MAS61006 Assessed Project

Semester 2, 2023

This project counts for 40% of the assessment for MAS61006.

1 Aim

In this project, the aim is to design and implement an experiment to investigate the performance of the MICE algorithm and multiple imputation approach as described in Chapter 5 of your notes. You will use the `mice` package in R to do this. You will write up your experiment and results in a short report.

2 Tasks

Your experiment should be conducted as follows.

2.1 Selecting a data set

You must find your own data set for this experiment.

- You should not use any data set that is described in the lecture notes or included in the `mice` package. You can use another data set available in R, if you wish.
- Your data set should be suitable for linear or logistic regression modelling.
- Aim for somewhere between 5-15 variables, with at least 100 complete case observations.

2.2 Fitting a benchmark model

- If your data set has any missing data, select the complete cases only using the command `na.omit()`
- Choose a dependent variable from your data set, and fit a linear model or logistic regression model as appropriate. All other variables should be included (assuming no problems with collinearity).
- Obtain the parameter estimates and standard errors for your model parameters.
- You do not need to consider any model selection, or analysis of your fitted model. The aim is simply to obtain a benchmark for comparison later on.

2.3 Modifying the data set to have missing data

At random, select elements of your data set to be missing (`NA`). You may use the following code to do this. You should experiment with different values of `probMissing`.

```r
makeMissing <- function(mydf, probMissing){

  # mydf: your data frame
  # probMissing: the probability that any single
  #  element of the data frame will be changed to NA

  R <- matrix(rbinom(nrow(mydf) * ncol(mydf),
                     1,
                     probMissing),
              nrow = nrow(mydf),
              ncol = ncol(mydf))
  mydf[R == 1] <- NA
  mydf
}
```

2.4 Comparing a complete case analysis with MICE

- Fit the same model as before, using two approaches
    1. A complete case analysis
    2. Multiple imputation with the MICE algorithm
- Obtain parameter estimates and standard errors in each case.
- Experiment with at least two of the "built-in" imputation methods in the `mice` function: see `?mice::mice` for details.

2.5 Analysing the results

Compare parameter estimates and standard errors for your chosen model using

1. the full data set (no missing data);
2. the modified data set and complete case analysis;
3. the modified data set and multiple imputation with MICE.

You do not need to consider anything apart from changes in parameter estimates and standard errors.

2.6 Scope of your experiment

You can experiment with different imputation methods and different proportions of missing data. You can also consider whether any results are robust to the random selection of missing values.

There is a hard page limit of 6 pages. Credit will be given for *concise* presentation of your results: presenting a clear comparison of different methods/scenarios with the smallest number of plots/tables possible.

3 The written report

- Your report must be prepared using R Markdown. Use the following YAML header, inserting your own registration number in the author field.

```
---
title: "MAS61006 Assessed Project"
author: "your student registration number"
date: "April 2023"
output: pdf_document
fontsize: 11pt
---
```

- You may import any LaTeX packages you wish, but you should not change the font size or margins.
- There is a 6 page limit: this page limit includes **everything**.
- Your report should not contain any R code or raw R output but you should submit your `.Rmd` file alongside your PDF report.
- Your target reader for this report is another student on this module, with the exception of the Conclusions section - see the instruction in the next section.

3.1 Report structure

The report should be structured with the following sections. No other sections (introduction, summary etc.) are needed

1. The data

Briefly describe your data. State what all the variables are, and what type each variable is (e.g. continuous, binary, categorical).

2. The model

Defining your notation carefully, describe the model you will fit to the data.

3. Exploratory data analysis

Include an example of an exploratory data analysis on a *single* data set in which missing data have been introduced. Only present results that you judge to be relevant to the imputation. Use the caption test on all your plots!

4. Method

Describe the experiment that you will carry out. You can use a bullet point/numbered list to do this. There should be enough detail such that another student could carry out your experiment, with all the same choices that you have made.

5. Convergence of the MICE algorithm

Give an *example* of the checks you have done for convergence of the MICE algorithm. You do not need to show this for *every* implementation of the MICE algorithm

6. Results

Present your results and give a brief discussion of them. Aim to present your results concisely, using the smallest number of plots/tables as possible.

7. Conclusion

Write a conclusion in 'plain English' for a non-expert reader: imagine your reader is a colleague at work, who has a basic understanding of regression modelling from an undergraduate degree, but has not studied statistics at MSc level.

8. References

Include a reference to the `mice` package, and any other references as appropriate. Use the command `citation("mice")` in R for details.

4 Assessment criteria

- The grading of your project will give equal weighting to presentation and technical content.
- For the presentation we will assess
  - how clearly the report is written;
  - whether you have followed all the instructions in this project brief;
  - formatting of plots, including use of "the caption test";
  - suitability of your Conclusions section for the target reader.
- For the technical content, we will assess
  - whether you have carried out the experiment appropriately, as described in this project brief;
  - the scope of your experiment: how much investigation you do, within the constraint of a 6 page report;
  - correctness of your application of the mice algorithm, convergence checking and use of multiple implementation;
  - how you have chosen to present your results: clarity and conciseness;
  - the insight provided in your written commentary: what you explain *beyond* what can be seen in any plots/tables.

5 Unfair Means

Your project submission must be entirely your own work: do not discuss your project with anyone apart from staff teaching this module. If you haven't already done so, you must work through the tutorial on unfair means available on the MSc Statistics organisation page on Blackboard.

6 Submitting your work

Upload both your pdf and Rmd file using the Assignment Dropbox on Blackboard. Use the file names

- `MAS61006ProjectReportxxxxxxxxxx.pdf`
- `MAS61000ProjectCodexxxxxxxxx.Rmd`

replacing `xxxxxxxxx` with your student registration number.