**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)))

*# 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))

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)

)

kar2 |>

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

purrr::map(

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

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

)

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)

house\_assoc <-

glm(

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

) |>

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

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)

How many individuals were included in the analysis?

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)

How many individuals were included in these analysis?

Is it right to compare the results from the two models?

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)

What do conclude about the association between bcg and leprosy?

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

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)).

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?