Class 1: An introduction to Missing Data Analysis

Andrew Parnell
andrew.parnell@mu.ie

Maynooth University
National University of Ireland Maynooth

https://andrewcparnell.github.io/mda_course

Let's get started

- Introduction from Novartis
- Timetable for the course
- 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

Pre-requisites

These things I'm assuming you already know:

1. Understanding of basic univariate probability distributions such as the normal and the binomial
2. Understanding of basic probability theory and e.g. Bayes' theorem
3. Understanding of linear regression, least squares, and interpreting the output of e.g. `lm`
4. Knowing how to use R and basic plots

It helps if you know a little bit about multivariate distributions such as the multivariate normal and multinomial

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
- 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

Structure of this class

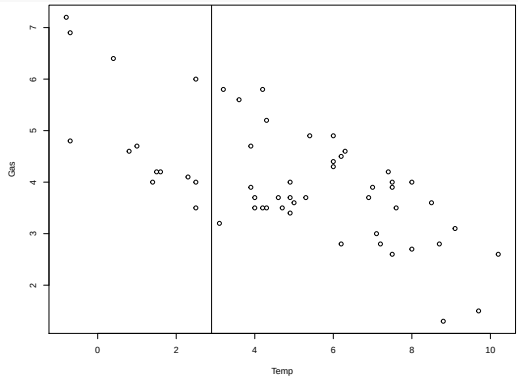
- Learn some of the missing data analysis jargon
- See an overview of how missing data analysis works (i.e. single and multiple imputation)
- Learn to think about types of missing data analysis that might be appropriate for your work

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  0.4 6.4
## 4  2.5 6.0
## 5  2.9  NA
## 6  3.2 5.8
```

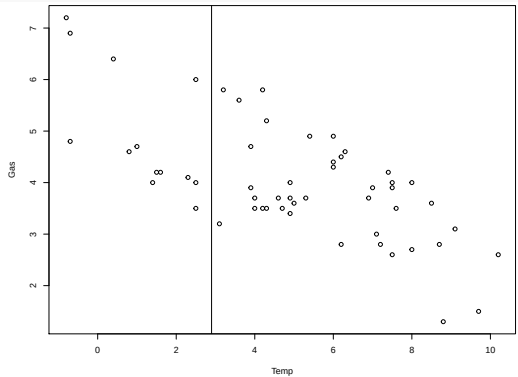
```
plot(whiteside2)
abline(v = 2.9)
```



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?

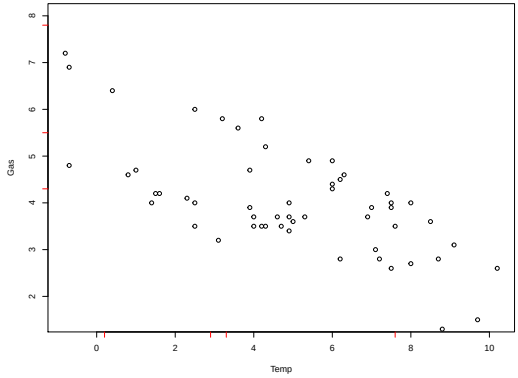
```
plot(whiteside2)
abline(v = 2.9)
```



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
     labels = FALSE, col.ticks = 'red'
axis(side = 2, at = c(7.8, 4.3, 5.5),
     labels = FALSE, col.ticks = 'red'
```



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
- Weight the observations so that the analysis is biased to those observations in classes most likely to be missing
- ...

These methods tend to have poor or unknown bias and calibration properties for predicting missing values so we 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:

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

(This can also be proved using Bayes' theorem)

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 information, 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

$$P(m = 1|Y_{\text{obs}}, Y_{\text{mis}}, \psi)$$

$$= P(m = 1|Y_{\text{obs}}, \psi)$$

- NMAR means

$$P(m = 1|Y_{\text{obs}}, Y_{\text{mis}}, \psi)$$

must depend on Y_{mis}

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:

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

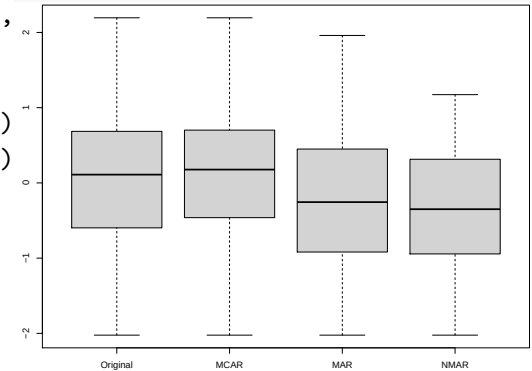
The different missingness mechanisms correspond to:

