RHUL Psychology Statistical modelling notebook

Matteo Lisi

2022-04-07

Contents

Chapter 1

About

This online book is created and maintained by Matteo Lisi and is meant to be a shared resource for staff and students at the Department of Psychology of Royal Holloway, University of London. It will contain a miscellanous set of tutorial, examples, case studies, workshops materials and any other useful material related to data analysis and modelling. These will be added and revised over time, based on the most common questions and requests that I receive.

This is a work in progress and may contain imprecisions and typos. If you spot any please let me know at matteo.lisi [at] rhul.ac.uk. The materials that will be included builds upon and draw from existing literature on statistics and modelling. I will endeavor to properly cite existing books and papers; but if any author feels that I have not given them fair acknowledgement, please let me know and I will make amend.

Chapter 2

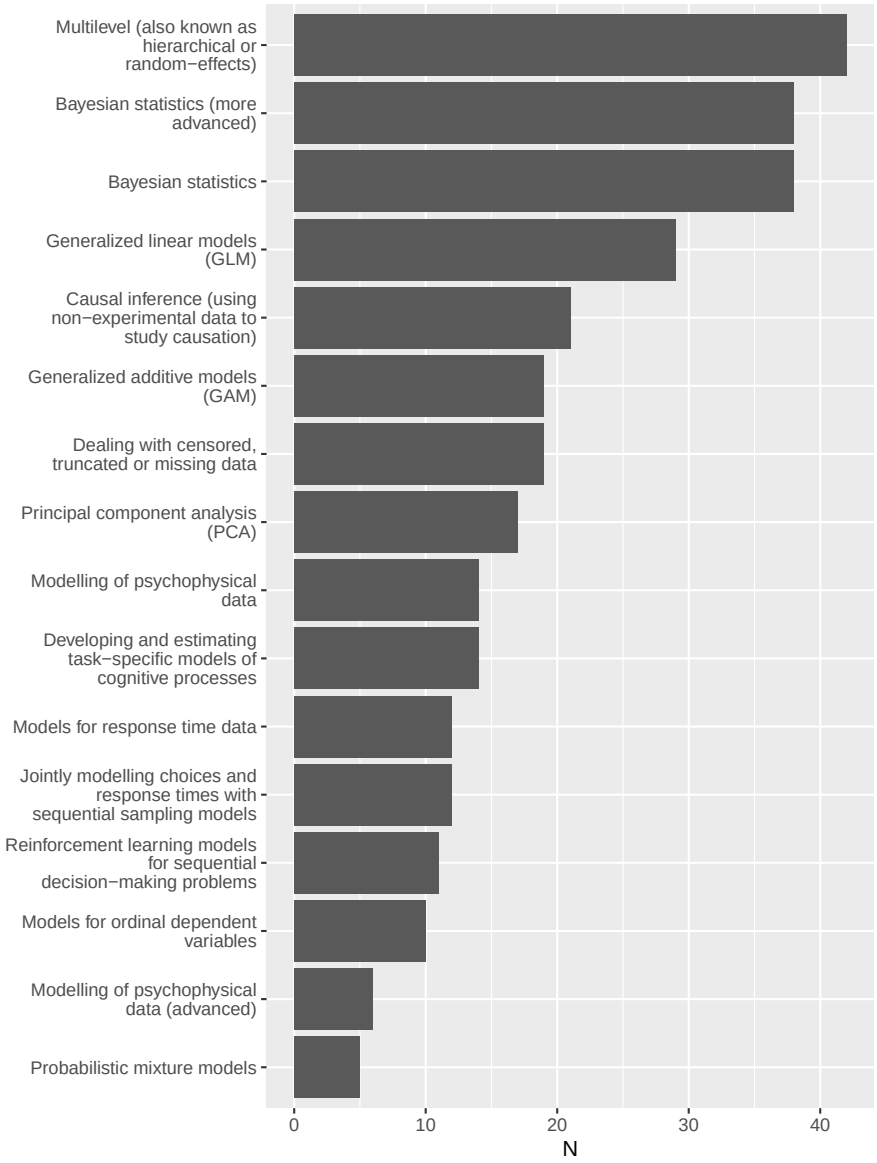
Departmental survey about statistical methods

I used an anonymous survey to ask colleagues some questions about which topics may be more interesting or useful in their research.

