Calculating weights

2024-03-14

# Loading R packages

setwd("C:/Users/swj88/Documents/Github/maicplus")  
devtools::load\_all()  
#> ℹ Loading maicplus  
#library(maicplus)

Additional suggested packages for this vignette:

library(dplyr) # this is used for data merging/cleaning  
  
#library(survival)  
#library(survminer) # ggsurvfit  
#library(sandwich) # robust standard error

# Preprocessing

## Preprocessing IPD

In this example scenario, age, sex, Eastern Cooperative Oncology Group (ECOG) performance status, smoking status, and number of previous treatments have been identified as imbalanced prognostic variables/effect modifiers.

This example reads in and combines data from three standard simulated data sets (adsl, adrs and adtte) which are saved as ‘.csv’ files.

adsl <- read.csv(system.file("extdata", "adsl.csv",  
 package = "maicplus",  
 mustWork = TRUE  
))  
adrs <- read.csv(system.file("extdata", "adrs.csv",  
 package = "maicplus",  
 mustWork = TRUE  
))  
adtte <- read.csv(system.file("extdata", "adtte.csv", package = "maicplus", mustWork = TRUE))  
  
# Data containing the matching variables  
adsl <- adsl %>%  
 mutate(SEX\_MALE = ifelse(SEX == "Male", 1, 0)) %>%  
 mutate(AGE\_SQUARED = AGE^2)  
  
# Could use built-in function for dummizing variables  
# adsl <- dummize\_ipd(adsl, dummize\_cols=c("SEX"), dummize\_ref\_level=c("Female"))  
  
# Response data  
adrs <- adrs %>%  
 dplyr::filter(PARAM == "Response") %>%  
 transmute(USUBJID, ARM, RESPONSE = AVAL, PARAM)  
  
# Time to event data (overall survival)  
adtte <- adtte %>%  
 dplyr::filter(PARAMCD == "OS") %>%  
 mutate(EVENT = 1 - CNSR) %>%  
 transmute(USUBJID, ARM, TIME = AVAL, EVENT)  
  
# Rename adsl as ipd  
ipd <- adsl   
head(ipd)  
#> X USUBJID ARM AGE SEX SMOKE ECOG0 N\_PR\_THER SEX\_MALE AGE\_SQUARED  
#> 1 1 1 A 45 Male 0 0 4 1 2025  
#> 2 2 2 A 71 Male 0 0 3 1 5041  
#> 3 3 3 A 58 Male 1 1 2 1 3364  
#> 4 4 4 A 48 Female 0 1 4 0 2304  
#> 5 5 5 A 69 Male 0 1 4 1 4761  
#> 6 6 6 A 48 Female 0 1 4 0 2304

## Preprocessing aggregate data

There are two ways of specifying aggregate data. One approach is to read in aggregate data using an excel spreadsheet. In the spreadsheet, possible variable types include mean, median, or standard deviation for continuous variables and count or proportion for binary variables. The naming should be followed by these suffixes accordingly: \_COUNT, \_MEAN, \_MEDIAN, \_SD, \_PROP. Then, process\_agd will convert the count into proportions.

Other way is to define data frame of aggregate data in R. If you do it this way, \_COUNT prefix should not be specified and only proportion is allowed for binary variables. Other suffix names would be the same as the first approach.

Possible missingness in the binary variables should be accounted for by subtracting the denominator by the missing count i.e. proportion = count / (N - missing).

# Through excel spreadsheet  
# target\_pop <- read.csv(system.file("extdata","aggregate\_data\_example\_1.csv", package = "maicplus", mustWork = TRUE))  
# agd <- process\_agd(target\_pop)  
  
# Second approach by defining a data frame in R  
agd <- data.frame(  
 STUDY = "Lung study",  
 ARM = "Total",  
 N = 300,  
 AGE\_MEAN = 51,  
 AGE\_MEDIAN = 49,  
 AGE\_SD = 3.25,  
 SEX\_MALE\_PROP = 147 / 300,  
 ECOG0\_PROP = 0.40,  
 SMOKE\_PROP = 58 / (300 - 5),  
 N\_PR\_THER\_MEDIAN = 2  
)

## Centering IPD

In the statistical theory section, we briefly explained why it is useful to transform IPD by subtracting the aggregate data means when performing optimization. The function center\_ipd centers the IPD using the aggregate data means.

#### prepare data  
ipd\_centered <- center\_ipd(ipd = ipd, agd = agd)  
head(ipd\_centered)  
#> X USUBJID ARM AGE SEX SMOKE ECOG0 N\_PR\_THER SEX\_MALE AGE\_SQUARED  
#> 1 1 1 A 45 Male 0 0 4 1 2025  
#> 2 2 2 A 71 Male 0 0 3 1 5041  
#> 3 3 3 A 58 Male 1 1 2 1 3364  
#> 4 4 4 A 48 Female 0 1 4 0 2304  
#> 5 5 5 A 69 Male 0 1 4 1 4761  
#> 6 6 6 A 48 Female 0 1 4 0 2304  
#> AGE\_CENTERED AGE\_MEDIAN\_CENTERED AGE\_SQUARED\_CENTERED SEX\_MALE\_CENTERED  
#> 1 -6 -0.5 -586.5625 0.51  
#> 2 20 0.5 2429.4375 0.51  
#> 3 7 0.5 752.4375 0.51  
#> 4 -3 -0.5 -307.5625 -0.49  
#> 5 18 0.5 2149.4375 0.51  
#> 6 -3 -0.5 -307.5625 -0.49  
#> ECOG0\_CENTERED SMOKE\_CENTERED N\_PR\_THER\_MEDIAN\_CENTERED  
#> 1 -0.4 -0.1966102 0.5  
#> 2 -0.4 -0.1966102 0.5  
#> 3 0.6 0.8033898 -0.5  
#> 4 0.6 -0.1966102 0.5  
#> 5 0.6 -0.1966102 0.5  
#> 6 0.6 -0.1966102 0.5

