**Missing Data Practical**

There is one exercise to complete.

|  |
| --- |
| In this exercises we will use the following R functions, for which you should briefly read the documentation:  jtools::summ()  glm()  mice::mice()  attributes()  mice::with()  mice::pool()  You may need to install the following packages:  install.packages ("jtools")  install.packages ("mice")  install.packages("tidyverse")    library(jtools)  library(mice)  library(tidyverse) |

In this practical we will explore the missing patterns in some raw data and also examine a multiply imputed dataset. We will use a leprosy dataset. Here is a description of the study:

*Between 1980 and 1984 a population of approximately 112 000 people living in Karonga District, Northern Malawi, were screened for leprosy. Individuals found to have leprosy were not followed further. The remaining population was followed until 1989. During the follow-up period 274 new cases of leprosy were identified, and 1096 controls without leprosy at baseline were selected at random from the screened population. Whether an individual was vaccinated with BCG was assessed  
by the presence or absence of a typical BCG scar when screened. BCG was introduced into Karonga district in mass vaccination campaigns in schools in the late 1970's.*

We want to assess the effect of BCG on the probability of contracting leprosy. The analysis is a logistic regression of leprosy (binary outcome) on BCG (binary baseline variable), adjusted for various confounders (age, sex, school, house).

1. **Exploring the pattern of missingness in the raw data**

Open the karonga.csv dataset. Explore the pattern of missingness in the raw dataset. Missing data are represented in R by “NA”. Which variables have missing data (use skim)? How many individuals have particular missing data patterns?

*# Read data*

kar <- readr::read\_csv(glue::glue("{dir}/{prac}/data/karonga\_2022.csv"))

Note that d is coded as “cases” and “controls”. The application of as.factor works by alphabetical order, so cases=0, and controls =1. We first need to avoid this, in order to get results which are interpretable as “odds of being a leprosy case”.

*# Recode variables and define categorical variables as factors*

kar2 <- kar %>%

dplyr::mutate(

d = case\_when(

d == "Control" ~ 0,

d == "Case" ~ 1,

TRUE ~ NA\_real\_

),

age = factor(

age,

levels = c("5-9", "10-14", "15-19", "20-24", "25-29", "30-44", "45+"),

labels = c("5-9", "10-14", "15-19", "20-24", "25-29", "30-44", "45+")

),

sex = factor(

sex, levels = c("Male", "Female"), labels = c("Male", "Female")

),

bcg = factor(

bcg, levels = c("Absent", "Present"), labels = c("Absent", "Present")

),

house = factor(

house,

levels = c(

"wattle and daub", "temporary shelter",

"sun-dried bricks or pounded mud", "burnt brick"

),

labels = c(

"wattle and daub", "temporary shelter",

"sun-dried bricks or pounded mud", "burnt brick"

)

),

school = factor(

school,

levels = c(

"none", "1-5yr primary", "6-8yr primary", "secondary/tertiary"

),

labels = c(

"none", "1-5yr primary", "6-8yr primary", "secondary/tertiary"

)

)

)

*# Inspect data*

print(kar2, n = 10)

*# Explore extent of missingness in each variable (simple)*

kar |>

purrr::map\_df(\(x) sum(is.na(x)))

# A tibble: 1 × 7

id d age sex bcg school house

<int> <int> <int> <int> <int> <int> <int>

1 0 0 0 0 0 99 123

*# Explore extent of missingness in each variable (a little more complex)*

kar |>

purrr::map\_df(\(x) sum(is.na(x))) |>

tidyr::pivot\_longer(

dplyr::everything(), names\_to = "variable", values\_to = "missing"

) |>

dplyr::mutate(percentage = round(missing / nrow(kar) \* 100, digits = 1))

# A tibble: 7 × 3

variable missing percentage

<chr> <int> <dbl>

1 id 0 0

2 d 0 0

3 age 0 0

4 sex 0 0

5 bcg 0 0

6 school 99 7.2

7 house 123 9

Generate missing data indicators for variables with missing data:

*# Create indicator variables for variables with missing data*

kar2 <- kar2 |>

dplyr::mutate(

dplyr::across(

c(school, house), \(x) dplyr::if\_else(is.na(x), 1, 0),

.names = "{.col}\_na"

