

Class 1: An introduction to Missing Data Analysis

Andrew Parnell
andrew.parnell@mu.ie



PRESS RECORD https://andrewcparnell.github.io/mda_course

Let's get started

- ▶ Introduction from Novartis
- ▶ Timetable for the week
- ▶ Pre-requisites

How this course works

- ▶ This course lives on GitHub, which means anyone can see the slides, code, etc, and make comments on it
- ▶ The timetable document (`index.html`) provides links to all the pdf slides and practicals
- ▶ The slides and the practicals are all written in `Rmarkdown` format, which means you can load them up in Rstudio and see how everything was created
- ▶ Let me know if you spot mistakes, as these can be easily updated on the GitHub page
- ▶ There is a `mda_course.Rproj` R project file from which you should be able to run all the code

Copyright statement

All the non-GitHub materials provided in the Introduction to Missing Data Analysis course are copyright of Andrew Parnell.

This means:

- ▶ As a user (the student) you have permission (licence) to access the materials to aid and support your individual studies.
- ▶ You are not permitted to copy or distribute any materials without the relevant permission
- ▶ As faculty we may reserve the right to remove a user in the event of any possible infringement

Course format and other details

- ▶ We do two lectures in the morning (separated by a coffee break) then have a practical class where we go through some R code together
- ▶ There is also an (optional) exercise sheet with further questions for those who are particularly keen
- ▶ If you want to send me a private message use the message board and I will try to answer them as we go
- ▶ Please ask lots of questions, but **MUTE YOUR MICROPHONE** when not asking them
- ▶ Some good books:
 - ▶ *Statistical Analysis with Missing Data* by Little and Rubin
 - ▶ *Bayesian Data Analysis* by Gelman et al (chapter on Missing Data Analysis)
 - ▶ *Flexible Imputation of Missing Data* by Stef van Buuren

Why care about missing data analysis?

Statistics is a missing data problem – Rod Little

- ▶ Most data sets I receive, whether small or large, contain missing data
- ▶ Usually I want to fit some kind of regression model and I don't want to have to throw away data
- ▶ I don't really care too much about the missing data values, but I do care about getting the parameter estimates or predictions correct

I will focus on missing data analysis in the case of regression models

Structure of this class

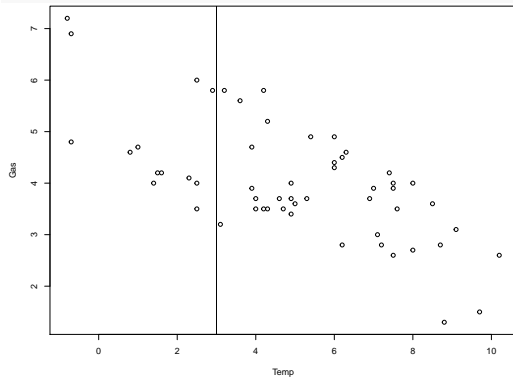
- ▶ Learn some of the missing data analysis jargon
- ▶ See an overview of how missing data analysis works (i.e. single and multiple imputation)
- ▶ Learn to think about types of missing data analysis that might be appropriate for your work

A simple example - Whiteside data

Task: understand the relationship between gas consumption and temperature

```
##    Temp Gas
## 1 -0.8 7.2
## 2 -0.7 6.9
## 3    NA 6.4
## 4  2.5 6.0
## 5  2.9 5.8
## 6  3.2 5.8
```

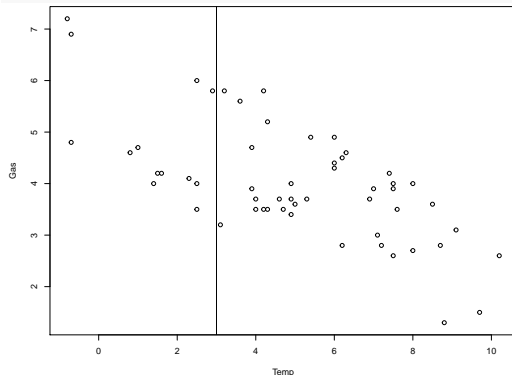
```
plot(whiteside2)
abline(v = 3.0)
```



Questions to think about

- ▶ Why is that observation missing?
Was it a coding mistake or was there something special about that particular value?
- ▶ How might we fill it in?
- ▶ Do I care about that actual value or am I more interested in a statistical model of this relationship?

