Imputation

Xinyi Lin

7/30/2019

The goal of this file is to show the process and results of multiple imputation in the `student_r` variable.

Import data

Import "training.dta" as target dataset.

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

There are 260769 observations and 58 variables in original dataset. As we are interested in the `student_r` variables, I choose observations with age between 15-19 and variables related with the `student_r` variables which are ageyrs_r, sex, student_r, area, educate_r, currmarr_r, sexp1yr, SEScat. This subset of dataset is named as `impu_data`.

```
## # A tibble: 6 x 8
##   ageyrs_r sex   student_r area  educate_r currmarr_r sexp1yr SEScat
##      <dbl> <fct> <fct>     <fct> <fct>     <fct>        <dbl> <fct>
## 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               NA 2
## 4       15 male  no        0     1         0               NA 1
## 5       16 male  no        0     1         0               NA 1
## 6       17 male  no        0     1         0                1 3
```

```
##     ageyrs_r         sex        student_r    area      educate_r
##  Min.   :15.00   female:25767   no  :21817   0:36210   0   : 1073
##  1st Qu.:16.00   male  :21923   yes :19665   1: 8862   1   :46609
##  Median :17.00                  NA's: 6208   2: 2618   NA's:    8
##  Mean   :17.17
##  3rd Qu.:18.00
##  Max.   :19.00
##
##  currmarr_r      sexp1yr        SEScat
##  0   :38369   Min.   : 0.000   0   : 9887
##  1   : 9317   1st Qu.: 1.000   1   :10993
##  NA's:    4   Median : 1.000   2   :12708
##               Mean   : 1.158   3   :13894
##               3rd Qu.: 1.000   NA's:  208
##               Max.   :30.000
##               NA's   :21243
```

Above are summary of `impu_data`. There are 6208 missing values in `student_r` variables.

Imputation

missForest

First, I use `missForest` package to do multiple imputation. This package use random forest algorithm to impute data and will give the optimal imputation result with lowest estimated error.

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
```

Following are summary and head of the imputed data. This data is stored in file "impu_forest.csv".

```
##  ageyrs_r       sex        student_r   area      educate_r currmarr_r
##  15: 7580   female:25767   no :25502   0:36210   0: 1079   0:38372
##  16: 9201   male  :21923   yes:22188   1: 8862   1:46611   1: 9318
##  17: 9086                             2: 2618
##  18:11310
##  19:10513
##
##
##     sexp1yr      SEScat
##   1     :20097   0:10012
##   0     :16864   1:10999
##   2     : 3196   2:12710
##  17     : 1069   3:13969
##   3     :  957
##   9     :  747
##  (Other): 4760
```

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

Percentages of "yes" in different sexes, ages and rounds are shown as following.

sex	visit	15	16	17	18	19
female	1	52.14	30.34	20.54	10.90	4.39
female	2	59.91	56.16	35.47	15.98	6.37
female	3	61.24	51.99	31.19	18.46	9.36
female	4	67.86	55.87	28.77	14.25	2.76
female	6	59.71	44.12	22.98	15.57	7.42
female	7	64.25	48.52	35.28	18.42	8.79
female	8	71.00	52.36	33.33	18.71	12.53
female	9	74.60	62.87	42.65	21.47	12.19
female	10	75.61	60.75	44.93	25.46	15.71
female	11	81.46	69.80	50.00	31.40	18.27
female	12	83.87	70.52	52.16	34.81	17.36
female	13	84.78	77.89	57.72	36.68	25.95
female	14	88.54	78.55	62.46	43.27	21.84

sex	visit	15	16	17	18	19
female	15	85.32	73.86	54.91	30.39	17.60
female	16	88.01	77.67	57.48	33.15	14.12
female	17	83.18	80.99	58.53	34.48	12.36
female	18	82.00	75.79	56.06	39.44	25.87
male	1	56.54	47.31	33.19	19.67	19.48
male	2	78.48	74.25	52.28	34.25	33.33
male	3	67.15	69.18	54.32	28.38	31.60
male	4	74.62	66.28	46.08	22.65	14.39
male	6	74.58	59.19	48.05	25.52	23.10
male	7	71.76	66.32	43.38	33.33	18.22
male	8	75.91	62.50	54.15	37.10	25.35
male	9	82.69	70.30	65.09	41.41	31.62
male	10	75.00	66.91	55.29	48.47	27.23
male	11	88.18	76.71	57.82	40.87	46.34
male	12	86.31	74.91	53.75	40.78	36.61
male	13	88.28	73.52	60.70	44.80	31.86
male	14	85.82	76.86	64.60	55.52	39.53
male	15	87.10	75.57	66.29	48.17	35.60
male	16	84.18	70.55	61.31	48.24	33.33
male	17	83.39	78.28	59.49	51.31	23.35
male	18	78.90	68.10	55.62	39.50	38.59

Hmisc

I also use the `Hmisc` package to impute data. The `Hmisc` package uses additive semiparametric models to do multiple imputation. Following show the summary and head of first imputation results. This imputed data is stored in file "impu_himsc.csv".

