Biostatistics & Epidemiological Data Analysis using R

7

Missing values

Jana Fehr & Stefan Konigorski

Digital Health & Machine Learning, HPI

November 30, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2022.10.19
	2	First steps in data analysis using R	2022.10.26
	3	Second steps in data analysis using R	2022.11.02
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2022.11.09
	5	Estimation	2022.11.16
	6	Hypothesis testing and study planning	2022.11.23
	7	Missing data	2022.11.30
Data analysis w/ regression models	8	Linear regression I	2022.12.07
	9	Linear regression II	2022.12.14
	10	Regression models for binary and count data	2023.01.11
	11	Analysis of variance & Linear mixed models I	2023.01.18
	12	Linear mixed models II & Meta analysis	2023.01.25
	13	Survival analysis	2023.02.01
	14	Causal inference & Data analysis challenge	2023.02.08

(see full schedule online)

Learning objectives

- Understand different types of missing values
- Get to know different ways how to deal with missing values
- Understand potential biases of naive/"traditional" approaches for dealing with missing values
- Get to know Multiple Imputation as one of the recommended methods, and apply it to examples

Introduction

Motivation & Introductory example

Example: sleep & wellbeing

- Study question: Is there an association between the length of nightly sleep of children and their psychological wellbeing?
- Question: How can you investigate this study question in the KiGGS dataset?
  $\longrightarrow$ see `R_7a_missing_values_intro_example.Rmd`
  $\longrightarrow$ use external rating of psychological wellbeing by parents
- To investigate the study question, you have to choose an appropriate statistical model, and also think about how to deal with missing values

When and how do missing values arise?

Example: questionnaire survey

- The question/variable was forgotten while filling out.
- Participant didn't know how to answer a question.
- The proper answer category was not available.
- No motivation to answer questions at the end of the questionnaire.
- Questionnaire has to be filled out in a given time, no time to fill out the questions at the end.
- Question is not relevant (e.g. pregnant, men).
- Proband cannot remember the answer, e.g. date of operation.

When and how do missing values arise?

Example: questionnaire survey

- Question is not answered due to fear that answering it will harm the participant.
- Participant was distracted by outdoor noise.
- In a study over multiple time points (using the same or different questionnaires), the participant could only come to one appointment.

$\longrightarrow$ Focus on values that (theoretically) should be there

Question: In which of these situations could it matter for the analysis (e.g. bias the results) how the missing values are dealt with?

When and how do missing values arise?

Example: in prospective cohort study

- Study: can a new biomarker xy and body composition estimated from MRI scans (at baseline) predict the risk of Alzheimer's disease?
- Which missing values can occur here (biomarker, MRI scans, Alzheimer)?
- When and which types of these missing values could affect the effect estimation (e.g. of odds ratios) and hypothesis testing?

Summary

All discussions how to deal with missing values, all statistical methods for dealing with missing values, and thoughts about if/when missing values can affect the study results, are based on the following question:

What is the reason that values are missing (for each variable), and is this reason independent of the (unknown) actual value of the variable (i.e. missing randomly) or is there a systematic reason that it is missing (and you can model this reason for missingness)?

Types (reasons, mechanisms) of missing values

Overview of MCAR, MAR, MNAR

MCAR

Whether the value of a variable is missing or not does neither depend on the value itself nor does it depend on any observed or unobserved factors.

MAR

Whether the value of a variable is missing or not only depends on observed factors.

MNAR

Whether the value of a variable is missing or not only depends on unobserved factors.

$\longrightarrow$ MCAR, MAR: "ignorable" (model of missingness mechanism, not missingness per se), "non-informative missingness"
$\longrightarrow$ MNAR: "non-ignorable", "informative missingness"

Example of MCAR, MAR, MNAR

Study: Measure blood pressure before and after a 2-week long intervention including physical activity.

MCAR

At the second measurement, only 50% of the participants were invited, randomly selected.

MAR

At the second measurement, only those participants whose blood pressure at the first examination was > 130 were invited.

MNAR

At the second measurement, only the blood pressure values > 130 were recorded.

MCAR - definition

The (not observed) value of a variable is **missing completely at random (MCAR)**, if the probability that the value is missing (i.e. whether the value is missing or not) neither depends on the value of the variable itself nor on other observed or unobserved data.

MCAR - examples

- At the second blood pressure measurement, only 50% of the participants were invited, randomly selected.
- When measuring biomarker levels, there are errors in 2% of the measurements due to a malfunctioning of the pipetting machine, which occur randomly and don't depend on the characteristics of the samples or biomarker levels.
- In a representative study to assess drug consumption on a population level, in the questionnaire that is sent out to a random sample drawn from the registration office, 80% don't respond because they just don't want to or don't have time - but this is unrelated to their drug consumption and to other variables.

