Manjkajoče vrednosti

A. Blejec

1. junij 2011

Kazalo

1 MCAR

Postavimo konstatnte za opis velikosti problema. Imejmo N enot, ki so opisane z n spremenljivkami. Nastavimo tudi delež enot z manjkajočimi vrednostmi `pMiss` in dopustno število spremnljivk, ki lahko manjkajo pri posamezni enoti (`vMiss`)

```
> N <- 1000
> n <- 12
> pMiss <- 0.03
> vMiss <- 1
```

Pripravimo testne podatke in jih opremimo z imeni enot in spremnljivk

```
> set.seed(123)
> X <- matrix(round(rnorm(N * n) * 10), N, n)
> dimnames(X) <- list(paste("E", 1:N, sep = ""), paste("x",
+     1:n, sep = ""))
> head(X)
    x1  x2 x3  x4  x5  x6 x7  x8  x9 x10 x11 x12
E1  -6 -10 -5  -2   2  -5 -7 -16   5  19  24   3
E2  -2 -10  2  -3   7  11 10   4  18  -6  -2  -4
E3  16   0 -5 -14   7 -11 -7  19 -17  -6   9   3
E4   1  -1 12  -7 -13  15 -1   6   3 -10  -6 -12
E5   1 -25  2  26 -20   9  6  17  -3  27   2   8
E6  17  10 -6   0  22   3 -6  -1  -4  -7  11   3
```

Zaradi ponovljivosti simulacij, nastavimo generator slučajnih števil na neko vrednost.

```
> set.seed(1234)
```

Najprej z enostavnim slučajnostnim vzorcem izberemo enote, ki bodo imele manjkajoče vrednosti. Indeksi izbranih enot bodo v `idMiss`

```
> idMiss <- sample(1:N, N * pMiss)
> nMiss <- length(idMiss)
```

Za vsako izbrano enoto določimo število manjkajočih vrednosti (`kolkMiss`)

```
> kolkMiss <- sapply(idMiss, function(x) sample(1:vMiss,
+     1))
> kolkMiss
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Zdaj za vsako enoto določim, katere spremenljivke bodo imele manjkajoče vrednosti

```
> varMiss <- lapply(kolkMiss, function(x) sample(1:n, x))
> head(varMiss)
[[1]]
[1] 11

[[2]]
[1] 1

[[3]]
[1] 4

[[4]]
[1] 1

[[5]]
[1] 3

[[6]]
[1] 9
```

Vse določitve so slučajne (nadziram jih lahko s postavitvami generatorja slučajnih števil (`set.seed()`).

Končno vpišemo manjkajoče vrednosti na ustrezna mesta v matriki podatkov Za to pripravimo matriko indeksov enot in indeksov spremnljivk.

```
> ids <- cbind(rep(idMiss, kolkMiss), unlist(varMiss))
> dimnames(ids)[[2]] <- c("row", "col")
> head(ids)
     row col
[1,] 114  11
[2,] 622   1
[3,] 609   4
[4,] 999   1
[5,] 858   3
[6,] 638   9
> tail(ids)
      row col
[25,] 214   3
[26,] 791  11
[27,] 513   5
[28,] 890   4
[29,] 809   2
[30,]  45  11
```

Vstavimo manjkajoče podatke na prava mesta

```
> Xmiss <- X
> Xmiss[ids] <- NA
```

Poglejmo, kaj dobimo

```
> Xmiss[idMiss, ]
```

```
      x1  x2  x3  x4  x5  x6  x7  x8  x9 x10 x11 x12
E114  -1  -9   3   3  -9 -13  -6  -8  12   0  NA   3
E622  NA  33   3  -6   0   6  -2   2   4  -2  -3   8
E609  -3  17   4  NA   6 -16  -1   9   9   6   6   9
E999  NA   5  12  -7  13   2  -6   5   1  11   9  -3
E858  -4   6  NA   2  -9   4   4  -6 -21 -19   3   1
E638  11  11  -4   7  32   1 -10   2  NA   2   1  15
E10   -4  -4  -6  NA  11  -3 -12   8  -8   3  -3 -16
E231  20  -9 -10   0  -1   4  NA   7  -8   9  -6   4
E661  NA  10   7 -10  -7   4  -8  -1  -5   8   5 -10
E510  -1   9  16  15 -21   8  NA -14   1   7  10  -8
E687  -2  NA  -5  -4  -8 -11 -14  10   2 -11  -6 -11
E539  -6   2   9  -8  12  19  -8  16  -5   3  NA   7
E280  NA  -1  -2 -11  -6   3  19   1  17  -7  -2  14
E912 -12 -10   6   8  -8  11  22  18 -22  NA   6   1
E289  -2  NA   6   8   4   8   3 -13  12  -7   4  -2
E825  -3   3  10  -5   7  10  NA   1   8  -3  17  14
E282   7  -1  -3 -17  NA -14  12  -9  11   1   1  -6
E263  NA  -3  -8   3  -6   2  -2   7  -9   9  -5   8
E184  -9  -5  11  NA  13 -13  -4   5  -8  -9  -6   8
E228  -7  12  -6  -6   0  20  16  17  NA   2  -8  16
E311  13  13   1  -2 -20  17 -11  14   1   2  12  NA
E297  19   3   6  -2 -13  NA  10  -9   5   0  14  -4
E156  -3  NA   0  -8 -10  -5   4   2  38 -16   3   3
E40   -4  18  -3  -2  14  18  NA  -1  -4  -9   5   2
E214  -5 -12  NA  20   4   0  -6   5  -3   5   2   5
E791   4   6   8 -18   3 -13 -11 -13   2   2  NA   1
E513   9   8   4  -3  NA -14 -10  14  19   0  -9  -1
E890   4  -8  -7  NA  -8   4  -8   2  -7 -15   3  10
E809   3  NA  -2   1 -15  20  -2  11   8 -14  -4  -8
E45   12   5   6   8  -2  -8  -7 -10   3   8  NA -13
```

Za nadzor poglejmo še indekse