)

)

Use the missing data indicators to explore whether the missingness depends on any of the other variables:

Eg, univariate approach

*# Univariable approach*

kar2 |>

dplyr::select(d, age, sex, bcg, house) |>

purrr::map(

\(x) glm(school\_na ~ x, data = kar2) |>

broom::tidy(exponentiate = TRUE, conf.int = TRUE)

)

$d

# A tibble: 2 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.06 0.00781 8.06 1.59e-15 1.05 1.08

2 x 1.05 0.0175 2.67 7.77e- 3 1.01 1.08

$age

# A tibble: 7 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1 0.0146 -1.71e-14 1.00e+ 0 0.972 1.03

2 x10-14 1.00 0.0217 1.90e- 1 8.49e- 1 0.962 1.05

3 x15-19 1.05 0.0234 2.05e+ 0 4.08e- 2 1.00 1.10

4 x20-24 1.11 0.0277 3.83e+ 0 1.34e- 4 1.05 1.17

5 x25-29 1.10 0.0304 2.99e+ 0 2.88e- 3 1.03 1.16

6 x30-44 1.17 0.0230 7.00e+ 0 4.09e-12 1.12 1.23

7 x45+ 1.16 0.0217 7.02e+ 0 3.57e-12 1.12 1.22

$sex

# A tibble: 2 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.07 0.0102 6.98 4.44e-12 1.05 1.10

2 xFemale 1.00 0.0140 0.142 8.87e- 1 0.975 1.03

$bcg

# A tibble: 2 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.10 0.00897 10.8 2.57e-26 1.08 1.12

2 xPresent 0.939 0.0142 -4.40 1.16e- 5 0.914 0.966

$house

# A tibble: 4 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.03 0.00824 3.93 0.0000886 1.02 1.05

2 xtemporary shelter 0.998 0.0383 -0.0548 0.956 0.926 1.08

3 xsun-dried bricks or pounded mud 1.06 0.0150 3.95 0.0000826 1.03 1.09

4 xburnt brick 1.01 0.0161 0.833 0.405 0.982 1.05

kar2 |>

dplyr::select(d, age, sex, bcg, school) |>

purrr::map(

\(x) glm(house\_na ~ x, data = kar2) |>

broom::tidy(exponentiate = TRUE, conf.int = TRUE)

)

$d

# A tibble: 2 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.09 0.00864 10.3 3.27e-24 1.08 1.11

2 x 1.00 0.0193 0.0944 9.25e- 1 0.965 1.04

$age

# A tibble: 7 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.10 0.0166 5.68 0.0000000168 1.06 1.14

2 x10-14 1.02 0.0248 0.698 0.485 0.969 1.07

3 x15-19 0.991 0.0267 -0.344 0.731 0.940 1.04

4 x20-24 1.01 0.0316 0.377 0.706 0.951 1.08

5 x25-29 0.974 0.0347 -0.751 0.453 0.910 1.04

6 x30-44 0.991 0.0262 -0.337 0.736 0.942 1.04

7 x45+ 0.976 0.0248 -0.982 0.326 0.930 1.02

$sex

# A tibble: 2 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.09 0.0113 7.57 7.08e-14 1.07 1.11

2 xFemale 1.01 0.0155 0.567 5.71e- 1 0.979 1.04

$bcg

# A tibble: 2 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.09 0.00997 8.41 1.02e-16 1.07 1.11

2 xPresent 1.01 0.0158 0.943 3.46e- 1 0.984 1.05

$school

# A tibble: 4 × 7

term estimate std.error statistic p.value conf.low conf.high

<chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

1 (Intercept) 1.04 0.0149 2.86 0.00430 1.01 1.07

2 x1-5yr primary 1.03 0.0180 1.85 0.0643 0.998 1.07

3 x6-8yr primary 1.03 0.0200 1.44 0.149 0.990 1.07

4 xsecondary/tertiary 1.02 0.0460 0.393 0.695 0.930 1.11

and multivariate approach:

