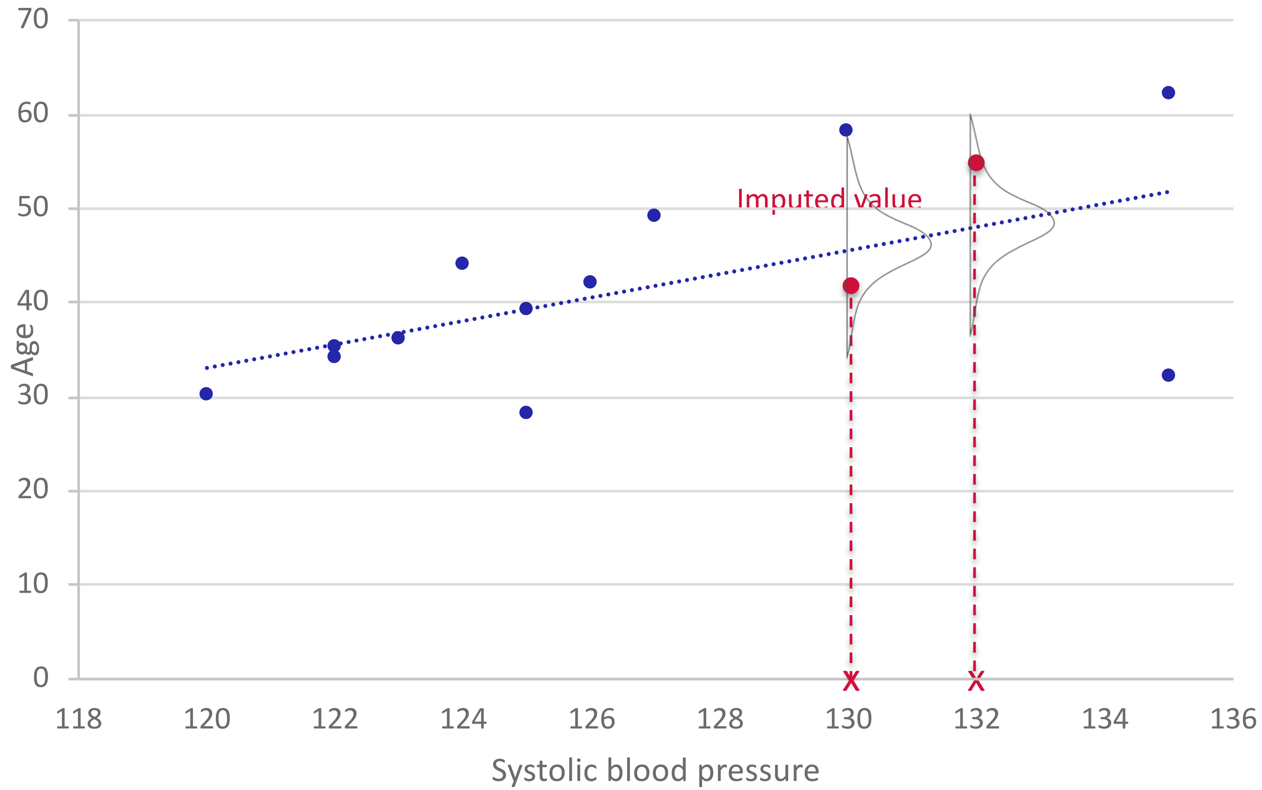
- We can adopt linear regression

  $\mu_{\text{age}} = (a + b_1\text{vacc}_i + b_2\text{cvd}_i + b_3\text{sex}_i + b_4\text{pulm}_i + b_5\text{DM}_i + b_6\log(\text{contact}_i) + b_7\text{hosp}_i)$

  This also yields a residual error variance term σ^2_{age}

- Imputation is then given by a *random sample* from $\text{Normal}\left(\mu_{\text{age}}, \sigma^2_{\text{age}}\right)$

Single imputation by regression



Single imputation by regression

After taking care of problem #1

- Under M(C)AR: unbiased regression coefficient (no bias)
  - SI regression without error term = biased result
  - With error term is best one can get with SI = unbiased result

- But SE still underestimated → too easily significant
  - Beta's of prediction model also estimated → not accounted for
    - Can be dealt with; we will return to this lateron
  - **All data are treated as if they were observed**
    - **Principle limitation of single imputation**

⇒ Need for more advanced imputation strategies!

⇒ **Next lecture**