Multiple Imputation and subsequent calculations

Simon Ress | Ruhr-Universität Bochum

September 22, 2022

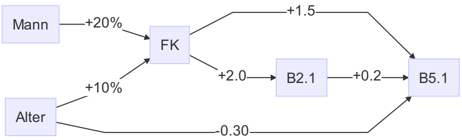
Workshop at hr&c, Bochum, 2021

Content

Examplary Data Set

Planing the interdependencies of variables

- Every variable is constructed by an *random term*
- Some variables are influenced by *values of other variables*
- **e.g. “Mann” = 1 increases the probability for FK=1 by 20%**



Data Set Creation

```r
#Create variables
set.seed(415)
Mann = ifelse(runif(5000,0,1) < 0.50, 1, 0)
Alter = as.numeric(cut(runif(5000,20,70),
                       c(20,30,40,50,60,70)))
FK = ifelse((Mann*0.2 + Alter*0.1 +
             runif(5000,0,0.6)) > 0.95, 1, 0)
B2.1 = as.numeric(cut(FK*2 +
                      rnorm(5000,2,0.35), c(0,1,2,3,4,6)))
set.seed(1015)
B5.1 = as.numeric(cut(FK*1.5 + B2.1*0.2 + Alter*(-0.30) +
                      rnorm(5000,2.5,0.30), c(-2,1,2,3,4,8)))

#Build data frame
df = data.frame(Mann, Alter, FK, B2.1, B5.1)
```

View Data Set

```r
head(df,10)
```

```
##    Mann Alter FK B2.1 B5.1
## 1     0     4  0    2    2
## 2     0     4  0    1    2
## 3     1     5  0    3    2
## 4     1     5  1    4    3
## 5     0     4  0    2    2
## 6     1     3  1    4    4
## 7     0     3  0    3    3
## 8     0     1  0    3    3
## 9     0     4  0    2    2
## 10    0     2  0    2    3
```

I. Check whether true effects can be estimated

Table 1: lm(FK~Mann) | Mann: +0.2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0596	0.0070	8.5438	0
Mann	0.2001	0.0099	20.1491	0

Table 2: lm(FK~Alter) | Alter: +0.1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.1406	0.0110	-12.7686	0
Alter	0.1002	0.0033	30.0803	0

Table 3: lm(B2.1~FK) | FK: +2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.4980	0.0079	318.0197	0
FK	2.0115	0.0197	101.8565	0

II. Check whether true effects can be estimated (!)

Table 4: lm(B5.1~FK) | FK: +1.5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6745	0.0088	302.2230	0
FK	1.4519	0.0222	65.2572	0

Table 5: lm(B5.1~B2.1) | B2.1: +0.2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3494	0.0283	47.6337	0
B2.1	0.5521	0.0096	57.5786	0

Table 6: lm(B5.1~Alter) | Alter: -0.3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.2286	0.0251	128.6139	0
Alter	-0.1088	0.0076	-14.3234	0

Excursus: Modern Causal Analysis

- Satisfaction of the Conditional Independence Assumption (CIA) necessary to estimate true causal effects
- Meet the CIA using an appropriate set of control variables
- Choose control variables by a Directed Acyclic Graph (DAG)

Excursus MCA: DAG (B5.1 <- FK)

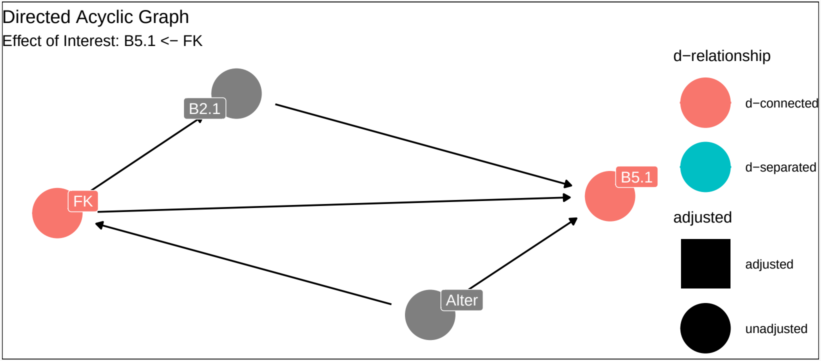


Table 7: lm(B5.1~FK+B2.1+Alter) | FK: +1.5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9909	0.0311	96.2506	0
FK	1.5033	0.0283	53.1769	0

Excursus MCA: DAG (B5.1 <- B2.1)

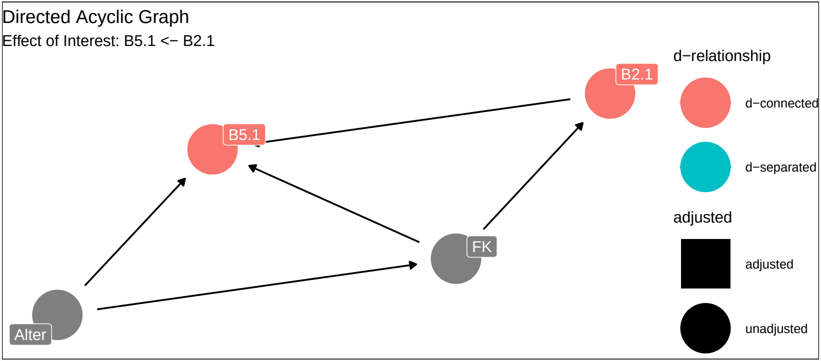


Table 8: lm(B5.1~B2.1+FK+Alter) | B2.1: +0.2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9909	0.0311	96.2506	0
B2.1	0.2031	0.0112	18.0842	0

