R101: A practical guide to making R your everyday statistical tool (PSY532)

Lecture 5

Handling missing data

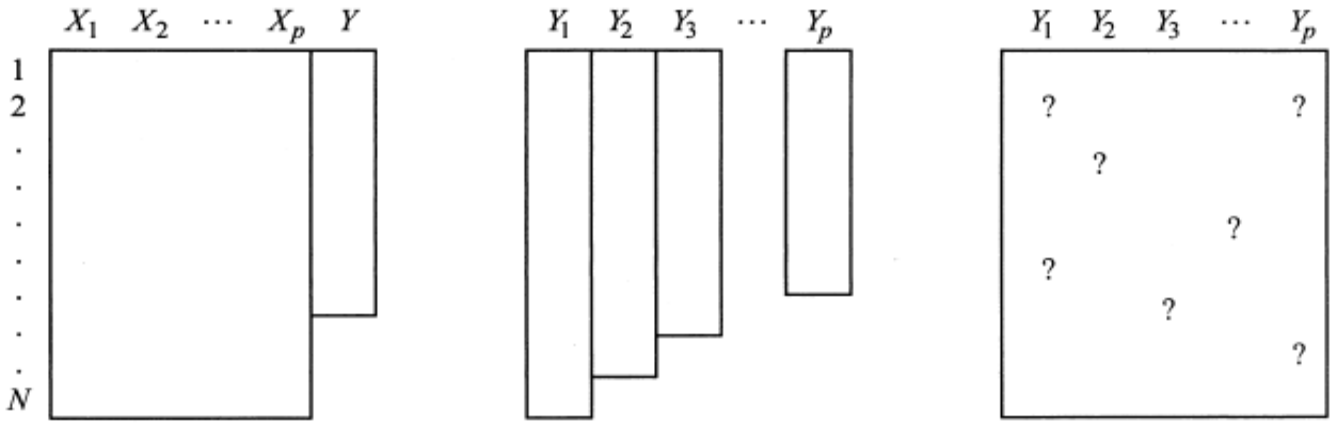
Programme (lecture and seminar)

- Types of missingness: MCAR, MAR and MNAR
  - Little's MCAR test
- Older (but still common) approaches to missing data
- EM *single* imputation: an improved and commonly used approach available in SPSS
- The advantages of *multiple* imputation
- Multiple imputation options in R:
  - `norm` package (and its problems)
  - `mice` package (the most commonly used option)
- Planned missingness: designing a study in which each participant answers only a subset of survey questions
  - Multiple imputation can then be used on the data
- Structural equation modelling packages often have their own Full Information Maximum Likelihood (FIML) algorithm for taking missing data into account when calculating model parameters. We do not discuss these algorithms here, but they are covered in the readings.

Non-response patterns and types of missingness

Schafer & Graham 2002

Causes of missingness (CoM): Being physically unable to show up, not being sure, concerns about privacy, etc.



	Univariate	Monotone	Arbitrary
MCAR	CoM not related to X or Y	CoM not related to any Y	CoM not related to any Y
MAR	CoM related to X (and possibly Y, but not once X is taken into account)	CoM related to one or more Ys prior to drop-out	Possibly different set of Ys related to CoM for each participant 1 to N
MNAR	CoM related to Y and possibly X	CoM related to one or more Ys after drop-out (unseen responses)	CoM is related to an unobserved variable, which is related to one or more Ys

- MCAR, pure MAR, and pure MNAR really never exist. Nevertheless, they are useful guidelines for thinking about the sources and consequences of missing data.
- A number of tests have been proposed for MCAR (see script):
  - Little's MCAR test, using the `LittleMCAR` function in the `BaylorEdPsych` package (suitable for datasets with up to 50 variables)
  - a visual approach, using the `marginplot` function in the `VIM` package. This approach is also useful for detecting whether certain variables in the data set predict patterns of missingness, implying MAR.
  - Hawkins test of MCAR available in the `MissMech` package; not covered here
- The multiple imputation methods in this lecture assume that data is not MNAR (i.e., it has to pass the test for MCAR or come from a study where the study design ensures against MNAR)
- The multiple imputation methods in the lecture are applied to a univariate non-response pattern, but are transferable to the monotonic and arbitrary patterns.

In the script, we conduct Little's MCAR test on a modified version of the SS data set. The version has a random 5% of data removed for all variables except the experimental condition and pre-experimental measures. Data is missing for the illusion of control measures, "It was all chance", the question about strategy (yes/no), the question about the number of wins in the next 100 rounds, etc. See lines 31-42.