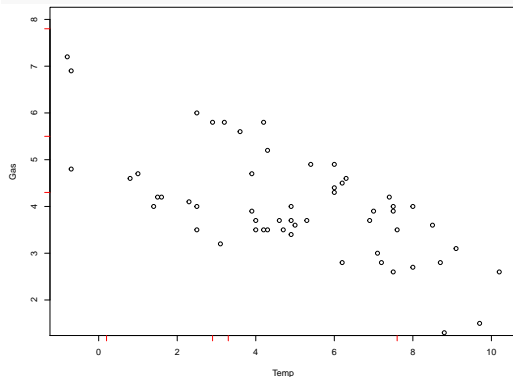
```
plot(whiteside2)  
abline(v = 3.0)
```



Missing data in regression models

- ▶ The pattern of missingness might be more complex
- ▶ We might have some missing response variables
- ▶ We might have various combinations of missing explanatory variables
- ▶ If these missing variables are at the edges of the parameter space, or are particularly influential on the response, they might have more influence on the fitted models

```
plot(whiteside2, ylim = c(1.5, 8))  
axis(side = 1, at = c(0.2, 2.9, 3.3, 7.8, 10.5))  
axis(side = 2, at = c(7.8, 4.3, 5.5),
```



Simple options for completing the analysis

- ▶ Delete all the missing observations (listwise deletion or `complete.cases`)
- ▶ Try to analyse just pairs of the data points that are complete (pairwise deletion or `pairwise.complete.obs`)
- ▶ Just take the overall mean/median of that variable
- ▶ Fit a model and then use it to predict the missing values (or the inverse of it for missing covariates)
- ▶ Fill in the missing values from above or below
- ▶ Set values to zero and include a missingness indicator (1 or 0) as an extra covariate
- ▶ ...

These methods tend to have poor or unknown bias and calibration properties for predicting missing values so will try to use more formal/established methods.

Two important bits of maths we need for this session

For events A, B, C :

$$p(A, B|C) = p(A|B, C) \times p(B|C)$$

(This comes from Bayes' theorem)

If $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$ then:

$$y_1|y_2 = a \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

(Proof of this is boring and complicated)

Missing data analysis (terrible) jargon

- ▶ Missing completely at random (**MCAR**) means the cause of the missingness was completely unrelated to the data. For example, the thermometer broke that day, or Mr Whiteside forgot to take a measurement
- ▶ Missing at random (**MAR**) means that the missingness only depends on the observed data. Given the observed, the data are MCAR. For example, there might be more missing gas consumption values on low temperature days, because Mr Whiteside didn't want to go outside to his gas meter.
- ▶ Not missing at random (**NMAR** or MNAR) means the missingness depends on unobserved data that we do not have. For example, we might be missing gas consumption data due to an outage that we weren't aware of. If that information is causally linked to temperature or gas consumption itself then it may change the relationship.

MAR is the most common scenario and we will focus on that today, and cover MNAR in more detail in later classes.

Missing data analysis mathematical notation

Let:

- ▶ Y be the variables we are interested in, as a matrix of n observations and p variables
- ▶ y_{obs} the observed data
- ▶ y_{mis} the missing data
- ▶ M an $n \times p$ matrix that defines which observations/variables are missing, with $m_{ij} = 1$ if observation i and variable j are missing, and 0 otherwise
- ▶ ψ some parameters governing the missing data mechanism

Now:

- ▶ MCAR means $P(m = 1 | y_{\text{obs}}, y_{\text{mis}}, \psi) = P(m = 1 | \psi)$
- ▶ MAR means $P(m = 1 | y_{\text{obs}}, \psi) = P(m = 1 | \psi)$
- ▶ NMAR means $P(m = 1 | y_{\text{obs}}, y_{\text{mis}}, \psi)$ is not reducible!