MAR - definition

The (not observed) value of a variable is **missing at random (MAR)**, if the probability that the value is missing (i.e. whether the value is missing or not) doesn't depend on the value of the variable itself and also doesn't depend on unmeasured variables, but depends on the values of other observed variables (and can be explained by them).

$\longrightarrow$ MAR $\neq$ randomly missing!

MAR - examples

- At the second measurement, only those participants whose blood pressure at the first examination was > 130 were invited.
- When measuring biomarker levels, there are errors in 2% of the measurements due to a malfunctioning of the pipetting machine, which depend on the outside temperature or whether the sample was contaminated (and both these variables are recorded) The errors are not associated with the biomarker level itself.
- In a representative study to assess drug consumption on a population level, in the questionnaire that is sent out to a random sample drawn from the registration office, 80% don't respond because they just don't want to or don't have time - which depends on age and sex (young men have lowest response) but this is unrelated to their drug consumption and to other unmeasured variables.

MNAR - definition

The (not observed) value of a variable is **missing not at random (MNAR)**, if the probability that the value is missing (i.e. whether the value is missing or not) depends on unmeasured variables (e.g. on the value of the variable itself) and also cannot be explained by other observed variables.

MNAR - examples

- At the second measurement, only the blood pressure values $>$ 130 were recorded.
- When measuring biomarker levels, levels outside of the detection limit cannot be assessed.
- In a representative study to assess drug consumption on a population level, in the questionnaire that is sent out to a random sample drawn from the registration office, 80% don't respond because they just don't want to or don't have time - which depends on age and sex (young men have lowest response). Of those people that take drugs, however, 99% don't respond.

Simulation of missing values in sleep & wellbeing study

- Aim: Investigate the association of sleep and wellbeing.
- See `R_7b_missing_values_simulation_study.Rmd`

Simulation of MCAR data

Delete values of psychological wellbeing (10%, 50% or 90%) randomly.

Simulation of MAR data

Delete values of psychological wellbeing (10%, 50% or 90%) depending on the values of physical wellbeing.

Simulation of MNAR data

Delete values of psychological wellbeing (10%, 50% or 90%) depending on the values of psychological wellbeing.

Analysis of missing values

Challenges for the analysis

1. Identify missing values in the dataset

- Coding of values (NA, <NA>, ".", 7, 99 ...) $\longrightarrow$ data dictionary
- Is the dataset a raw dataset or has it already been somehow corrected/processed and missing values have been transformed in some form (deleted, corrected, imputed)?
- Identify "missing" values through data quality control checks ("wrong values")

Challenges for the analysis

2. Identify the type and pattern of missing values

- MCAR?
- Missing values because divided by 0, or read in data incorrectly?
- Systematically missing (MAR or MNAR)? $\longrightarrow$ knowingly systematic, unknowingly systematic, does it affect the analysis of the study question?

3. Choose how to deal with missing values

Overview: Methods for dealing with missing values

- Correct/replace missing values, if the values can be determined from external information
- No imputation:
  1. Complete-case analysis (row-wise deletion)
  2. Pairwise deletion
- Simple imputation: impute with a single value
  1. Mean imputation
  2. Median imputation
  3. Regression-based imputation
  4. ...
- More complex and recommended approaches
  1. Model-based: Full Information Maximum Likelihood (FIML)
  2. Data-based: Multiple Imputation (MI)
  3. ...

Evaluation of all methods in the sleep & wellbeing example

- For illustration, apply all methods above on the simulated sleep & wellbeing data.
- Question: if/when do the approaches yield biased estimates and wider confidence intervals/tests with lower power (i.e. larger p-values)?
- See `R_7b_missing_values_simulation_study.Rmd`.

Baseline results when analyzing the complete data (ground truth):

n	r	95% CI	p-value
7559	0.10	(0.08, 0.13)	1.9×10^{-19}

No imputation: Complete-case analysis

Approach

Exclude all missing values by row-wise deletion, i.e. only keep those observations that don't have any missing values in all variables used in the analysis.

Complete-case analysis

Evaluation in the sleep & wellbeing example (see `R_7b_missing_values_simulation_study.Rmd`):

Scenario	n	r	95% CI	p-value
MCAR 10%	6804	0.11	(0.09, 0.13)	9.1×10^{-20}
MCAR 50%	3780	0.10	(0.07, 0.13)	6.8×10^{-10}
MCAR 90%	756	0.10	(0.03, 0.17)	4.9×10^{-3}
MAR 10%	6659	0.10	(0.08, 0.12)	4.6×10^{-16}
MAR 50%	3940	0.12	(0.09, 0.15)	7.6×10^{-15}
MAR 90%	728	0.17	(0.09, 0.24)	6.6×10^{-6}
MNAR 10%	6956	0.06	(0.03, 0.08)	1.6×10^{-6}
MNAR 50%	3808	0.03	(0.00, 0.06)	6.7×10^{-1}
MNAR 90%	636	NA	(NA, NA)	NA

Complete-case analysis

Advantages and disadvantages

- Simple, doesn't need any special statistical methods.
- Only uses the existing data for the analysis.
- Unbiased estimates under MCAR.
- Biased estimates under MAR (here: partially only slightly biased) and MNAR.
- Loss of statistical power, wider confidence intervals since more missing values.
- When analyzing multiple variables with missing values, the sample size can quickly decrease (see introductory example).