2.1 March 2022

2.1.1 Question 1

In the first question people indicated topics of interests. The winner are multi-level models, followed closely by Bayesian statistics.



There were some additional suggestions.

```
#> [1] "power analyses using Shiny apps"
#> [2] "agent-based models"
#> [3] "this may be covered in the above, but approaches to analysing experience sampli
#> [4] "Methods for longitudinal analyses"
#> [5] "Network modelling"
#> [6] "Neural networks, Markov processes"
```

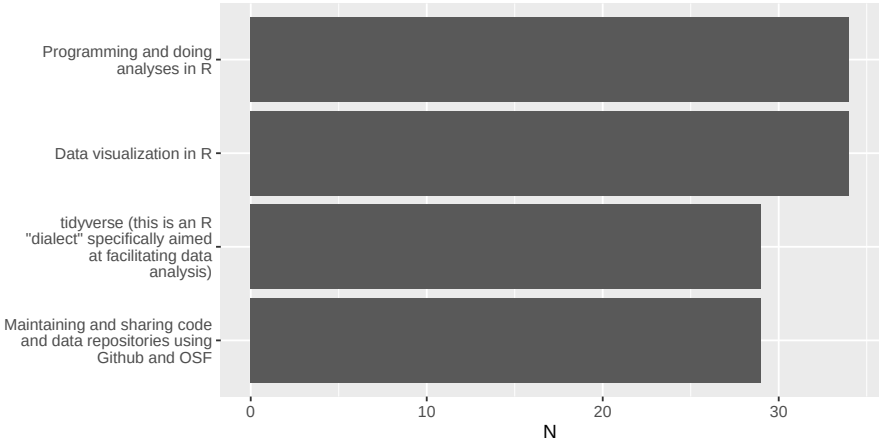
```
#> [7] "Random forests and related"
#> [8] "causal modelling using regression models - path models etc"
#> [9] "prediction modelling"
```

A few other topics were mentioned in the comment section:

- Shiny apps
- Network modelling
- Longitudinal analyses
- Random forests
- Neural network

2.1.2 Question 2

Here people indicated their interest for topics related to data analysis.