*# Multivariable approach*

school\_assoc <-

glm(

school\_na ~ d + age + sex + bcg + house, data = kar2, family = "binomial"

) |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 1247 (123 missing obs. deleted)

Dependent Variable: school\_na

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(12) = 82.880, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.199

Pseudo-R² (McFadden) = 0.170

AIC = 430.236, BIC = 496.906

Standard errors: MLE

----------------------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

---------------------------- -------------- ------- ------- -------- -------

(Intercept) 0.000 0.000 Inf -0.031 0.976

d 1.368 0.748 2.503 1.017 0.309

age10-14 1283599.220 0.000 Inf 0.022 0.983

age15-19 8581308.267 0.000 Inf 0.025 0.980

age20-24 26405238.471 0.000 Inf 0.027 0.979

age25-29 12727757.459 0.000 Inf 0.025 0.980

age30-44 32137511.238 0.000 Inf 0.027 0.979

age45+ 27653180.599 0.000 Inf 0.027 0.979

sexFemale 0.699 0.405 1.205 -1.289 0.197

bcgPresent 0.767 0.312 1.883 -0.579 0.563

housetemporary shelter 0.759 0.096 6.000 -0.262 0.794

housesun-dried bricks or 3.101 1.690 5.691 3.653 0.000

pounded mud

houseburnt brick 1.642 0.761 3.543 1.263 0.207

----------------------------------------------------------------------------

house\_assoc <-

glm(

house\_na ~ d + age + sex + bcg + school, data = kar2, family = "binomial"

) |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 1271 (99 missing obs. deleted)

Dependent Variable: house\_na

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(12) = 30.586, p = 0.002

Pseudo-R² (Cragg-Uhler) = 0.061

Pseudo-R² (McFadden) = 0.049

AIC = 619.433, BIC = 686.351

Standard errors: MLE

---------------------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

------------------------------ ----------- ------- ------- -------- -------

(Intercept) 0.076 0.034 0.170 -6.293 0.000

d 1.149 0.617 2.139 0.437 0.662

age10-14 1.198 0.672 2.138 0.613 0.540

age15-19 0.666 0.314 1.415 -1.058 0.290

age20-24 0.828 0.349 1.966 -0.428 0.669

age25-29 0.248 0.056 1.095 -1.840 0.066

age30-44 0.247 0.081 0.750 -2.468 0.014

age45+ 0.203 0.067 0.612 -2.831 0.005

sexFemale 1.143 0.723 1.804 0.572 0.568

bcgPresent 1.113 0.685 1.808 0.432 0.666

school1-5yr primary 1.236 0.624 2.449 0.609 0.543

school6-8yr primary 1.163 0.534 2.536 0.380 0.704

schoolsecondary/tertiary 1.386 0.281 6.845 0.401 0.688

---------------------------------------------------------------------------

9% of individuals have missing data on house. Around half of these individuals also have missing data on school. There are more missing values for house in younger individuals (although not significant).

7% of individuals have missing data on school. There are more missing values for school in older individuals.

1. **Perform a complete-case analysis**

Use the data to explore whether people vaccinated with BCG are less likely to have leprosy.

*# Fit logistic model of `bcg` against `d`*

glm(d ~ bcg, data = kar2, family = "binomial") |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 1370

Dependent Variable: d

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(1) = 81.126, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.091

Pseudo-R² (McFadden) = 0.059

AIC = 1293.977, BIC = 1304.422

Standard errors: MLE

---------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

----------------- ----------- ------- ------- --------- -------

(Intercept) 0.381 0.327 0.444 -12.376 0.000

bcgPresent 0.247 0.176 0.345 -8.166 0.000

---------------------------------------------------------------

How many individuals were included in the analysis? 1370

Now adjust for the age, sex, schooling, housing to assess whether the association between BCG and the probability of contracting leprosy is changed.