Go to script

Older (but still common) approaches

Reading: All references marked *

	Description	Problems
Available case analysis (listwise deletion)	Delete all participants who are missing values on any variable included in an analysis. Default option when running ANOVAs, regressions etc. in SPSS.	• Possible only with MCAR, or with MAR if the analysis includes the variable relating to the CoM as a covariate • Reduces power of the analyses • *N*s are different across analyses
Averaging the available responses	If a person is missing some but not all responses on a scale that has a mean score, calculate that mean based only on the available items (e.g., the average of 6 items rather than 8)	• May introduce bias under MCAR by adding additional variability to the survey • Conceptually problematic if some items are more likely to be missing (i.e., MAR or MNAR), yet the survey score is redefined as the average of available items rather than defined items

	Description	Problems
Mean substitution	Substitute each missing value with the average of other participants' values on the item	The average of the variable is preserved, but other aspects of its distribution are altered: variance, quantiles, and correlation with the responses of other participants.
Hot deck imputation	Substitute each missing value with a value randomly drawn from other participants' values	Correlations in responses across participants increase.
Regression-based single imputation	First, divide participants into those with a variable Y, and those for whom Y is missing. Then, estimate a regression model in the first group (X_1 , X_2 ... X_p predicting Y) and calculate Y for the second group based on that regression model. Available in SPSS.	• The imputed data points do not depart from the regression line, which makes them different (less variable) to the observed data points. SPSS adds some error to each data point to partially correct for this. • Unless the Xs and Y are strongly related, the relationship between them is inflated.

EM single imputation

Reading: All except #

EM single imputation is currently one of the most common approaches to missing data because of its easy implementation in SPSS.

E-step

Regression-based single imputation: First, divide participants into those with a variable Y, and those for whom Y is missing. Then, estimate a regression model in the first group (X_1 , X_2 ... X_p predicting Y) and calculate Y for the second group (the "missingness" group) based on the regression equation.

M-step

Regression models can also be fitted based on means and estimates of "variance" and "covariance". These can be calculated from a data set. At the M-step, these are calculated from the filled-in data set of the E-step, resulting in a new regression model. There is then another E-step to calculate Y in the "missingness" group.

The steps are repeated until they produce the same estimates for Y in the "missingness" group (i.e., "converge").

Advantages of *multiple* imputation

Reading: All except #. In particular, Graham (2009).

What is it?

1. Use EM-based or regression-based procedures to generate a number of imputed data sets. Each data set might contain a different imputed value for each missing value. Researchers tend to generate anywhere between 5 and 50 data sets, and it has been recommended that the number correspond to the percentage of missing cases.
2. Conduct your planned analysis (ANOVA, regression, generalized linear model, etc.) on each data set from Step 1.
3. "MI inference": Combine the results of the analyses into a pooled result using "Rubin's rules" (e.g., one rule is that point estimates for the parameters are simply averaged across analyses).

