Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

How does R handle missing data?

- Suppose the vector *y* contains three numbers, for example, 3,4, and 7. If we are to compute the mean of y, we simply do

```
y=c(3,4,7); mean(y)
[1] 4.666667
```

- Now let us suppose that the last number is missing, `R` indicates this by the expression `NA`, which stands for “not available”.

```
y=c(3,4,NA); mean(y)
[1] NA
```

- The mean is now undefined.
- If we look at the help of the function mean, by typing `help(mean)`, we notice the extra argument `na.rm`, which by default is set to `FALSE`.
- If we instead set it to `TRUE`, this will remove any missing values before calculating the mean.

```
y=c(3,4,NA); mean(y,na.rm=TRUE)
[1] 3.5
```

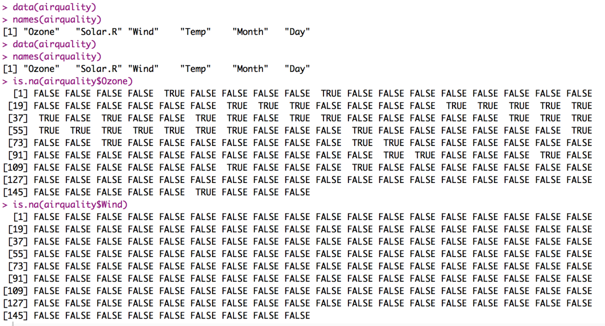
- This makes possible to compute a result but, of course, the set of observations on which the computations are based has changed.
- This may cause problems when conducting statistical inference.

How does R handle missing data?

- Let us investigate the popular `lm` function, where lm stands for linear model/modelling.

- Only for illustrative purposes, let us use the built-in dataset `airquality`. Suppose we want to predict daily ozone concentration (ppb) from wind speed (mph).

- We start by investigating whether each of the variables has any missing data.



```
> data(airquality)
> names(airquality)
[1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
> data(airquality)
> names(airquality)
[1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
> is.na(airquality$Ozone)
  [1] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [19] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
 [37]  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
 [55]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
 [73] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
 [91] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE
[109] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[127] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
> is.na(airquality$Wind)
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [19] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [55] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [91] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[127] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

How does R handle missing data?

↪ We then fit the regression model

```
fit=lm(Ozone~Wind, data=airquality)
fit
Call:
lm(formula = Ozone ~ Wind, data = airquality)

Coefficients:
(Intercept)         Wind
     96.873       -5.551
```

↪ We got no error message. This is because the `lm` function automatically excludes missing values, before fitting the model.

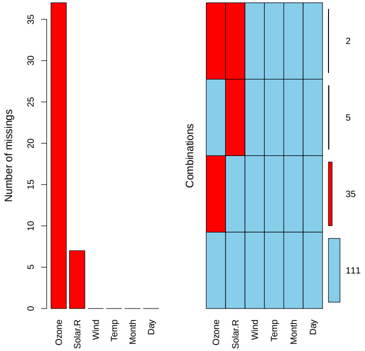
↪ We can check this by typing