Excursus MCA: DAG (B5.1 <- Alter)

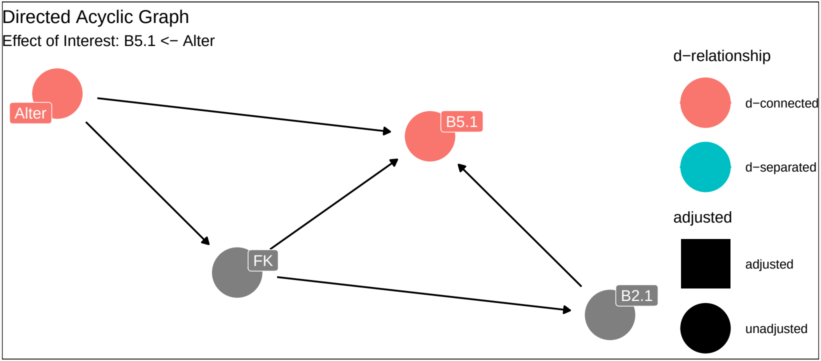


Table 9: lm(B5.1~Alter+FK+B2.1) | Alter: -0.3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9909	0.0311	96.2506	0
Alter	-0.3007	0.0044	-68.9450	0

Missing Patterns

Missing Patterns: Missing completely at random (MCAR)

- Values are randomly missing in the dataset
  - Missing data values do not relate to any other data
  - There is no pattern to the actual values of the missing data themselves
- For instance, when smoking status is not recorded in a random subset of patients
- This is easy to handle, but unfortunately, data are almost never missing completely at random

MCAR: Inserting Missing values in data frame

Missing Patterns: Missing at random (MAR)

- Confusing and would be better stated as *missing conditionally at random*
- Missing data do have a relationship with other variables in the dataset
  - Whether a value is missing or not depends on other variables
- The actual values that are missing are random
- For example, smoking status is not documented in female patients because the doctor was too shy to ask

Missing Patterns: Missing not at random (MNAR)

- The pattern of missingness is related to other variables in the dataset
- In addition, the values of the missing data are not random
    - Whether a value is missing or not depends on other variables
- For example, when smoking status is not recorded in patients admitted as an emergency, who are also more likely to have worse outcomes from surgery

Main Repos

- Official GitHub Repo of Metropolis (formerly mtheme); older version in TeXLive
- My GitHub Repo for a local Ubuntu package of Metropolis – formerly mtheme
- Manuel

Slide with Bullets

- Bullet 1
- Bullet 2
- Bullet 3

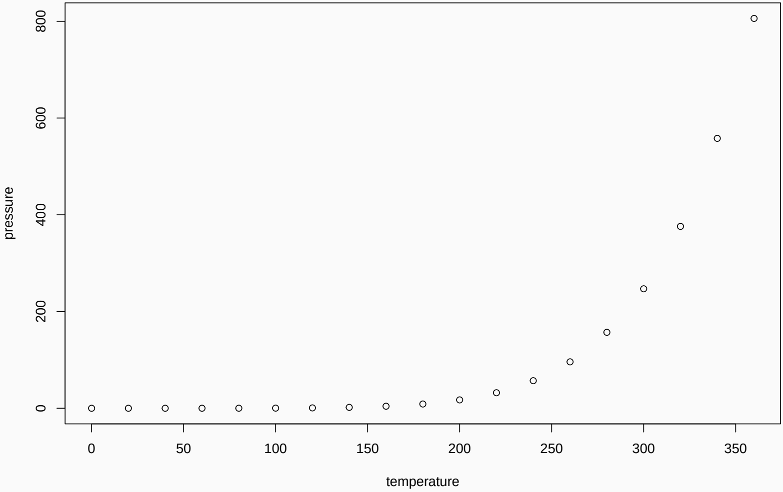
Slide with R Output

Slide with R Output

```r
summary(cars)
```

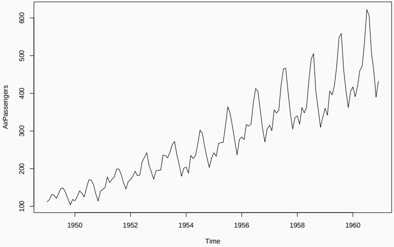
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Slide with Plot



Two column layout

Here is some text above which goes over to whole slide



- Description of plot
- Second point

and here some text below which goes over to whole slide

Breakout page

Figures caption

Figures caption

Figure 1: Figure: Here is a really important caption.

Using LaTeX Parts: Blocks

As one example of falling back into LaTeX, consider the example of three different block environments are pre-defined and may be styled with an optional background color.

Default

Block content.

Alert

Block content.

Example

Block content.