*# Fit logistic model of `bcg` against `d` adjusting for other variables*

glm(d ~ bcg + age + sex + house + school, data = kar2, family = "binomial") |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

> summ(glm(d~bcg+ age +sex +house+school, data = df2, family="binomial"), confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 1186 (184 missing obs. deleted)

Dependent Variable: d

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(14) = 120.716, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.155

Pseudo-R² (McFadden) = 0.103

AIC = 1076.001, BIC = 1152.176

Standard errors: MLE

---------------------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

------------------------------ ----------- ------- ------- -------- -------

(Intercept) 0.278 0.153 0.504 -4.214 0.000

bcgPresent 0.346 0.228 0.524 -5.015 0.000

age15-19 1.081 0.567 2.061 0.237 0.812

age20-24 1.293 0.629 2.660 0.699 0.485

age25-29 2.661 1.394 5.081 2.966 0.003

age30-44 1.483 0.853 2.577 1.397 0.162

age45+ 1.264 0.741 2.159 0.859 0.390

age5-9 0.658 0.378 1.144 -1.484 0.138

sexMale 0.927 0.668 1.286 -0.457 0.648

housesun-dried bricks or 1.093 0.646 1.848 0.332 0.740

pounded mud

housetemporary shelter 1.259 0.484 3.275 0.472 0.637

housewattle and daub 1.422 0.905 2.235 1.528 0.127

school6-8yr primary 0.470 0.301 0.736 -3.299 0.001

schoolnone 1.149 0.797 1.656 0.744 0.457

schoolsecondary/tertiary 0.113 0.015 0.862 -2.103 0.035

---------------------------------------------------------------------------

How many individuals were included in these analysis? 1186

Is it right to compare the results from the two models? No – because different numbers of individuals are used in each one.

Re-fit the unadjusted model, restricted to individuals with complete confounders:

*# Fit logistic model of `bcg` against `d` removing individuals with any missing*

*# data*

glm(

d ~ bcg + age + sex + house + school, data = drop\_na(kar2),

family = "binomial"

) |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

> summ(glm(d~bcg, data = subset(df2,house\_na!=1 & school\_na!=1), family="binomial"), confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 1186

Dependent Variable: d

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(1) = 69.938, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.091

Pseudo-R² (McFadden) = 0.060

AIC = 1100.779, BIC = 1110.936

Standard errors: MLE

---------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

----------------- ----------- ------- ------- --------- -------

(Intercept) 0.370 0.313 0.437 -11.721 0.000

bcgPresent 0.245 0.170 0.352 -7.579 0.000

What do conclude about the association between bcg and leprosy?

The unadjusted association between bcg and leprosy is 0.245 (95%CI: 0.170-0.352). This suggests that individuals who have had a BCG vaccination have a 75% lower odds of leprosy. This association slightly attenuates to the null, to a 65% lower odds, after adjustment for potential confounders.

1. **Basic multiple imputation**

Performing full multiple imputation is beyond the scope of our teaching today. Generally,

1. You must include at least all variables in your analysis model (including the outcome variables) in the imputation model.
2. Select an appropriate imputation model for each variable which needs imputing (eg, logistic for binary variables, linear regression for normal quantitative variables, multinomial [polytomous logistic regression] for categorical variables, ordered logistic model (proportional odds model) for ordered categorical variables).
3. Select number of imputations (default is m=5, higher is always better, rule of thumb: m=% missing data).
4. Careful consideration is needed for: inclusion of survival outcomes, effect-modification, composite variables, non-normal quantitative variables, data are missing not at random, proxy/surrogate variables.

We can simply impute the school and house variables in the karonga dataset as follows. We will assume that school is an ordered categorical variable, house is a categorical variable, and there are no missing values in other variables. We need to set a seed number (which can be any whole number) because the approach uses a random number generator. If the seed is not selected or set to a different number, your results will differ each time.