```
deleted=na.action(fit)
naprint(deleted)
[1] "37 observations deleted due to missingness"
```

↪ **Message**: It is mandatory, before using any built-in function, to check how it handles missing values.

How does R handle missing data?

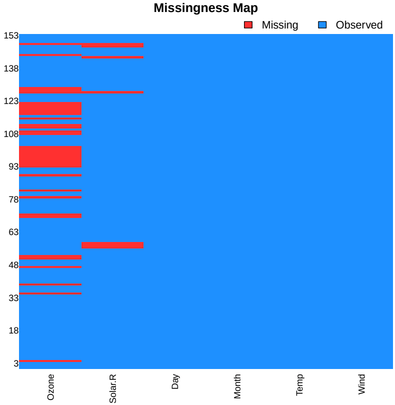
- The R packages VIM, which stands for *Visualization and Imputation of Missing values*, and Amelia, *A Program for Missing Data* have, amongst other very useful functions, helpful exploratory tools.
- For instance, using the function aggr in VIM, we can produce the following plot for the airquality dataset:



How does R handle missing data?

- ↪ The plot on the left simply shows the number of missing observations for each variable in the dataset.
- ↪ The plot on the right shows the amount of missing values in certain combination of variables.
- ↪ For instance, the number 111 on the bottom row means that there are 111 observations in the dataset with no missing in all variables.
- ↪ By opposition, the number 2 on the top row, means that there are 2 observations with missing values in both the ozone and solar radiation variables.
- ↪ Additionally, the function `missmap` in `Amelia` shows where missingness occurs in a dataset.

How does R handle missing data?



↪ On the *x*-axis we have the variables contained in the dataset and on the *y*-axis nn we have the observations.

Example

↪ This example is adapted from Schafer and Graham (2002).

↪ Suppose that systolic blood pressure of n individuals is recorded both in January (Y_1) and in February (Y_2).

↪ Data were simulated for $n = 30$ subjects from a bivariate normal distribution, with means $\mu_1 = \mu_2 = 125$, standard deviations $\sigma_1 = \sigma_2 = 25$, and correlation $\rho = 0.6$.

↪ We will impose missing on Y_2 using the three distinct mechanisms learned in the last lecture: MCAR, MAR, and MNAR.

↪ See associated `R` (`example.simulated.blood.pressure`) script.

Example

↪ The 10 measurements in February were randomly selected from those in January; this missing mechanism is obviously MCAR.

↪ To induce a MAR mechanism we do the following: those who have measurements in February are those whose January's measurement exceed 140 (Y_1>140), a threshold used for diagnosing high blood pressure or hypertension.

↪ Finally, to induce a MNAR mechanism, the following was implemented: all individuals with measurements in February are those whose February measurement itself exceeded 140. This could happen for instance if all individuals had their measurements in February but the staff person only recorded it in case it was in the hypertensive range.

↪ MNAR could be induced in other ways, e.g., February measurement only recorded if it is substantial different from the January one.

Example

Y_1	Y_2	Y_2-MCAR	Y_2-MAR	Y_2-MNAR
96	126			
130	128	128		
102	111			
161	160		160	160
148	117		117	
111	102	102		
140	131			
142	141	141	141	141
126	150			150
110	127			
161	157	157	157	157
137	131			
103	119			
69	82	82		
158	142		142	142
132	116	116		
121	129	129		
138	155			155
145	142		142	142
128	148			148
141	150		150	150
149	136		136	
123	130	130		
93	68			
123	155	155		155
102	146			146
126	117	117		
104	80			
108	121			
136	133			

Example

- The above table was obtained fixing the random number generator seed equal to one, `set.seed(1)`. Changing the seed will change the values in the table.

- The mean and standard values of the variables are as follows:

	Y_1	Y_2	Y_2-MCAR	Y_2-MAR	Y_2-MNAR
Mean	125.43	128.33	125.7	143.13	149.64
Std Dev.	22.12	22.89	22.97	13.44	6.53

- We can notice that as we move from MCAR to MAR to MNAR, the observed Y_2 values become and increasingly select and unusual group relative to the population; the sample mean increases and the standard deviation decreases.

- Although this phenomenon is not a universal feature of MCAR, MAR, and MNAR, it does happen in many realistic examples.

Example

- ↪ Let us now apply the (unpaired) *t*-test to check for the plausibility of MCAR.

- ↪ Remember that data on the MCAR variable were actually generated from a MCAR mechanism. However, due to our reduced sample size, we might fail to reject the equality of means in both groups.

- ↪ We define two groups of Y_1, one for which Y_2 is observed (20 observations) and another one for which Y_2 is not observed (10 observations). We then postulate that the means of both groups are equal.

- ↪ Using the `t.test` command (see associated `R` script), we obtain a *p*-value much greater than 0.05, thus failing to reject the null hypothesis of equality of means.