```
## Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5
Iteration 6
Iteration 7
Iteration 8
```

```
##
## Imputed Values:
##
## X[[i]]
##        n  missing distinct
##     6208        0        2
##
## Value         no   yes
## Frequency   3969  2239
## Proportion 0.639 0.361
##
##
##  8 values imputed to 1
##
```

```
##
## Imputed Values:
##
## [1] 0 0 0 1
## Levels: 0 1
##
##
## Imputed Values:
##
## X[[i]]
##        n  missing distinct
##      208        0        4
##
## Value          0     1     2     3
## Frequency     53    51    67    37
## Proportion 0.255 0.245 0.322 0.178
```

```
##  ageyrs_r       sex        student_r   area      educate_r currmarr_r
##  15: 7580   female:25767   no :25786   0:36210   0: 1073   0:38372
##  16: 9201   male  :21923   yes:21904   1: 8862   1:46617   1: 9318
##  17: 9086                             2: 2618
##  18:11310
##  19:10513
##  SEScat
##  0: 9940
##  1:11044
##  2:12775
##  3:13931
##
```

```
##   ageyrs_r  sex student_r area educate_r currmarr_r SEScat
## 1       19 male        no    2         1          0      0
## 2       18 male       yes    0         1          0      1
## 3       16 male        no    0         1          0      2
## 4       15 male        no    0         1          0      1
## 5       16 male        no    0         1          0      1
## 6       17 male        no    0         1          0      3
```

Percentages of "yes" in different sexes, ages and rounds are shown as following.

sex	visit	15	16	17	18	19
female	1	52.14	30.34	20.54	10.90	4.39
female	2	55.86	44.18	26.42	18.18	10.08
female	3	59.33	44.37	29.83	12.82	10.10
female	4	65.00	50.70	27.74	14.51	9.94
female	6	59.71	44.12	22.98	15.57	7.42
female	7	64.25	48.52	35.28	18.42	8.79
female	8	71.00	52.36	33.33	18.71	12.53
female	9	74.60	62.87	42.65	21.47	12.19
female	10	75.61	60.75	44.93	25.46	15.71
female	11	81.46	69.80	50.00	31.40	18.27
female	12	83.87	70.52	52.16	34.81	17.36
female	13	84.78	77.89	57.72	36.68	25.95
female	14	88.54	78.55	62.46	43.27	21.84
female	15	85.32	73.86	54.91	30.39	17.60

sex	visit	15	16	17	18	19
female	16	88.01	76.67	55.48	32.60	17.29
female	17	83.18	75.76	52.84	28.33	18.26
female	18	82.00	75.79	56.06	39.44	25.87
male	1	56.54	47.31	33.19	19.67	19.91
male	2	70.89	57.84	47.21	38.08	38.14
male	3	62.77	59.86	43.62	27.84	24.10
male	4	73.85	58.72	51.61	31.76	24.35
male	6	74.58	59.19	48.05	25.52	23.10
male	7	71.76	66.32	43.38	33.33	18.22
male	8	75.91	62.50	54.15	37.10	25.35
male	9	82.69	70.30	65.09	41.41	31.62
male	10	75.00	66.91	55.29	48.47	27.23
male	11	88.18	74.89	56.40	40.48	38.05
male	12	86.31	74.91	52.96	40.07	36.61