### How to handle standard deviation aggregate summary

As described in NICE DSU TSD 18 Appendix D [1], balancing on both mean and standard deviation for continuous variables may be considered in some cases. If a standard deviation is provided in the comparator population, preprocessing is done so that in the target population, is calculated using the variance formula . This in the target population is then matched with the calculated in the internal IPD.

### How to handle median aggregate summary

If a median is provided, IPD is preprocessed to categorize the variable into a binary variable. All the values in the IPD that are higher than the comparator population median is assigned a value of 1. Conversely, all values that are lower are assigned a value of 0. Comparator population median is replaced by 0.5 to adjust to the categorization in the IPD data. The newly created IPD binary variable is matched so that the proportion is 0.5.

# Calculating weights

We use the centered IPD and use the function estimate\_weights to calculate the weights. Before running this function, we need to specify the column that are centered, i.e. covariates that will be used in the optimization.

# list variables that are going to be used to match  
centered\_colnames <- c("AGE", "AGE\_SQUARED", "SEX\_MALE", "ECOG0", "SMOKE", "N\_PR\_THER\_MEDIAN")  
centered\_colnames <- paste0(centered\_colnames, "\_CENTERED")  
  
weighted\_data <- estimate\_weights(  
 data = ipd\_centered,  
 centered\_colnames = centered\_colnames  
)  
#> initial value 500.000000   
#> iter 10 value 215.753747  
#> iter 20 value 199.844445  
#> final value 199.842237   
#> converged  
  
# Alternatively, you can specify the numeric column locations for centered\_colnames  
# weighted\_data <- estimate\_weights(ipd\_centered, centered\_colnames = c(14, 16:20))

Following the calculation of weights, it is necessary to determine whether the optimization procedure has worked correctly and whether the weights derived are sensible.

The approximate effective sample size is calculated as:

A small ESS, relative to the original sample size, is an indication that the weights are highly variable and that the estimate may be unstable. This often occurs if there is very limited overlap in the distribution of the matching variables between the populations being compared.

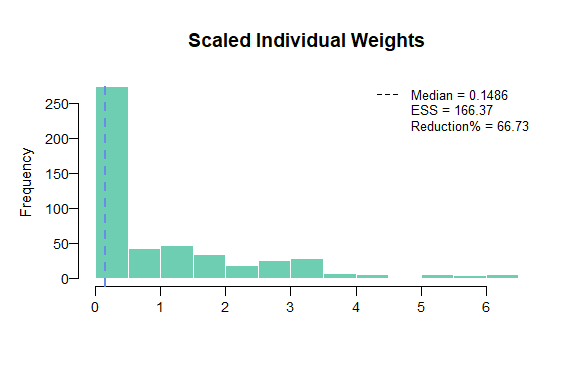
In this example, the ESS reduction is 66.73% of the total number of patients in the intervention arm (500 patients in total). As this is a considerable reduction, estimates using this weighted data may be unreliable.

weighted\_data$ess  
#> [1] 166.3675

Also, it is useful to visualize the weights using a histogram to check if there are any extreme weights. Scaled weights are weights that are relative to the original unit weights of each individual. Scaled weights are calculated as

.

plot(weighted\_data)



#ggplot option is also available  
#plot(weighted\_data, ggplot = TRUE, bin\_col = "black", vline\_col = "red")

Another check would be to look at whether the weighted summary of covariates in internal IPD match the external aggregate data summary.

outdata <- check\_weights(weighted\_data, agd)  
outdata  
#> covariate match\_stat internal\_trial internal\_trial\_after\_weighted  
#> 1 AGE Mean 59.850 51.00  
#> 2 AGE SD 9.011 3.25  
#> 3 SEX\_MALE Prop 0.380 0.49  
#> 4 ECOG0 Prop 0.410 0.40  
#> 5 SMOKE Prop 0.320 0.20  
#> 6 N\_PR\_THER Median 3.000 2.00  
#> external\_trial sum\_centered\_IPD\_with\_weights  
#> 1 51.00 0.0001  
#> 2 3.25 0.0125  
#> 3 0.49 0.0000  
#> 4 0.40 0.0000  
#> 5 0.20 0.0000  
#> 6 2.00 0.0000

# References

[1] Phillippo D, Ades T, Dias S, Palmer S, Abrams KR, Welton N. NICE DSU technical support document 18: Methods for population-adjusted indirect comparisons in submissions to NICE: Appendix d: Worked example of MAIC and STC 2016.