**Supplementary analytic data pre-processing R code for main supplement of Phillips et al. 2023, Practical considerations for specifying a super learner**

**Contents**

R code for example 6 analytic dataset preprocessing (page 2)

R code for example 7 analytic dataset preprocessing (page 3)

R session information (page 4)

##############################################################################

# EXAMPLE 6 ANALYTIC DATAET PREPROCESSING R CODE

# The file "IST dataset, corrected (csv)" was first downloaded from the

# database (https://datashare.ed.ac.uk/handle/10283/124) and stored in the

# user's "Downloads" folder.

##############################################################################

library(tidyverse)

d <- read.csv("~/Downloads/IST\_corrected.csv")

####### set 14-day mortality to "N" if recurrent stroke occurred after 14 days

d$DDEAD <- as.character(d$DDEAD)

d$DDEAD[(d$DDEAD != "N" & d$DDEADD > 14)] <- "N"

####### specify covariates and outcome and filter data to include relevant columns

X <- c("RDELAY", "RCONSC", "SEX", "AGE", "RSLEEP", "RATRIAL", "RCT",

"RVISINF", "RHEP24", "RASP3", "RSBP", "RDEF1", "RDEF2",

"RDEF3", "RDEF4", "RDEF5", "RDEF6", "RDEF7", "RDEF8", "STYPE",

"RXHEP", "RXASP")

Y <- "DDEAD"

d <- dplyr::select(d, c(X,Y))

####### find missing data

d[d == ""] <- NA

d[d == "NULL"] <- NA

d[d == " "] <- NA

d[d == " "] <- NA

d[d == "[]"] <- NA

colSums(is.na(d))

####### create binary indicator columns denoting missingness and impute with U

all.equal(is.na(d$RATRIAL), is.na(d$RASP3)) # missingness pattern is the same!

d$MISSING\_RATRIAL\_RASP3 <- rep(0, nrow(d))

d$MISSING\_RATRIAL\_RASP3[is.na(d$RATRIAL)] <- 1

d$MISSING\_RHEP24 <- rep(0, nrow(d))

d$MISSING\_RHEP24[is.na(d$RHEP24)] <- 1

d$RATRIAL[is.na(d$RATRIAL)] <- "U"

d$RASP3[is.na(d$RASP3)] <- "U"

d$RHEP24[is.na(d$RHEP24)] <- "U"

####### remove the observations with missing outcome (a negligible proportion)

d <- d[!(is.na(d$DDEAD) | d$DDEAD == "U"),]

d$DDEAD[d$DDEAD == "Y"] <- 1 # change “Y” (14-day morality occurred) to 1

d$DDEAD[d$DDEAD == "N"] <- 0 # change “N” (14-day morality didn't occur) to 0

write.csv(d, file = "~/data\_example6.csv", row.names = FALSE)

#############################################################################

# EXAMPLE 7 ANALYTIC DATASET PREPROCESSING R CODE

# File "Additional File 1: Data set from acupuncture headache trial (XLS 254 KB)"

# was first downloaded from Vickers AJ. Whose data set is it anyway? Sharing

# raw data from randomized trials. Trials. 2006 Dec;7(1):1-6.

# https://doi.org/10.1186/1745-6215-7-15. This downloaded data was stored in

# the user's "Downloads" folder with name "13063\_2006\_152\_MOESM1\_ESM.xls".

##############################################################################

library(xlsx)

library(data.table)

d <- read.xlsx("~/Downloads/13063\_2006\_152\_MOESM1\_ESM.xls", sheetName = "Data")

data.table::setDT(d)

####### define nodes in observed data

# W: baseline covariates, variables measured before or at assignment of A

W <- c("age", "sex", "migraine", "chronicity", "pk1", "f1", "pf1", "rlp1",

"rle1", "ef1", "ewb1", "sf1", "p1", "gen1", "hc1", "painmedspk1",

"prophylacticdose1","prophmqs1","allmedsbaseline")