- ↪ If we do the same but for the MAR data, we obtain a *p*-value much smaller than 0.05, thus rejecting the null hypothesis of mean equality (and thus reject that data are MCAR). This result is consistent with the fact that the data was generated under a MAR mechanism.

A more formal description of the missing data mechanisms

- ↪ In the previous slides, we have introduced the concepts of MCAR, MAR, and MNAR. We will now look at more precise definitions of these mechanisms. To do so, we need to introduce some notation and terminology.
- ↪ The complete data consist of the values one would have obtained if there were no missing data and it is partially a hypothetical entity because some of its values are missing.
- ↪ However, in principle, each individual has a value on each variable.
- ↪ Despite this notion being sensible or intuitive in most situations (e.g., a student's test score is missing because he/she did not attend school on that day), it can also be kind of unnatural in some other situations (e.g., a cancer patient has his/her quality of life score missing because he/she already died).
- ↪ Nevertheless, we need to assume that a complete set of values does exist, at least, hypothetically, and we will denote it as Y_{com}.

A more formal description of the missing data mechanisms

- ↪ In most practical settings, some portion of the hypothetical complete data will be missing.
- ↪ We can then think of the complete data as consisting of observed and missing data, denoted by Y_{obs} and Y_{mis}, respectively.
- ↪ The fundamental idea behind Rubin's (1976) theory is that missingness is a variable that has a probability distribution.
- ↪ Concretely, Rubin defines a binary missing data indicator variable, typically denoted in the literature by R, taking the value 1 if a value is observed and the value 0 if the value is missing.
- ↪ In the BMI/glucose level example, a single variable can summarise the distribution of missing data because the glucose level variable is complete (i.e., we only had missing values on one variable, the BMI).

A more formal description of the missing data mechanisms

- However, when more variables are available, it tends to be the case that several variables will have missing data.

- In such case R becomes a matrix of binary missing value indicators.

- When all variables have missing values, then R will be a matrix with the same dimension of the data matrix and whose entries are given by

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ has been observed,} \\ 0 & \text{if } Y_{ij} \text{ is missing,} \end{cases}$$

  for $i = 1, \ldots, n$ (n is the number of observations/sample size) and $j = 1, \ldots, p$ (p is the number of variables).

- The important point to retain is that R has a probability distribution and that this distribution might depend on Y_{obs} and Y_{mis}.

A more formal description of the missing data mechanisms

- ↪ Let ψ contain the parameters of the missing data model. Then, the general expression of the missing data model is

$$\Pr(R = r \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi).$$

- ↪ The data are said to be MCAR if

$$\Pr(R = r \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = r \mid \psi).$$

- ↪ So, under MCAR the probability of being missing is completely unrelated to the data. It depends only on some parameters ψ, the overall probability of missingness.

- ↪ The essential feature of MCAR is that the observed data can be thought of as a random sample of the complete data that would have been obtained if no data were missing.

- ↪ As already noted and explained, the validity of MCAR can be checked empirically from the data at hand against the alternative of MAR, but only under the unverifiable assumption that the missing data mechanism is not MNAR.

A more formal description of the missing data mechanisms

- The data are said to be MAR if

$$\Pr(R = r \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = r \mid Y_{\text{obs}}, \psi).$$

- So, under MAR, the missingness probability depends on observed information, but is further unrelated to the specific missing values or other unmeasured variables.

- MAR assumption has the following important implication for the distribution of missing data: upon stratification on Y_{obs}, the distribution of Y_{mis} is the same as the distribution of the corresponding observations in the target population. That is, data are MCAR after controlling for Y_{obs}.

- It should be noted that the validity of MAR assumption cannot be checked empirically from the data at hand against MNAR.

