Student_r variable

Xinyi Lin

7/3/2019

```
library(tidyverse)
```

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))
```

Analysis

```
plot_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(visit, ageyrs_r, sex, occup1_r, occup2_r, student_r)
```

table:

```
 table = plot_data %>%
  group_by(visit, student_r, sex, ageyrs_r) %>%
  summarize(count = n()) %>%
  spread(key = student_r, value = count) %>%
  mutate(na = ifelse(is.na(`<NA>`), 0, `<NA>`),
         sum = no + yes + na) %>%
  select(-`<NA>`) %>%
  mutate(no_prc = round(no/sum, 4)*100,
         yes_prc = round(yes/sum, 4)*100,
         na_prc = round(na/sum, 4)*100) %>%
  select(sex, visit, ageyrs_r, no, no_prc, yes, yes_prc, na, na_prc, sum) %>%
  ungroup()
```

Numbers and percentages of missing values in round 1, 2, 3, 4, 11, 12, 16, 17.

```
table %>%
  filter(visit %in% c(1,2,3,4,11,12,16,17)) %>%
  select(sex, visit, ageyrs_r, na_prc) %>%
  spread(key = ageyrs_r, value = na_prc) %>%
  knitr::kable(digits = 3)
```

sex	visit	15	16	17	18	19
female	1	0.00	0.00	0.00	0.00	0.00
female	2	24.77	46.23	50.94	58.13	61.80
female	3	13.88	42.38	50.51	60.51	70.69
female	4	8.57	53.05	63.36	67.10	77.07
female	11	0.00	0.00	0.00	0.00	0.00
female	12	0.00	0.00	0.00	0.00	0.00
female	16	0.00	4.33	16.61	12.71	14.41
female	17	0.00	27.00	50.50	41.13	47.75
male	1	0.00	0.00	0.00	0.00	0.43
male	2	31.01	50.75	45.18	52.88	53.61
male	3	13.14	48.39	63.79	64.05	66.78
male	4	7.69	43.02	70.05	72.35	77.86
male	11	0.00	3.65	2.84	3.97	12.68
male	12	0.00	0.00	0.79	1.06	0.79
male	16	0.00	1.29	20.83	20.00	26.88
male	17	0.00	26.27	56.27	58.31	60.78

Conclusion: In the table above, we can find that there is only one missing value in round 1 and missing values only exit in male in round 11 and 12. In round 16 and 17, all ages have missing value except age 15. There is no missing value in other rounds except round 1, 2, 3, 4, 11, 12, 16, 17.

```
liner_table = table %>%
  filter(na_prc > 0)
lm = lm(na_prc ~ sex + visit + ageyrs_r, liner_table)
summary(lm)
```

```
##
## Call:
## lm(formula = na_prc ~ sex + visit + ageyrs_r, data = liner_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.182  -4.998   1.789   6.777  14.565
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -188.841     23.887  -7.906 5.51e-10 ***
## sexmale        2.946      3.078   0.957   0.3438
## visit2        68.186     11.280   6.045 2.90e-07 ***
## visit3        70.068     11.280   6.212 1.65e-07 ***
## visit4        74.668     11.280   6.619 4.14e-08 ***
## visit11       20.065     11.791   1.702   0.0959 .
## visit12       10.257     12.105   0.847   0.4014
## visit16       30.385     11.304   2.688   0.0101 *
## visit17       61.754     11.304   5.463 2.06e-06 ***
```

```
## ageyrs_r       9.807      1.119   8.763 3.33e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 44 degrees of freedom
## Multiple R-squared:  0.8529, Adjusted R-squared:  0.8228
## F-statistic: 28.35 on 9 and 44 DF,  p-value: 1.724e-15
```