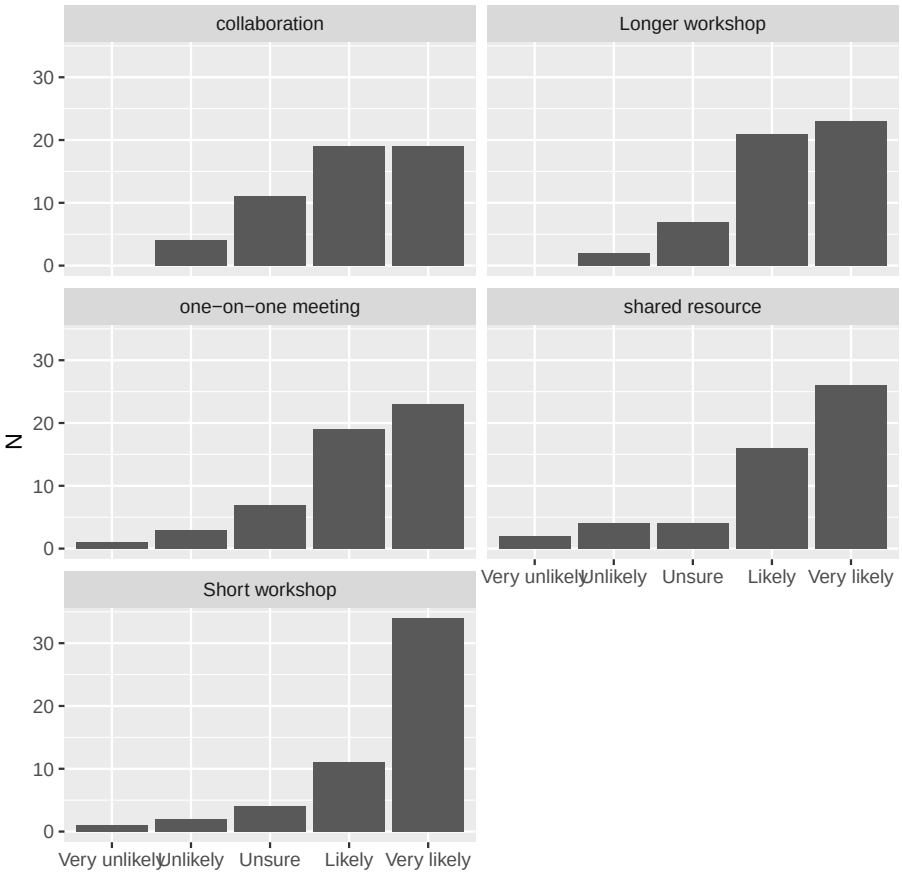


Other things mentioned in the comments were:

- SPM
- Docker
- Python

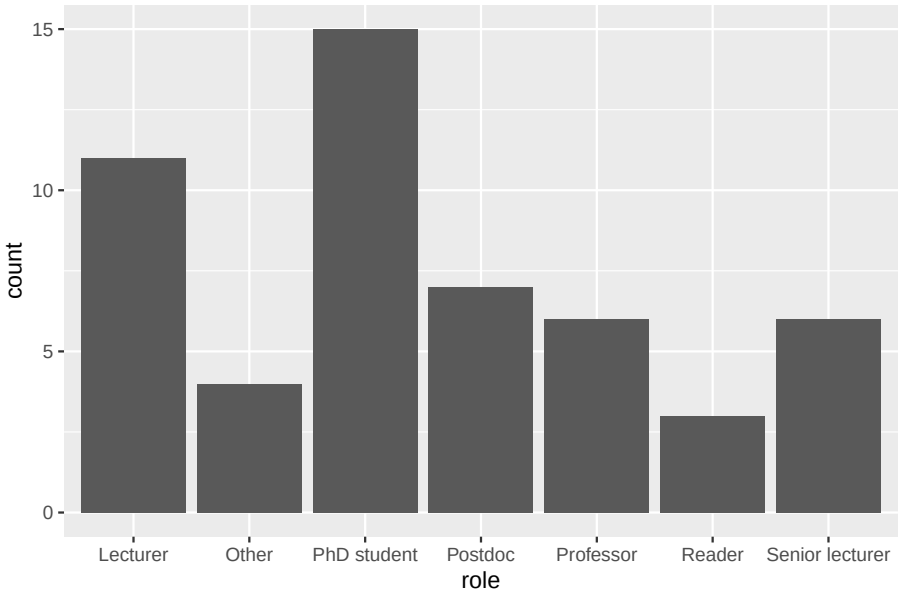
2.1.3 Question 4

This question was about likelihood of using different formats of support



2.1.4 Respondents' status

The final questions asked about the status / career level.


15
10
5
0
count
Lecturer
Other
PhD student
Postdoc
Professor
Reader
Senior lecturer
role

Chapter 3

Meta-analyses

For running meta-analyses, we recommend the `metafor` package for R (see link 1, link 2).

A comprehensive, hands-on guide on how to use this package is provided in the book by Harrer and colleagues (Harrer et al., 2021), freely available at this link.

An alternative to the `metafor` package is to Bayesian multilevel modelling (also discussed in the book by Harrer and colleagues). A more technical discussion of Bayesian multilevel modelling for meta-analyes is provided in this paper by Williams, Rast and Bürkner (Williams et al., 2018).

Chapter 4

Missing data

4.1 Types of missing data

Following the work of Rubin(RUBIN, 1976), missing data are typically grouped in 3 categories:

- Missing completely at random (**MCAR**). This assumes that the probability of being missing is the same for all cases; this implies that the mechanisms that causes missingness is not related in any way to the data. For example, say, there's a known unpredictable error on the server side that prevented recording some responses for some respondents to a survey. As the missingness is entirely independent on the respondents' characteristics, this would be MCAR. When the data are MCAR we can ignore a lot of the complexities and just do a *complete-case* analysis (that is, simply exclude incomplete observations from the dataset). A part from possible loss of information, doing a complete case analysis should not introduce bias in the results. In practice, however, it is difficult to establish whether the data are truly MCAR. Ideally, to argue that data are MCAR, one should have a good idea of the mechanisms that caused missingness (more on this below). Formally, data is MCAR if

  $$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = 0|\psi)$$

  where R is an indicator variable that is set to 0 for missing data and 1 otherwise; $Y_{\text{obs}}, Y_{\text{mis}}$ indicate observed and missing data, respectively; and ψ is simply a parameter that determine the overall (fixed) probability of being missing.

