Homework 7

Xinyi Lin

10/28/2019

```
library(tidyverse)
```

```
nhanes_data = read.csv("./nhanes.csv")
```

Question 1

The variable `WTMEC2YR` represents the sampling weights. The minimum, 1st, 5th, 10th, 25th percentile, median, 75th, 90th, 95th, 99th percentile, and maximum of `WTMEC2YR` are shown below. The minimum equals to 0th percentile and the maximun equals to 100th percentile.

```
quantile(nhanes_data$WTMEC2YR, c(0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95, 0.99, 1))
```

```
##         0%         1%         5%        10%        25%        50%
##   1363.174   2984.029   6616.504   9591.616  20053.870  32093.159
##        75%        90%        95%        99%       100%
##  70187.100  92689.524 102437.128 116640.874 156152.181
```

```
sum_weight = sum(nhanes_data$WTMEC2YR)
sum_weight
```

```
## [1] 210152592
```

The sum of the sampling weight equals to 2.1015259×10^8 which is much larger than the sample size, so the weight variable hasn't been normalized.

Question 2

Variables `SDMVPSU` and `SDMVSTRA` contain the sampling error codes. There are 15 strata selected from stratum and in each stratum, 2 PSUs are selected. Sample sizes of each PSU are shown below.

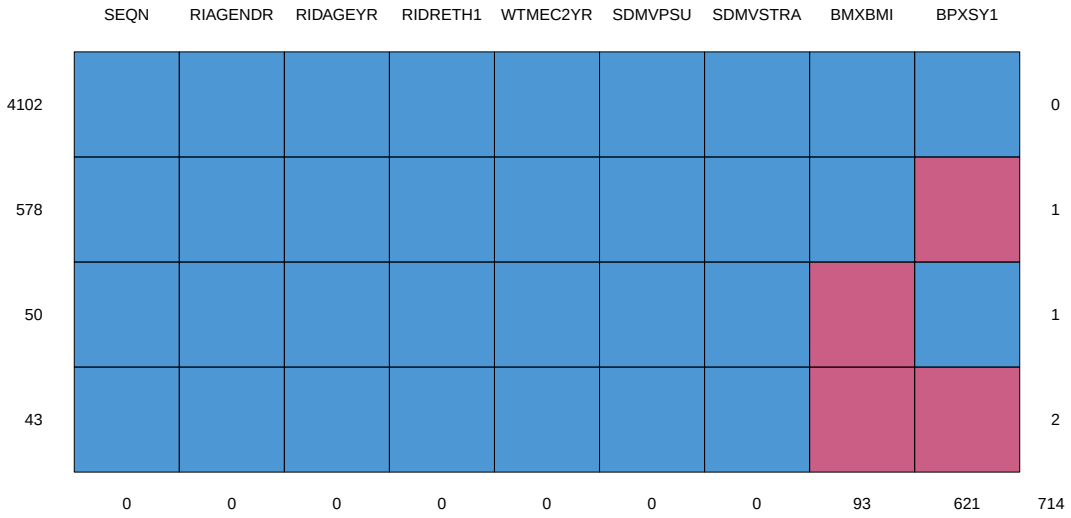
```
nhanes_data %>%
  group_by(SDMVSTRA,SDMVPSU) %>%
  summarize(n = n()) %>%
  knitr::kable()
```

SDMVSTRA	SDMVPSU	n
44	1	170
44	2	257
45	1	161
45	2	139
46	1	185
46	2	215

SDMVSTRA	SDMVPSU	n
47	1	149
47	2	185
48	1	143
48	2	135
49	1	198
49	2	211
50	1	151
50	2	205
51	1	139
51	2	127
52	1	152
52	2	186
53	1	127
53	2	135
54	1	148
54	2	159
55	1	184
55	2	117
56	1	161
56	2	148
57	1	148
57	2	103
58	1	115
58	2	120

Question 3

```
BPXSY1_mis = nrow(nhanes_data[is.na(nhanes_data$BPXSY1),])/nrow(nhanes_data)
BMXBMI_mis = nrow(nhanes_data[is.na(nhanes_data$BMXBMI),])/nrow(nhanes_data)
library(mice)
md.pattern(nhanes_data)
```




```
##      SEQN RIAGENDR RIDAGEYR RIDRETH1 WTMEC2YR SDMVPSU SDMVSTRA BMXBMI
## 4102    1        1        1        1        1       1        1      1
## 578     1        1        1        1        1       1        1      1
## 50      1        1        1        1        1       1        1      0
## 43      1        1        1        1        1       1        1      0
##        0        0        0        0        0       0        0     93
##      BPXSY1
## 4102      1   0
## 578       0   1
## 50        1   1
## 43        0   2
##         621 714
```

There are 13 percentage of cases has missing data on blood pressure and 1.9 percentage of cases has missing data on BMI. The summary of the missing data partterns is shown above. Missing values only exist in `BMXBMI` and `BPXSY1`. There are 93 missing values in `BMXBMI` and 621 missing values in `BPXSY1`.