A more formal description of the missing data mechanisms

- Finnaly, the data are MNAR if

$$\Pr(R = r \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi),$$

does not simplify, so here the probability of missingness also depends on unobserved information, including Y_{mis} itself.

- A complicated form of MNAR is when missingness depends on a completely unobserved/unmeasured variable.

- For a concrete example, think again on the BMI/glucose level example. Suppose that the true missing mechanism for BMI is MAR, hence meaning that individuals with missing values of BMI may be more likely to have extreme blood glucose levels. However, the MAR missing values in BMI would become MNAR if we had no measurements of glucose at all.

Numerical illustration

- This example is from van Buuren (2012, pp. 31).
- The aim is to simulate data from MCAR, MAR and MNAR mechanisms.
- Let the data $Y = (Y_1, Y_2)$ be simulated from a standard bivariate normal distribution with correlation $\rho_{Y_1, Y_2} = 0.5$.
- Missing data are created in Y_2 using the missing data model

$$\Pr(R_2 = 0) = \psi_0 + \frac{e^{Y_1}}{1 + e^{Y_1}}\psi_1 + \frac{e^{Y_2}}{1 + e^{Y_2}}\psi_2,$$

with different parameter settings for $\psi = (\psi_0, \psi_1, \psi_2)$.

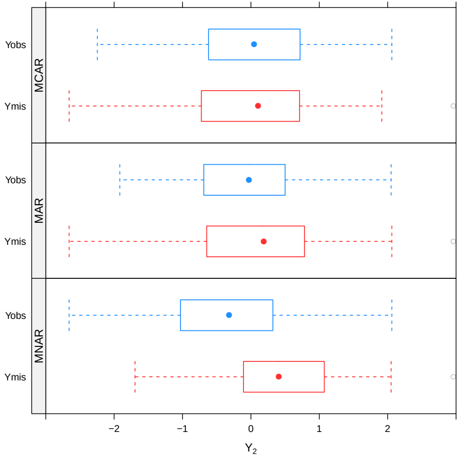
Numerical illustration

- For MCAR we set $\psi_{\text{MCAR}} = (0.5, 0, 0)$, for MAR we set $\psi_{\text{MAR}} = (0, 1, 0)$, and for MNAR we set $\psi_{\text{MNAR}} = (0, 0, 1)$. Thus, we obtain the following models:

  1. MCAR: $\Pr(R_2 = 0)$=0.5
  2. MAR: $\Pr(R_2 = 0) = \frac{e^{Y_1}}{1+e^{Y_1}}$, or equivalently, $\text{logit}\{\Pr(R_2 = 0)\} = Y_1$.
  3. MNAR: $\text{logit}\{\Pr(R_2 = 0)\} = Y_2$.

- Here, $\text{logit}(p) = \log\{p/(1-p)\}$, for $0 < p < 1$, is the logit function.

Numerical illustration



Numerical illustration

- ↪ The figure in the previous slide displays the distribution of Y_{obs} and Y_{mis} under the three different missing data models. Note that this is only possible because we are simulating data; with real data we do not get to know the Y_{mis} values.

- ↪ As our intuition would dictate, the distribution of Y_{obs} and Y_{mis} are similar under MCAR, but become more and more distinct as we move to the MNAR case.

- ↪ Associated `R` script: `simulated.example.2`.

A little more on testing MCAR versus MAR

- ↪ As we will see later, simple techniques usually only work under MCAR, but this assumption is too restrictive and, most of the times, unrealistic.
- ↪ The majority of this course will be focused on methods that can handle the more general MAR data. If time permits, we will briefly cover methods for MNAR data.
- ↪ Several tests have been proposed in the literature to test MCAR versus MAR.
- ↪ These tests are not routinely used and their practical value is still not clear. Remember that we can never rule out the possibility of MNAR.
- ↪ Enders (2010, pp. 17–21) contains a good discussion of this topic.

A little more on testing MCAR versus MAR

- ↪ We have already seen the unpaired *t*-test. We have seen it in a very simple context, where we had two variables, one of which had missing values.

- ↪ If we have several variables and missing in one of them, we need to perform a series of *t*-tests.

- ↪ This approach separates the missing and observed values on a particular variables and uses a *t*-test to examine group mean differences on the other variables in the dataset.

- ↪ The MCAR mechanism implies that 'subjects' with observed values should be the same as the 'subjects' with missing values, on average.

- ↪ As a consequence, a non significant *t*-test provides evidence that data are MCAR, whereas a significant *t* statistic (or, alternatively, a large mean difference) suggests that data are MAR or MNAR.

- ↪ However, as the number of variables grows, performing and assessing *t*-tests gets tedious.

A little more on testing MCAR versus MAR

- ↪ The main advantage for implementing the t-test approach is to identify (auxiliar) variables that can later adjust for in the missing data handling procedure.

- ↪ As a final remark, it is important to mention that mean comparisons do not provide a conclusive test of MCAR, because MAR and MNAR mechanisms can produce missing data subgroups with equal means.

- ↪ Little (1988) proposed a multivariate version of the t-test that simultaneously evaluates mean differences on every variable in the data set.

- ↪ For details see Little (1988) or Enders (2010, pp. 19–20). This can be carried out by the `LittleMCAR` function in the `BaylorEdPsych` `R` package.

