Imputation_missForest

Xinyi Lin

7/30/2019

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(missForest)
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
## Loading required package: foreach
```

```
##
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
```

```
## Loading required package: itertools
```

```
## Loading required package: iterators
```

Import data

```
library(haven)
training <- read_dta("training.dta")
```

Creat the variable

Notation:

Variable name: student_r

Created from: occup1_r, occup2_r

Label: student_r — yes-is a student, no-not a student, NA-occup1_r and occup2_r are NAs

```
student_data = training %>%
  as_data_frame() %>%
  mutate(student_r = ifelse(occup1_r == 8 | occup2_r == 8, "yes", "no"),
         student_r = ifelse(is.na(student_r) & occup1_r == 20, "no", student_r))
head(student_data,10)
```

```
## # A tibble: 10 x 59
##    study_id curr_id visit area  female ageyrs_r religion_r educate_r
##    <chr>    <chr>   <dbl> <dbl> <dbl+>    <dbl> <dbl+lbl>  <dbl+lbl>
##  1 E060914  002009~    12 1     0           37 3          1
##  2 B039195  008602~     7 0     0           45 3          1
##  3 A063907  010081~    12 0     0           26 2          1
##  4 E105637  012038~    15 2     0           19 3          1
##  5 G059950  006256~    11 0     0           18 3          1
##  6 H036023  009034~    11 0     0           37 2          1
##  7 E122573  008019~    18 0     0           16 5          1
##  8 A105390  012023~    18 2     0           38 1          1
##  9 J045501  009034~     7 0     0           15 2          1
## 10 E012406  009108~     7 0     0           32 2          1
## # ... with 51 more variables: educyrs_r <dbl+lbl>, occup1_r <dbl>,
## #   occup2_r <dbl>, student <dbl+lbl>, evermarr_r <dbl+lbl>,
## #   currmarr_r <dbl+lbl>, eversex_r <dbl+lbl>, sexyear_r <dbl+lbl>,
## #   comm_num <dbl>, currrltn <dbl>, rltnlst1 <dbl>, rltnlst2 <dbl>,
## #   rltnlst3 <dbl>, rltnlst4 <dbl>, cndever1 <dbl>, cndever2 <dbl>,
## #   cndever3 <dbl>, cndever4 <dbl>, sexp1yr <dbl>, arvmed <dbl>,
## #   cuarvmed <dbl>, rltnage1 <dbl>, rltnyrs1 <dbl>, occup11 <dbl>,
## #   occup21 <dbl>, rltnhh1 <dbl>, rnyrcon1 <dbl>, rltnage2 <dbl>,
## #   rltnyrs2 <dbl>, occup12 <dbl>, occup22 <dbl>, rltnhh2 <dbl>,
## #   rnyrcon2 <dbl>, rltnage3 <dbl>, rltnyrs3 <dbl>, occup13 <dbl>,
## #   occup23 <dbl>, rltnhh3 <dbl>, rnyrcon3 <dbl>, rltnage4 <dbl>,
## #   rltnyrs4 <dbl>, occup14 <dbl>, occup24 <dbl>, rltnhh4 <dbl>,
## #   rnyrcon4 <dbl>, childmarr <dbl+lbl>, R1_18_comm <dbl>,
## #   hiv_prev <dbl+lbl>, SEScat <dbl+lbl>, agecat <dbl>, student_r <chr>
```

```
# need data cleaning in variables
impu_data = student_data %>%
  filter(ageyrs_r <= 19) %>%
  mutate(student_r = as.factor(student_r),
         visit = as.factor(visit),
         sex = ifelse(female==1, "female", "male")) %>%
  select(ageyrs_r, sex, student_r, area, educate_r, currmarr_r, sexp1yr, SEScat)
head(impu_data)
```

```
## # A tibble: 6 x 8
##   ageyrs_r sex   student_r area      educate_r currmarr_r sexp1yr SEScat
##      <dbl> <chr> <fct>     <dbl+lbl> <dbl+lbl> <dbl+lbl>    <dbl> <dbl+lbl>
## 1       19 male  no        2         1         0                1 0
```

```
## 2       18 male  yes       0         1         0                1 1
## 3       16 male  no        0         1         0               98 2
## 4       15 male  no        0         1         0               98 1
## 5       16 male  no        0         1         0               98 1
## 6       17 male  no        0         1         0                1 3
```

```
summary(impu_data)
```

```
##     ageyrs_r         sex            student_r         area
##  Min.   :15.00   Length:47690       no  :21817   Min.   :0.0000
##  1st Qu.:16.00   Class :character   yes :19665   1st Qu.:0.0000
##  Median :17.00   Mode  :character   NA's: 6208   Median :0.0000
##  Mean   :17.17                                   Mean   :0.2956
##  3rd Qu.:18.00                                   3rd Qu.:0.0000
##  Max.   :19.00                                   Max.   :2.0000
##
##    educate_r        currmarr_r        sexp1yr          SEScat
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.00   Min.   :0.000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 1.00   1st Qu.:1.000
##  Median :1.0000   Median :0.0000   Median : 2.00   Median :2.000
##  Mean   :0.9775   Mean   :0.1954   Mean   :42.68   Mean   :1.645
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:98.00   3rd Qu.:3.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :99.00   Max.   :3.000
##  NA's   :8        NA's   :4        NA's   :1392    NA's   :208
```

According to table above, there are 6208 missing values in `student_r` variables.

Imputation