Imputation

Xinyi Lin

7/27/2019

```
library(tidyverse)
library(missForest)
library(Hmisc)
library(mice)
library(ggplot2)
#library(VIM)
#library(rms)
```

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>     <dbl>
##  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>, occup1_r <dbl>,
## #   occup2_r <dbl>, student <dbl>, evermarr_r <dbl>, currmarr_r <dbl>,
## #   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>, R1_18_comm <dbl>, hiv_prev <dbl>, 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>     <dbl>      <dbl>   <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
```

```
md.pattern(impu_data, plot = F)
```

```
##       ageyrs_r sex area currmarr_r educate_r SEScat sexp1yr student_r
## 41313        1   1    1          1         1      1       1         1    0
## 4779         1   1    1          1         1      1       1         0    1
## 1378         1   1    1          1         1      1       0         0    2
## 160          1   1    1          1         1      0       1         1    1
## 36           1   1    1          1         1      0       1         0    2
## 12           1   1    1          1         1      0       0         0    3
## 5            1   1    1          1         0      1       1         1    1
## 1            1   1    1          1         0      1       1         0    2
## 2            1   1    1          1         0      1       0         0    3
## 4            1   1    1          0         1      1       1         1    1
##              0   0    0          4         8    208    1392      6208 7820
```

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

Imputation

MICE

```
mice_data = mice(impu_data,seed = 123)
```

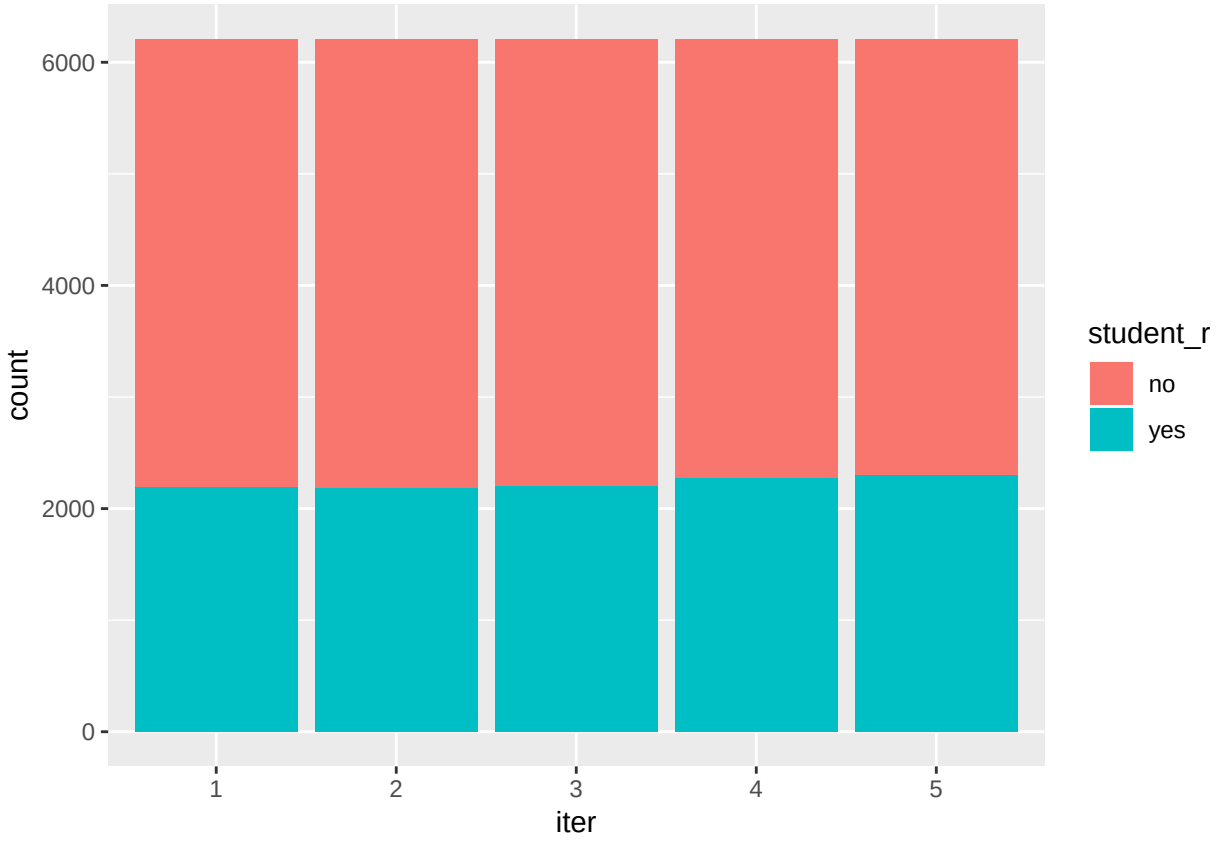
```
##
##  iter imp variable
##   1   1  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   1   2  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   1   3  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   1   4  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   1   5  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   2   1  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   2   2  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   2   3  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   2   4  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   2   5  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   3   1  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   3   2  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   3   3  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   3   4  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   3   5  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   4   1  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   4   2  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   4   3  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   4   4  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   4   5  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   5   1  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   5   2  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   5   3  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   5   4  student_r  educate_r  currmarr_r  sexp1yr  SEScat
##   5   5  student_r  educate_r  currmarr_r  sexp1yr  SEScat
```

```
## Warning: Number of logged events: 1
```

```
summary(mice_data)
```

```
## Class: mids
## Number of multiple imputations:  5
## Imputation methods:
##   ageyrs_r        sex  student_r       area  educate_r currmarr_r
##         ""         ""   "logreg"         ""       "pmm"      "pmm"
##    sexp1yr     SEScat
##      "pmm"      "pmm"
## PredictorMatrix:
##            ageyrs_r sex student_r area educate_r currmarr_r sexp1yr SEScat
## ageyrs_r          0   0         1    1         1          1       1      1
## sex               1   0         1    1         1          1       1      1
## student_r         1   0         0    1         1          1       1      1
## area              1   0         1    0         1          1       1      1
## educate_r         1   0         1    1         0          1       1      1
## currmarr_r        1   0         1    1         1          0       1      1
## Number of logged events:  1
##   it im dep     meth out
## 1  0  0     constant sex
```

```
#mice_data$imp$student_r
student_r = mice_data$imp$student_r %>%
  gather(key = iter, value = student_r, 1:5) %>%
  mutate(student_r = as.factor(student_r))
ggplot(data = student_r, aes(x = iter)) + geom_bar(aes(fill = student_r))
```



For each iteration `mice` function give same percentage of yes and no.

Add data back

```
completeData = complete(mice_data, 3)
```

```
## Warning in bind_rows_(x, .id): Vectorizing 'labelled' elements may not
## preserve their attributes

## Warning in bind_rows_(x, .id): Vectorizing 'labelled' elements may not
## preserve their attributes

## Warning in bind_rows_(x, .id): Vectorizing 'labelled' elements may not
## preserve their attributes

## Warning in bind_rows_(x, .id): Vectorizing 'labelled' elements may not
## preserve their attributes
```

```
head(completeData)
```

```
##   ageyrs_r  sex student_r area educate_r currmarr_r sexp1yr SEScat
## 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
```

Prediction

```
#fit = with(data = impu_data, expr = glm(student_r ~ ageyrs_r + sexp1yr + SEScat))
#combine = pool(fit)
#summary(combine)
```