# Multiple imputation using the `mice` package

imputed <-

mice::mice(

kar2[,2:7], method = c("", "", "", "", "polr", "polyreg"),

m = 5, seed = 2022

)

Explore the structure of the “imputed” object.

summary(imputed)

attributes(imputed

1. **Analysing a multiply imputed dataset**

Analyse each imputed dataset separately with the command:

*# Run logistic regression across five datasets*

fit <-

mice::with(

data = imputed,

exp = glm(d ~ age + sex + bcg + house + school, family = "binomial")

)

What does the output provide?

model\_fit

The model\_fit provides the output for the fitted model for each imputed dataset. For example, for imputation 1, the model fit is:

analyses :

[[1]]

Call: glm(formula = d ~ age + sex + bcg + house + school, family = "binomial")

Coefficients:

(Intercept) age15-19 age20-24

-1.36319 0.00175 0.27511

age25-29 age30-44 age45+

0.90161 0.45812 0.22283

age5-9 sexMale bcgPresent

-0.53135 -0.01566 -1.01172

housesun-dried bricks or pounded mud housetemporary shelter housewattle and daub

0.17595 0.26694 0.43852

school6-8yr primary schoolnone schoolsecondary/tertiary

-0.78402 0.22500 -1.20306

Degrees of Freedom: 1369 Total (i.e. Null); 1355 Residual

Null Deviance: 1371

Residual Deviance: 1228 AIC: 1258

Using Rubin’s rules as defined in the lecture, combine the coefficients [NB not the odds ratios] for bcgPresent to obtain the pooled coefficient and convert this to an odds ratio (OR=exp(coef)).

Coef = ((-1.012)+(-1.044)+(-1.044)+(-1.038)+ (-1.046))/5 = -1.0367

OR = exp(coef) = 0.355

The R command pool(model\_fit) applies Rubin’s rules to the imputation specific model fits to obtain a single pooled result. How do the results compare with those you calculated by hand, and the analysis on complete-cases?

Class: mipo m = 5

term m estimate ubar b t dfcom df riv

1 (Intercept) 5 -1.28885756 0.07789418 2.202907e-03 0.08053766 1355 967.57096 0.033936916

2 age15-19 5 -0.01786370 0.09451874 1.780382e-04 0.09473239 1355 1347.63978 0.002260355

3 age20-24 5 0.25457726 0.10828017 1.479837e-04 0.10845775 1355 1349.56734 0.001640009

4 age25-29 5 0.91677700 0.09383295 9.210594e-05 0.09394347 1355 1350.78086 0.001177914

5 age30-44 5 0.51727803 0.06579607 1.746926e-03 0.06789239 1355 999.01016 0.031860732

6 age45+ 5 0.29130263 0.06300206 1.681538e-03 0.06501990 1355 996.46314 0.032028250

7 age5-9 5 -0.47646178 0.07104876 9.827512e-04 0.07222806 1355 1222.47848 0.016598480

8 sexMale 5 -0.05770271 0.02278934 5.937058e-04 0.02350178 1355 1008.12008 0.031262295

9 bcgPresent 5 -1.03677772 0.03799045 2.049089e-04 0.03823634 1355 1325.87578 0.006472434

10 housesun-dried bricks or pounded mud 5 0.13325344 0.05764026 3.771784e-03 0.06216640 1355 471.17752 0.078523942

11 housetemporary shelter 5 0.12149715 0.21740243 1.379551e-02 0.23395704 1355 488.49787 0.076147313

12 housewattle and daub 5 0.32515765 0.04407995 4.451507e-03 0.04942176 1355 266.71516 0.121184549

13 school6-8yr primary 5 -0.59044509 0.04018659 1.267826e-02 0.05540050 1355 50.32078 0.378581778

14 schoolnone 5 0.10533972 0.03003987 5.906339e-03 0.03712748 1355 99.75951 0.235939996

15 schoolsecondary/tertiary 5 -1.66477184 0.59507708 2.053823e-01 0.84153585 1355 44.46798 0.414162780

lambda fmi

1 0.032823005 0.034816011

2 0.002255257 0.003732697

3 0.001637324 0.003113572

4 0.001176528 0.002652134

5 0.030876969 0.032811327

6 0.031034277 0.032973249

7 0.016327469 0.017932837

8 0.030314591 0.032232633

9 0.006430811 0.007926164

10 0.072806860 0.076717603

11 0.070759191 0.074540452

12 0.108086174 0.114699921

13 0.274616845 0.301825119

14 0.190899232 0.206646695

15 0.292867827 0.322661899

>

The pooled estimate from the multiple imputation approach is identical to the hand-calculation. It is also very similar to the result from the complete-cases.