A simple example

Suppose that the data have $n = 100$ and $p = 2$ and we use a joint normal distribution to model the data. Suppose that the first variable y_1 is complete, and the second variable y_2 contains missing values)

The missingness probability is:

$$P(m = 1) = \psi_0 + \psi_1 \frac{e^{y_{\text{obs}}}}{1 + e^{y_{\text{obs}}}} + \psi_2 \frac{e^{y_{\text{mis}}}}{1 + e^{y_{\text{mis}}}}$$

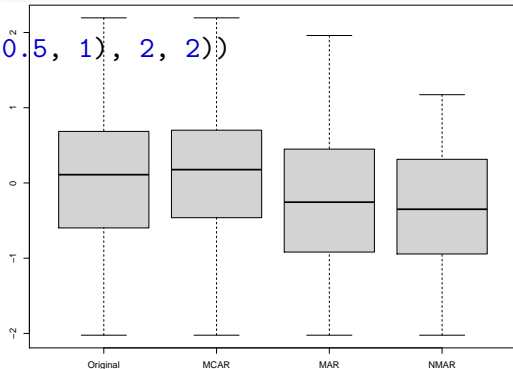
The different missingness mechanisms correspond to:

- ▶ MCAR: $\psi_1 = \psi_2 = 0$
- ▶ MAR: $\psi_2 = 0$
- ▶ NMAR: $\psi_2 \neq 0$

Simple example (cont)

```
n = 100
p = 2
y = mvrnorm(n, mu = rep(0,p),
            Sigma = matrix(c(1, 0.5, 0.5, 1), 2, 2))
m_mcar = rbinom(n, 1, 0.5)
m_mar  = rbinom(n, 1, plogis(y[, 1]))
m_nmar = rbinom(n, 1, plogis(y[, 2]))
y_mcar = y_mar = y_nmar = y
y_mcar[which(m_mcar == 1),2] = NA
y_mar[which(m_mar == 1),2] = NA
y_nmar[which(m_nmar == 1),2] = NA
```

```
boxplot(y[,2], y_mcar[,2], y_mar[,2],
        names = c('Original', 'MCAR',
```



Fitting a model and imputing data at the same time

- ▶ Usually, we want to estimate a model (e.g. a linear regression model) at the same time as imputing the missing values.
- ▶ Assume θ are the parameters associated with that model, and that ψ are the missingness parameters as before
- ▶ We need to find

$$p(y_{\text{obs}}, M | y_{\text{mis}}, \theta, \psi) = p(m | y_{\text{obs}}, y_{\text{mis}}, \psi) \times p(y_{\text{obs}} | y_{\text{mis}}, \theta)$$

- ▶ The first term is the missingness probability model, the second term is the model (e.g. linear regression) that we want to fit.
- ▶ The trick is to think about these in the different MCAR, MAR, and NMAR circumstances

Fitting models under missingness assumptions

- If MCAR:

$$p(y_{\text{obs}}, M | y_{\text{mis}}, \theta, \psi) = p(m | \psi) \times p(y_{\text{obs}} | y_{\text{mis}}, \theta)$$

so the models are completely separate

- If MAR:

$$p(y_{\text{obs}}, M | y_{\text{mis}}, \theta, \psi) = p(m | y_{\text{obs}}, \psi) \times p(y_{\text{obs}} | y_{\text{mis}}, \theta)$$

so separate *provided* there is no link between ψ and θ , otherwise you need both models

- If NMAR. No simplification. So you need both models to be able to fit anything

Ignorability

- ▶ Most importantly, if you are in an MCAR or MAR situation where the parameters ψ and θ are not linked, the missingness is known as **ignorable**
- ▶ This means you just need to fit the statistical bit of the model: $p(y_{\text{obs}}|y_{\text{mis}}, \theta)$
- ▶ This is not a free lunch! Fitting this model might be extremely fiddly as you still might need to estimate the missing data y_{mis} to get any results
- ▶ If you are in an MAR situation with ψ and θ linked, or in an NMAR situation, then it is said that the missingness is **non-ignorable**

One final bit of jargon

- ▶ The pattern of missingness can sometimes be important, especially in imputation
- monotone missing ignorable non-ignorable

Missing data analysis common notation

Other examples where missing data occurs

Bivariate MAR

book P22

How people usually analyse missing data

Book P23 (p35)

Available case vs complete case analysis

Complete case (Ch 3) Available case (p61/73)

Simple imputation

P67 Ch 4

Explicit vs implicit imputation

P68/80

A warning about imputation

P67 Ch 4

A simple demonstration of the pitfalls of imputation

P81/69

Regression imputation

P71/83

General guidelines on imputation

P81/93

Estimating uncertainties with imputed data

P90/102

From simple to multiple imputation

P95/107