```
> head(ids)
```

```
     row col
[1,] 114  11
[2,] 622   1
[3,] 609   4
[4,] 999   1
[5,] 858   3
[6,] 638   9
```

Delež manjkajočih za posamezne spremenljivke

```
> naByVar <- apply(Xmiss, 2, function(x) sum(is.na(x))/length(x))
> naByVar
   x1    x2    x3    x4    x5    x6    x7    x8    x9   x10   x11
0.005 0.004 0.002 0.004 0.002 0.001 0.004 0.000 0.002 0.001 0.004
  x12
0.001
> naTot <- sum(naByVar)
> naTot
[1] 0.03
```

Če dopustimo pri posamezni enot le eno manjkajočo vrednost ($m = 1$) je delež simuliranih manjkajočih vrednosti (`naTot`) približno enak predvidenemu deležu enot z manjkajočimi vrednostmi (`pMiss`).

1.1 Funkcija `naMCAR()`

Vse tole bi kazalo zapreti v funkcijo.

```
> naMCAR <- function(X, pMiss, vMiss = 1) {
+     N <- dim(X)[1]
+     n <- dim(X)[2]
+     idMiss <- sample(1:N, N * pMiss)
+     nMiss <- length(idMiss)
+     kolkMiss <- sapply(idMiss, function(x) sample(1:vMiss,
+         1))
+     kolkMiss
+     varMiss <- lapply(kolkMiss, function(x) sample(1:n,
+         x))
+     head(varMiss)
+     ids <- cbind(rep(idMiss, kolkMiss), unlist(varMiss))
+     dimnames(ids)[[2]] <- c("row", "col")
+     Xmiss <- X
+     Xmiss[ids] <- NA
+     return(Xmiss)
+ }
```

Primer uporabe

```
> set.seed(234)
> N <- 10
> n <- 4
> X <- matrix(round(runif(N * n) * 10), N, n)
> dimnames(ids)[[2]] <- c("row", "col")
> X
      [,1] [,2] [,3] [,4]
 [1,]    7    6    5    6
 [2,]    8    5    6    3
 [3,]    0    6    9    2
 [4,]    8    6    6    7
 [5,]    1    0    5    5
 [6,]    6    4    4    7
 [7,]    9    3    7    7
 [8,]    7    7    2    2
 [9,]    9    1    8    5
[10,]    3    9    9    4
> Xmiss <- naMCAR(X, pMiss = 0.5, vMiss = 3)
> Xmiss
      [,1] [,2] [,3] [,4]
 [1,]    7    6    5    6
 [2,]   NA   NA    6    3
 [3,]    0    6    9    2
 [4,]    8    6    6   NA
 [5,]    1    0    5    5
 [6,]   NA   NA   NA    7
 [7,]   NA    3    7   NA
 [8,]    7    7    2    2
 [9,]    9    1    8   NA
[10,]    3    9    9    4
```

1.2 Dodaten vir za MCAR

Za način nadzora deleža manjkajočih vrednosti po posameznih spremenljivkah je nekaj napisanega v 'Multivariatne metode in manjkajoči podatki' (mva07.pdf) na spletni strani http://ablejec.nib.si/R/#MVA. Tam je tudi omenjenih nekaj uporabnih funkcij.

2 NMAR

Za eno spremenljivko

```
> x <- rnorm(N)
> pMiss <- 0.03
> meja <- quantile(x, 1 - pMiss)
> which(x > meja)
[1] 3
> x[which(x > meja)] <- NA
> length(which(is.na(x)))
[1] 1
```

SessionInfo

Windows XP (build 2600) Service Pack 3

- R version 2.10.0 (2009-10-26), `i386-pc-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250`, `LC_CTYPE=Slovenian_Slovenia.1250`, `LC_MONETARY=Slovenian_Slovenia.1250`, `LC_NUMERIC=C`, `LC_TIME=Slovenian_Slovenia.1250`
- Base packages: base, datasets, graphics, grDevices, methods, splines, stats, utils
- Other packages: Hmisc 3.7-0, patchDVI 1.5, survival 2.35-8
- Loaded via a namespace (and not attached): cluster 1.12.1, grid 2.10.0, lattice 0.18-3, tools 2.10.0

Project path: `C:/_Y/R/!KrNeki`

View as vignette

Project files can be viewed by pasting this code to R console:

```
> projectName <-"!KrNeki";  mainFile <-"missingValues"


> commandArgs()
> library(tkWidgets)
> openPDF(file.path(dirname(getwd()), "doc", paste(mainFile,
+     "PDF", sep = ".")))
> viewVignette("viewVignette", projectName, file.path("../doc",
+     paste(mainFile, "RNW", sep = ".")))
```