A little more on ignorability versus nonignorability

- The ψ parameters associated to the probability of missing data have no scientific interest (e.g., had the data been complete there would be no reason to worry about ψ) and are generally unknown.

- It would simplify greatly the analysis if we could just ignore these parameters. However, in some situations, these parameters may influence the estimates of the parameters of interest (those associated with Y_{com}), say θ.

- The practical importance of Rubin's distinction between MCAR, MAR, and MNAR is that it clarified the conditions that need to exist in order to accurately estimate θ without the need to know ψ.

A little more on ignorability versus nonignorability

- ↪ Rubin showed that likelihood based analyses (e.g., maximum likelihood) and multiple imputation do not require information about ψ if:

    1. the data are MAR or MCAR, and
    2. the parameters θ and ψ are distinct, in the sense that the joint parameter space of (ψ, θ) is the product of the parameter space of θ and the parameter space of ψ.

- ↪ Schafer (1997, p.11) says that in many situations the second condition is, at least, reasonable from an intuitive point of view, given that knowing θ will provide little information about ψ and vice-versa.

- ↪ For this reason, missing data literature often describes MAR (and MCAR!) data as ignorable.

A little more on ignorability versus nonignorability

- As a cautionary message it should be noted that the term ‘ignorable’ does not mean that we can simply ‘ignore’ the missing data.

- For inferences to be valid, we need to condition on these variables/factors that influence the missing data rate.

- For example, in the previous example, in the MAR scenario, a valid estimate of the mean of Y_2 (variable with missing values) cannot be made without Y_1, so somehow we should make use of Y_1 in the computation of the mean of Y_2.

How to prevent MNAR missingness?

- ↪ As we had already quoted, the ideal solution to the missing data problem would be to have none.

- ↪ Missing data prevention requires a careful experiment's design and a very careful execution as well.

- ↪ Most of the methods we will cover assume MAR data, However, we cannot be sure whether the data are really missing at random, or whether the missingness depends on unobserved variables or the missing data themselves.

- ↪ The idea is to start the study with a data collection strategy that will turn MNAR missingness into MAR missingness.

- ↪ This, so called inclusive analysis strategy, incorporates variables that are known to be correlated with the missing prone variables. Then, missing values will be more likely to be MAR than MNAR.

- ↪ These correlates variables are called auxiliary variables in the missing data literature.

How to prevent MNAR missingness?

- ↪ Note that auxiliary variables might not be of substantial interest in the sense that they would not have been included in the analysis had the data been complete.

- ↪ Theory and past research, as well as the MCAR tests, can help to identify auxiliary variables.

- ↪ Not that the inclusion of auxiliary variables *per se* does not guarantee that the MAR assumption is satisfied, but it certainly improves the chances of it.

- ↪ For instance, it may be a strong assumption that nonresponse to an income question in a survey depends only on gender, race and education, but this is certainly a lot more plausible than assuming the probability of nonresponse is constant, or that it depends only on one of these variables.

Planned missing data designs

- At the very beginning, we have seen an example (nutrition study) where missing, instead of out of the researcher control, was 'induced' on purpose.
- The idea of planned missing data design is to intentionally generate MCAR or MAR data.
- A classic example of intentional MAR data occurs in selection designs where values on one variable determine whether respondents provide data on a second variable.
- As a concrete example, we mention the third US National Health and Nutrition Examination Survey wherein cystanin-C (a marker of renal disease) was measured amongst individuals, whose ages range from 12 to 59 years old, in a random sample, plus all individuals with high serum creatinine (>1.2mg/dl in men and >1.0 mg/dl in women).
- Cystanin-C is then MAR because it is randomly missing within those without high creatinine and missing with probability zero amongst those with high creatinine.
- In this example, creatinine is the conditioning variable.