- 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], y_nmar[,2],
        names = c('Original', 'MCAR',
                  'MAR', 'NMAR'))
```



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**

Some final bits of jargon

- The pattern of missingness can sometimes be important, especially in MAR models where you need to include the missing data as parameters
- A clever trick occurs when the missing data are **monotone missing** which means that you can arrange it in a 'staircase' format so that once a variable is missing, it is always missing for all of the other variables.
- This means that, if the data are MAR, you can write out the model in a chain of equations:

$$p(Y|Y_{\text{mis}}, \theta) = p(Y_1|Y_2, \ldots, Y_p, \theta) \times p(Y_2|Y_3, \ldots, Y_p, \theta) \times \ldots \times p(Y_p|\theta)$$

- Why does this help? Because if the data are **multivariate normal**, all the conditional distributions are known
- Monotone missing is common in longitudinal data studies where, e.g. participants drop out of the study and so henceforth all their data is missing

Imputation

- Filling in the missing values in a data set is called **imputation**
- If you fill in the missing values with 'best' values, this is called **single imputation**
- If, by contrast, you give lots of different options for the missing values, this is called **multiple imputation**
- Imputation methods tend to focus on replacing the missing data values (i.e. estimating Y_{mis}) and less on estimating model parameters θ

The methods we will consider all involve multiple imputation.

Regression imputation

- One of the main methods we will focus on is **regression imputation**
- In this method, each of the variables is used as the response in turn, and regression against the complete cases of the other variables, i.e.

$$\hat{y}_{ij} = \hat{\beta}_0^{(j)} + \sum_{k \neq j} \hat{\beta}_k^{(j)} y_{ij}$$

- In more complex versions, this might not be a linear regression, but a spline or other complicated function
- The simplest version is a single imputation method

Stochastic regression imputation

- The simple regression imputation method can be improved with simulation methods
- For example, For a linear regression model with residual variance σ^2, we can include that uncertainty in the prediction of the missing value
- We also have uncertainty in the estimation of the regression parameters

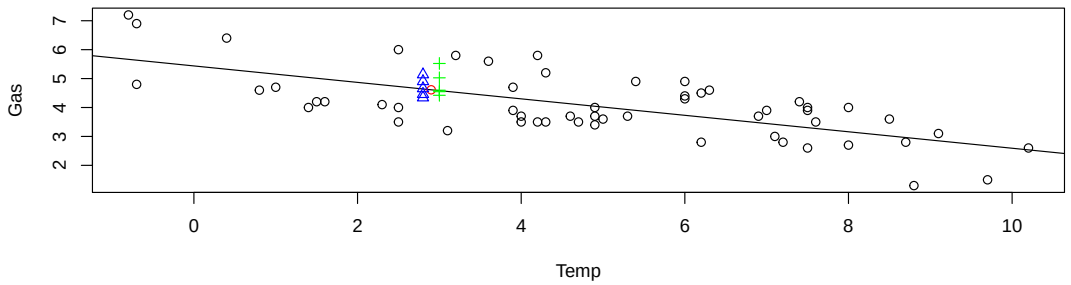
So

$$Var(y_j) = \sigma^2 \left(1 + X^*(X^T X)^{-1}(X^*)^T\right)$$

where X^* is the matrix of explanatory variables associated with the missing value and $X = Y_{i,-j}$

Whiteside example

```
library(mice)
plot(whiteside2)
abline(m<-lm(Gas~Temp, data=whiteside2))
points(2.9, predict(m, data.frame(Temp = 2.9)), col = 'red', pch = 1)
imp_2 = mice(whiteside2, m=5, meth="norm.nob", maxit=1, print = FALSE)
points(rep(2.8, 5), imp_2$imp$Gas, col = 'blue', pch = 2)
imp_3 = mice(whiteside2, m=5, meth="norm", maxit=1, print = FALSE)
points(rep(3.0, 5), imp_3$imp$Gas, col = 'green', pch = 3)
```



Other types of MI - predictive mean matching

- An alternative idea to that on the previous slide is called **predictive mean matching**
- This is where we find the closest data points to missing values and use those as the predicted missing values
- A further extension is to use the distance to the regression line as a weight and sample from the probability distribution to replace values as above.
- This works well if the sample size is large and so there are lots of candidate donors
- (Also known as **Hot Deck** imputation, cf **Cold Deck** where external data is used to impute)

General guidelines on imputation

- If the missingness is ignorable, we need to fit a model of the form: $p(Y_{\text{obs}}|Y_{\text{mis}}, \theta)$
- We should try to use as much of the observed information as possible to build that model
- We should avoid using single imputation
- Any models we build using the imputed data must take account of the uncertainty in it

(More later on how to evaluate imputations)

Summary so far

- The different types of imputation are MCAR (missing completely at random), MAR (missing at random), and NMAR (not missing at random)
- If the missingness mechanism is ignorable (MCAR/MAR + an extra assumption) we can fit a standard statistical model to the data to try to estimate the missing values and/or estimate the model parameters
- In the next session we will talk about how to build these models using both the likelihood and `mice` techniques