Simple mean imputation

Advantages and disadvantages

- Simple, doesn't need complex statistical methods.
- All follow-up analysis will use the full sample so that the estimates will not have larger standard errors because of missing values.
- The mean of all variables stays the same.
- Mean imputation can already lead to biased estimates under MCAR.
- Mean imputation can yield more strongly biased estimates compared to complete-case analysis.

Conclusion

Don't use it (for estimation & testing), if you are interested in an unbiased estimator.

Full Information Maximum Likelihood (FIML)

- In some situations, FIML is better suited than Multiple Imputation, but they are asymptotically equivalent.
- For more details and references, see Graham (2012).

Multiple Imputation

Approach

1. Imputation: Generate several (eg $m = 5$) imputed datasets, in which the missing values are each replaced by "plausible" values.
2. Analysis: Analyze all m imputed datasets separately with "normal" statistical methods.
3. Pooling: Combine the parameter estimates and standard errors from the separate analyses using "Rubin's rules" (Rubin, 1987), and do e.g. Wald (hypothesis) tests.

Multiple Imputation

Implementations in R

- mice (Multivariate Imputation by Chained Equations):
  - `https://www.doi.org/10.18637/jss.v045.i03`
  - `https://cran.r-project.org/web/packages/mice/index.html`
  - Predicts missing values of each variable based on (potentially) all other variables (models joined distribution by chained marginals using MCMC).
- missForest (Nonparametric Missing Value Imputation using Random Forest):
  - `https://doi.org/10.1093/bioinformatics/btr597`
  - `https://cran.r-project.org/web/packages/missForest/index.html`
  - Imputation using random forests (nonparametric).
- Hmisc (Harrell Miscellaneous)
  - `https://cran.r-project.org/web/packages/Hmisc/index.html`
  - Imputation using bootstrapping and additive models.
- mi (Multiple imputation with diagnostics):
  - `https://doi.org/10.7916/D8VQ3CD3`
  - `https://cran.r-project.org/web/packages/mi/index.html`
  - More diagnostics, imputation in Bayesian framework.

Downloads: 1.5M, 220k, 6.6M, 750k (`https://cranlogs.r-pkg.org/badges/grand-total/mice`)

Multiple Imputation using `mice`

- `md.pattern`: Function to show patterns of missing values.
- (`aggr` Function in `VIM` package: nice visualization to show patterns of missing values).
- `mice`: Function for generating multiple imputed datasets using the mice algorithm. Default is "pmm" (predictive mean modeling, Little 1988), a semiparametríc method. Another implemented method is "norm", which is faster and imputes based on Bayesian regression models.
- The statistical analysis can be applied to the imputed datasets using the `with` function.
- The results can be pooled using the `pool` function.

Multiple Imputation

Evaluation in the sleep & wellbeing example using the default "pmm" method (see `R_7b_missing_values_simulation_study.Rmd`):

Scenario	n	r	95% CI	p-value
MCAR 10%	7559	0.11	(0.08, 0.13)	1.3×10^{-17}
MCAR 50%	7559	0.10	(0.06, 0.14)	1.4×10^{-6}
MCAR 90%	7559	0.10	(0.05, 0.15)	1.2×10^{-4}
MAR 10%	7559	0.10	(0.07, 0.12)	1.5×10^{-13}
MAR 50%	7559	0.10	(0.05, 0.14)	7.7×10^{-5}
MAR 90%	7559	0.04	(0.00, 0.08)	6.8×10^{-1}
MNAR 10%	7559	0.06	(0.04, 0.09)	1.7×10^{-7}
MNAR 50%	7559	0.04	(0.01, 0.06)	4.7×10^{-3}
MNAR 90%	7559	NA	(NA, NA)	NA

Multiple Imputation

Advantages and disadvantages

- Multiple imputation is best suited for analyses that are based on parameter estimates and standard errors. Other models (e.g. ANOVA) can be reformulated for this.
- The imputed datasets can be analyzed using standard statistical methods.
- You can use the imputed datasets for different analyses.
- Imputation and estimation/inference are independent.
- Is a more complex procedure compared to simple imputation or complete-case analysis, but is implemented for many standard statistical models.
- Considers the randomness of observations.