male	13	88.28	73.52	60.70	44.80	31.86
male	14	85.82	76.86	64.60	55.52	39.53
male	15	87.10	75.57	66.29	48.17	35.60
male	16	84.18	70.55	58.63	43.24	34.41
male	17	83.39	72.65	54.02	44.02	29.94
male	18	78.90	68.10	55.62	39.50	38.59

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(missForest)
library(Hmisc)
library(haven)
training <- read_dta("training.dta")
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))
# for sexp1yr > 92, make them as NAs
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) %>%
  mutate(sex = as.factor(sex),
         area = as.factor(area),
         educate_r = as.factor(educate_r),
         currmarr_r = as.factor(currmarr_r),
         sexp1yr = ifelse(sexp1yr > 92, NA, sexp1yr),
         SEScat = as.factor(SEScat))
head(impu_data)
summary(impu_data)
# need to first change dataset into matrix, then change it into data frame
# how to decide parameters? cross-validation?
impu_data = as.data.frame(as.matrix(impu_data))
set.seed(123)
impu_forest = missForest(impu_data, variablewise = TRUE)
```

```r
#impu_forest$OOBerror
impu_forest_df = impu_forest$ximp
summary(impu_forest_df)
head(impu_forest_df)
write.csv(impu_forest_df, file = "impu_forest.csv")
visit_data = student_data %>%
  filter(ageyrs_r <= 19)
impu_forest_df$visit = visit_data$visit

table1 = impu_forest_df %>%
  group_by(visit, student_r, sex, ageyrs_r) %>%
  dplyr::summarize(count = n()) %>%
  spread(key = student_r, value = count) %>%
  mutate(sum = no + yes) %>%
  mutate(no_prc = round(no/sum, 4)*100, yes_prc = round(yes/sum, 4)*100) %>%
  select(sex, visit, ageyrs_r, no, no_prc, yes, yes_prc, sum) %>%
  ungroup()

table1 %>%
  select(sex, visit, ageyrs_r, yes_prc) %>%
  spread(key = ageyrs_r, value = yes_prc) %>%
  knitr::kable(digits = 3)
set.seed(123)
impu_himsc = aregImpute(~ ageyrs_r + sex + student_r + area + educate_r + currmarr_r + SEScat, data = i
impu_himsc_l = impute.transcan(impu_himsc, data=impu_data, imputation=1, list.out=TRUE, pr=FALSE, check
impu_himsc_df = as.data.frame(impu_himsc_l)
summary(impu_himsc_df)
head(impu_himsc_df)
write.csv(impu_himsc_df, file = "impu_himsc.csv")
impu_himsc_df$visit = visit_data$visit
table2 = impu_himsc_df %>%
  group_by(visit, student_r, sex, ageyrs_r) %>%
  dplyr::summarize(count = n()) %>%
  spread(key = student_r, value = count) %>%
  mutate(sum = no + yes) %>%
  mutate(no_prc = round(no/sum, 4)*100, yes_prc = round(yes/sum, 4)*100) %>%
  select(sex, visit, ageyrs_r, no, no_prc, yes, yes_prc, sum) %>%
  ungroup()

table2 %>%
  select(sex, visit, ageyrs_r, yes_prc) %>%
  spread(key = ageyrs_r, value = yes_prc) %>%
  knitr::kable(digits = 3)
```