A <- "group" # treatment, randomization to acupuncture (A = 1) or usual care (A = 0)

Delta <- "completer" # missingness indicator (if Y observed Delta=1 and if not Delta=0)

Y <- "pk5" # Y: outcome, 12-month headache score

####### subset data to nodes of interest

d <- d[, c(Y, Delta, A, W), with = FALSE]

colnames(d)[1:3] <- c("Y", "Delta", "A")

####### fix prophylacticdose1 column: it should be 0, not NA, when prophmqs1 = 0

d$prophylacticdose1 <- ifelse(d$prophmqs1 == 0, 0, d$prophylacticdose1)

####### examine missingness of all data

colSums(is.na(d))

####### create indicators of whether W values are missing before imputing them

# ef1, ewb1, sf1, gen1 and hc1 missingness in just one row, and it's the same

d[which(is.na(d$hc1)),]

d$missing\_ef1\_ewb1\_sf1\_gen1\_hc1 <- ifelse(is.na(d$ef1), 1, 0)

d$missing\_p1 <- ifelse(is.na(d$p1), 1, 0)

d$missing\_rle1 <- ifelse(is.na(d$rle1), 1, 0)

d$missing\_pf1 <- ifelse(is.na(d$pf1), 1, 0)

d$missing\_rlp1 <- ifelse(is.na(d$rlp1), 1, 0)

####### imputation

# all W with NA are continuous, we impute with median

imputedW <- apply(d[ , -c("Y", "Delta", "A"), with = F], 2, function(W){

ifelse(is.na(W), median(W, na.rm = T), W)

})

sum(is.na(imputedW)) == 0

####### put it all together

d <- data.table(Y = d$Y, Delta = d$Delta, A = d$A, imputedW)

write.csv(d, file = "~/data\_example7.csv", row.names = FALSE)

##############################################################################

# R SESSION INFORMATION FOR ANALYTIC DATAET PREPROCESSING R CODE

##############################################################################

R version 4.1.2 (2021-11-01)

Platform: x86\_64-apple-darwin17.0 (64-bit)

Running under: macOS Monterey 12.2.1

Matrix products: default

LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib

locale:

[1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8

attached base packages:

[1] stats graphics grDevices utils datasets methods base

other attached packages:

[1] forcats\_0.5.1 stringr\_1.4.0 dplyr\_1.0.8 purrr\_0.3.4 readr\_2.1.2

[6] tidyr\_1.2.0 tibble\_3.1.6 ggplot2\_3.3.5 tidyverse\_1.3.1 data.table\_1.14.2

[11] xlsx\_0.6.5

loaded via a namespace (and not attached):

[1] Rcpp\_1.0.8.3 cellranger\_1.1.0 pillar\_1.7.0 compiler\_4.1.2 dbplyr\_2.1.1 tools\_4.1.2

[7] lubridate\_1.8.0 jsonlite\_1.8.0 lifecycle\_1.0.1 gtable\_0.3.0 pkgconfig\_2.0.3 rlang\_1.0.2

[13] reprex\_2.0.1 rstudioapi\_0.13 DBI\_1.1.2 cli\_3.2.0 haven\_2.4.3 rJava\_1.0-6

[19] xml2\_1.3.3 withr\_2.5.0 httr\_1.4.2 fs\_1.5.2 generics\_0.1.2 xlsxjars\_0.6.1

[25] vctrs\_0.3.8 hms\_1.1.1 grid\_4.1.2 tidyselect\_1.1.2 glue\_1.6.2 R6\_2.5.1

[31] fansi\_1.0.2 readxl\_1.3.1 tzdb\_0.2.0 modelr\_0.1.8 magrittr\_2.0.2 backports\_1.4.1

[37] scales\_1.1.1 ellipsis\_0.3.2 rvest\_1.0.2 assertthat\_0.2.1 colorspace\_2.0-3 utf8\_1.2.2

[43] stringi\_1.7.6 munsell\_0.5.0 broom\_0.7.12 crayon\_1.5.0