Subject: Inquiry regarding data merging with DHS 2018 using R
Posted by woojae1995 on Thu, 09 Feb 2023 03:23:36 GMT
View Forum Message <> Reply to Message

I am currently analyzing Nigeria 2018 DHS dataset for a malaria project. I am using the household data (HR) & household member data (PR)

Because I am interested in ITN data, I referred to the github file: https://github.com/DHSProgram/DHS-Indicators-R/blob/ed302f3e5afc73e44d6113a64183173d725f8fbd/Chap12_ML/ML_EXISTING_ITN.R to produce indicators for ITN

However, I am struggling with data merging / changing household data to long format and then filtering it to only include data for children of age 6-59 months which is the scope of my research

After conducting this part of reshaping the dataset presented from the above URL;

```
# Reshaping the dataset to a long format to tabulate among nets
myvars <- c(paste("hhid"),
            paste("hml10_", 1:7, sep = ""))
HRdata_long1 <- reshape::melt(as.data.frame(HRdata[myvars]), id = c("hhid"))
HRdata_long1$idx <- str_sub(HRdata_long1$variable,-1,-1)
HRdata_long1$variable <- NULL
names(HRdata_long1)[names(HRdata_long1) == c("value")] <- c("hml10")

myvars <- c(paste("hhid"),
            paste("hml21_", 1:7, sep = ""))
HRdata_long2 <- reshape::melt(as.data.frame(HRdata[myvars]), id = c("hhid"))
HRdata_long2$idx <- str_sub(HRdata_long2$variable,-1,-1)
HRdata_long2$variable <- NULL
names(HRdata_long2)[names(HRdata_long2) == c("value")] <- c("hml21")

HRdata_long <- merge(HRdata_long1,
                     HRdata_long2, by = c("hhid", "idx"))

myvars <- c("hhid","hv005","hv025", "hv024", "hv270")

HRdata_long3 <- (as.data.frame(HRdata[myvars]))

HRdata_long <- merge(HRdata_long,
                  HRdata_long3, by = c("hhid"))
```

I get an exploded number of data entries of more than 200000 compared to something around 40,000 in the original HR data. I presume this is because by these commands the household data was expanded for all household members (1 household data -> n household members data)

But my question is; how do I select only children 6-59 months from this data?

I have tried the following for the last 3 lines of code;

```
myvars <- c("hhid","hv005","hv025", "hv024", "hv270", "hv014", "hc1")

HRdata_long3 <- (as.data.frame(HRdata[myvars]))

HRdata_long <- merge(HRdata_long,
                     HRdata_long3, by = c("hhid"))

HRdata_long <- filter(HRdata_long, hc1>=6, hc1 <=59)
```

But this will still give me 268121 observations for HRdata_long, which is far greater than what I got from my previous coding I identified. It was 11590 observations for HR data when restricted to children of 6~59 months by using the PR data as the following:

```
# keep relevant vars
PRtemp =subset(PRdata, select=c(hv001, hv002, hc1), 'NA'= TRUE)
#perform merge
HRdata <- merge(HRdata,PRtemp,by=c("hv001", "hv002"))
HRdata <- filter(HRdata, hc1 >=6, hc1 <=59)
rm(PRtemp)
```

Can anyone explain me the differences between the two approach and why I can't restrict the above long-format to just children?

From my guess, I think it is because when expanded to the long format, even for household members that are not children, they will still have the hc1 variable between 6-59 as long as they had a children of 6-59 months age in their original household. Is this the case?

If so, how can I work around to only restrict to the actual children data when I want to expand the HR data to the long format?