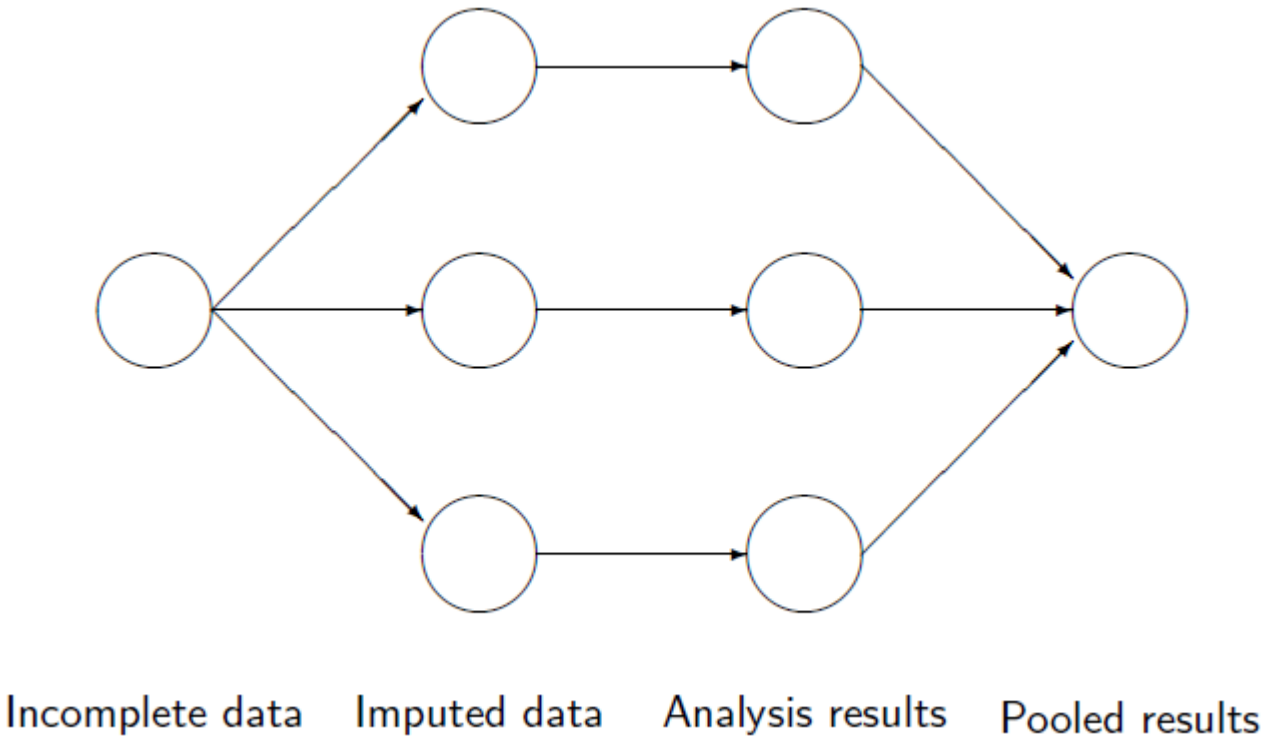
Its advantages over single EM

- Like single-imputation EM, these procedures overcome the problem of lack of variance in regression-based single imputation.
- Additionally, they overcome the problem that, in both EM and regression-based single imputation, the regression model is based on a single sample.

Cautionary note

Include as many variables without missing values as possible in the procedure to avoid predicting missing values based only on the variables used as predictors in the analyses at Step 2.

A diagram of the steps in multiple imputation



From publicly available course material by Stef van Buuren, author of the `mi` package

Multiple imputation options in R

`norm` package: Multiple imputation through data augmentation

Reading: All references marked *da*

- Multiple imputation procedure based on the single-imputation EM approach.
- As we shall see in applying norm to our modified *SS* data set with 5% of cases missing, the procedure can impute negative values where the observed values are only positive. This is a problem if the analyses in Step 2 of the multiple imputation process involve generalized linear modelling with Poisson or negative binomial random components. These are not suitable for outcome variables with negative values.
- If there are categorical variables among those missing, the `mix` package needs to be used, for which you will need the package author's, J. S. Schafer's, book "Analysis of Incomplete Multivariate Data" (1997; Chapters 7-9).
- The online manual for the package is also very useful.

Go to script

`mice` package: Multiple imputation through chained equations

Reading: All references marked *ce*

Depending on whether the scale of the variable with missing values is numeric or categorical (factor-based), this procedure uses one of the following regression methods to impute missing values. The pmm and logreg methods are defaults for numeric and categorical variables respectively. Most of the regression methods are Bayesian and all the other variables in the data set (even those with missing values) act as predictors.

Method	Description	Scale type
pmm	Predictive mean matching	numeric*
norm	Bayesian linear regression	numeric
norm.nob	Linear regression, non-Bayesian	numeric
norm.boot	Linear regression with bootstrap	numeric
mean	Unconditional mean imputation	numeric
2L.norm	Two-level linear model	numeric
logreg	Logistic regression	factor, 2 levels*
logreg.boot	Logistic regression with bootstrap	factor, 2 levels
polyreg	Multinomial logit model	factor, $>$ 2 levels*
polr	Ordered logit model	ordered, $>$ 2 levels*
lda	Linear discriminant analysis	factor
sample	Simple random sample	any

Go to script

Reading

Azur, M. J., Stuart, E. A., Frangakis, C., Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*, 40-49. Will be made available online as course material. *ce*

Graham, J. W. (2012). *Missing Data: Analysis and Design*. Springer Science+Business Media: New York. Chapter 2 "Analysis of missing data". Available as "sample pages" online here: http://www.springer.com/statistics/social+sciences+%26+law/book/978-1-4614-4017-8 * *da*

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576. Will be made available online as course material. * *da*

Graham, J. W., Cumsille, P. E. and Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of Psychology.* Volume One, 87–114. Available as a html file through a Masaryk University computer: http://onlinelibrary.wiley.com/doi/10.1002/0471264385.wei0204/full * *da*

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, 7, 199–204. Will be made available online as course material. #

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods,7, 147–177. Will be made available online as course material. * *da*

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45. *ce*