Multiple Imputation

Advantages and disadvantages

- Default "pmm" Methode is relatively slow, especially for many missing values, e.g. "norm" is much faster.
- In theory: unbiased under MAR, but this is not always the case in finite samples.
- Also in theory: always at least as good as simple imputation and complete-case analysis, which is also almost always the case in the sleep & wellbeing simulated datasets here (Grahahm 2009).

Multiple Imputation

Practical recommendations

- Use as many variables as possible for the imputation (also the outcome!).
- At least, use all variables in the imputation model that are in the analysis model (and possible more) (Graham 2012).
- When a variable is not considered in the imputation, this contains the assumption that it does not play a role in the analysis model (and can thereby create bias if this is not true).
- The imputation model should be at least as complex as the analysis model, i.e. if interactions will be analyzed later, then interactions should be incorporated in the imputation model (Graham 2012).

Exercise

1. Use `mice` and multiple imputation to investigate the association of sleep and psychological wellbeing in children (external rating by parents) in the KiGGS dataset. Is the estimated association different from the complete-case analysis?
2. Use `mice` and multiple imputation to investigate the association of sleep and psychological wellbeing in children (self rating by the children). Is the estimated association between sleep and psychological wellbeing different if you use the self-rating compared to the external rating by the parents?
3. Investigate the association between sleep and physical wellbeing, using different ways how to deal with missing values.
4. See `R_7c_missing_values_exercise.Rmd`.

Conclusions, overview, exercises

Conclusions 1

- Recommendation: complete-case analysis or multiple imputation.
- Don't use simple imputation methods! They can yield biased estimates also under MCAR.
- Under MCAR, both complete-case analysis and multiple imputation "should be" valid. Under MAR, multiple imputation "should be" better. Under MNAR, both can be invalid. That means: under MAR and MNAR, both methods can have bias, and can also not have bias.
- Further discussions: `https://doi.org/10.1093/ije/dyz032`
- Application: `https://doi.org/10.1093/ije/dyy267`

What do you do for data that is MNAR?

You have to model the missingness mechanism explicitely, e.g. using:

- Reweighting (eg Rezvan et al. 2015)
- Selection models (eg Little 1993)
- Pattern-mixture models (eg Little 1993)

Insert: Missing values = censored?

Aim

Analysis of (systolic, diastolic) blood pressure.

Challenge

Some participants take blood pressure-reducing drugs, i.e. the "true" blood pressure (and the effect of the drug) is unknown and not measured.

Methods

Before the analysis, model the effect of the blood pressure-reducing drugs, e.g. using a censored regression (Tobin et al. 2005, Konigorski et al. 2014). Use the knowledge/assumption that the true unobserved blood pressure is higher than the observed blood pressure.

Conclusions 2

- Do a detailed analysis of non-responders!
- Before planning and conducting your study, think about assessing predictors of missingness.
- Look at more detailed categorizations under which scenarios missing values can affect your results, and check if these are true in your analysis and sample.

Application example: imputation in genetics

- In genetics, specific imputation algorithms have been developed for the imputation of genetic variants that have not been measured experimentally.
- This is nowadays standard procedure so that microarray chips for measuring DNA variation have been specifically designed for this, e.g. to measure 500,000 genetic variants from which tens of millions of genetic variants can be imputed and then analyzed.
- The background for this is that the correlation between genetic markers has been investigated in large studies and specific statistical imputation algorithms have been developed that model the joint distribution of the observed and not-measured (missing) genetic variants.
- If you are interested in more information, see e.g. Li et al. (2009).

Questions?

References

- Graham JW (2012). Missing Data. Analysis and Design. Springer.
- Rubin DB (1987). Multiple imputation for nonresponse in surveys. Wiley.
- van Buuren S, Groothuis-Oudshoorn K (2011). mice: Multivariate Imputation by Chained Equations in R. J Stat Softw 45(3):1-67.
- Gelman A, Hill J (2011). Opening Windows to the Black Box. J Stat Softw 40.
- Stekhoven DJ, Buehlmann P (2012). MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
- Little RJA (1988). Missing data adjustments in large surveys. J Bus Econ Stat 6, 287-301.

References

- Tobin MD et al. (2005) Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Stat Med 24, 2911-35.
- Konigorski et al. (2014). Bivariate genetic association analysis of systolic and diastolic blood pressure by copula models. BMC Proc 8(Suppl 1):S72.
- Graham JW (2009). Missing data analysis: making it work in the real world. Annu Rev Psychol 60, 549–576.
- Li Y, et al. (2009). Genotype Imputation. Annu Rev Genomics Hum Genet 10, 387–406.
- Little RJA (1993). Pattern-mixture models for multivariate incomplete data. JASA 88, 125–34.
- Hayati Rezvan P et al. (2015). Evaluation of a weighting approach for performing sensitivity analysis after multiple imputation. BMC Med Res Methodol 15, 83.

Homework

Homework

See file `R_7_homework.Rmd`.