- Missing at random (**MAR**). A less strong assumption about missingness is that it is systematically related to the observed but not the unobserved data. For example, data are MAR if in a study male respondents are less likely to complete a survey on depression severity than female respondents

- that is, the probability of reaching the end of the survey is related to their sex (fully observed) but not the severity of their symptoms. Formally, data is MAR if

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = 0|Y_{\text{obs}}, \psi)$$

When data are missing at random (MAR) the results of complete case analyses may be biased and a common approach to deal with this is to use imputation . Stef van Buuren has a freely available online book on this topic(van Buuren, 2018). Among other things, it illustrates how to do multiple imputation in R with examples.

- Missing not at random (**MNAR**). This means that the probability of being missing varies for reasons that are unknown to us, and may depends on the missing values themselves. Formally this means that $\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi)$ does not simplify in any way. This case is the most hard to handle: a complete case analyses may or may not be biased, but there is no straightforward way to find out and we may have to find more information about what caused missingness.

4.2 Deciding whether the data are MCAR

As MCAR is the only scenario in which it is safe to do a complete case analysis, it would seem useful to have way to test this assumption. Some approaches have been proposed to test whether the data are MCAR, but they are not widely used and it's not clear how useful they are in practice. For example one could run a logistic regression with "missingness" as dependent variable (e.g. an indicator variable set to 1 if data is missing and 0 otherwise), and all other variables as predictors - if the data are MCAR then none of the predictors should predict missingness. A popular alternative, implemented in several software packages is Little's test(Little, 1988).

Technically these approaches can help determine whether the missingness depends on some observed variables (that is, if they are MAR), but strictly speaking cannot exclude missingness due to unobserved variables (MNAR scenario). Nevertheless, if one has good reasons to believe that the data are MCAR, and want to add some statistical test that corroborate this assumption, these could be reasonable tests to do.

However, statistical tests alone cannot tell whether data are missing completely at random. The terms MCAR, MAR and MNAR refers to the *causal* mechanisms that is responsible for missing data and, strictly speaking, causal claims cannot be decided uniquely on the basis of a simple statistical test. If the data "pass" the test it would provide some additional support to the assumption that they are MCAR, but in and of itself the test alone does not fully satisfy the assumptions of MCAR. Note that MCAR (as formally defined above) assumes also that there should be no relationship between the missingness on a particular variable and the values of that same variable: but since this is a question

about what is missing from the data, it cannot be tested with any quantitative analysis of the available data.

Furthermore, since these are null-hypothesis significance test, a failure to reject the null hypothesis does not in itself provide evidence for the null hypothesis (that the data are MCAR). It may be also that we don't have enough power to detect the pattern in the missingness. Thus, if we think the data are MCAR it is important to discuss openly possible reasons and mechanisms of missingness, and explain why it is plausible that the data are MCAR.

4.3 Causal analysis and Bayesian imputation

The best and most principled approach to deal with missingness (at least in my opinion) is to think hard about the causal mechanisms that may determine missingness, and use our assumption about the causal mechanisms to perform a full Bayesian imputation (that is , treating the missing data as parameter and estimating them).

I plan to creat and include here a worked example of how to do this; in the meantime interested readers are referred to Chapter 15 (in particular section 15.2) of the excellent book by Richard McElreath *Statistical Rethinking*(McElreath, 2020) which present a very accessible worked example of how to do this in R.

Bibliography

Harrer, M., Cuijpers, P., A, F. T., and Ebert, D. D. (2021). *Doing Meta-Analysis With R: A Hands-On Guide*. Chapman & Hall/CRC Press, Boca Raton, FL and London, 1st edition.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202.

McElreath, R. (2020). *Statistical Rethinking, A Course in R and Stan*. Chapman & Hall/CRC Press, 2nd edition.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman & Hall/CRC Press, 2nd edition.

Williams, D. R., Rast, P., and Bürkner, P. C. (2018). Bayesian meta-analysis with weakly informative prior distributions.