Subject: Discrepancy in resident status between individual files and merged household file
Posted by desktop on Mon, 11 Apr 2022 10:48:40 GMT
View Forum Message <> Reply to Message

Hi,

After merging the individual questionnaires with the household member (PR) datasets per Tom Pollum's response in this thread   (https://userforum.dhsprogram.com/index.php?t=msg&th=6693&start=0&)

, I noticed that the usual versus visiting residents differed between hv102 and (m)v135. See the R code below.

Discrepancies between women (1 = Usual, 2 = Visitor) and merged (PR+IR+MR) dataset (0 = Visitor, 1 = Usual)

table(women$V135, combined$HV102[combined$HV104 == 2], useNA = "ifany")

```
      0     1
 1 21537 655926
 2   686  21537
```


Discrepancies between men (1 = Usual, 2 = Visitor) and merged (PR+IR+MR) dataset (0 = Visitor, 1 = Usual)

table(men$MV135, combined$HV102[combined$HV104 == 1], useNA = "ifany")

```
      0     1
 1  1884 108320
 2    34   1884
```


Have I missed something, or are these discrepancies due to (m)v135 being reported by the individual themselves and hv102 being reported for all members by one person?

Subject: Re: Discrepancy in resident status between individual files and merged household file
Posted by desktop on Tue, 12 Apr 2022 14:58:25 GMT
View Forum Message <> Reply to Message

After cross-referencing my merge in R with what Tom did in STATA, I noticed several errors. Residency now checks out. Concatenating variables from the men's and women's questionnaire (such as (M)V35) has to be done after the datasets have been merged.

Below is the R code for anyone that wants to merge IR+MR+PR and does not have access to STATA.


```
# Import women's questionnaire
women <- read_sav("Your data location",
          col_select = c("V001", "V002", "V003", "V005", "V135")

# Change colnames to match household members (PR) dataset
colnames(women)[which(names(women) == "V001")] <- "HV001"
colnames(women)[which(names(women) == "V002")] <- "HV002"
colnames(women)[which(names(women) == "V003")] <- "HVIDX"

#Sort by
attach(women)
women <- women[order(HV001, HV002, HVIDX), ]
detach(women)

men <- read_sav("Your file location",
          col_select = c("MV001", "MV002", "MV003", "MV005", "MV135"))

#Change colnames to match household members (PR) dataset
colnames(men)[which(names(men) == "MV001")] <- "HV001"
colnames(men)[which(names(men) == "MV002")] <- "HV002"
colnames(men)[which(names(men) == "MV003")] <- "HVIDX"

#Sort by
attach(men)
men <- men[order(HV001, HV002, HVIDX), ]
detach(men)

household <- read_sav("Your file location",
              col_select = c("HV001", "HV002", "HVIDX", "HV005", "HV104", "HV027", "HV102"))

attach(household)
household <- household[order(HV001, HV002, HVIDX), ]
detach(household)

irpr <- merge(household, women, by = c("HV001", "HV002", "HVIDX"), all.x = T)

attach(irpr)
irpr <- irpr[order(HV001, HV002, HVIDX), ]
detach(irpr)

combined <- merge(irpr, men, by = c("HV001", "HV002", "HVIDX"), all.x = T)

# Weights
combined <- combined %>%
```

```
  mutate(weight = case_when(HV104 == 1 ~ MV005,
                HV104 == 2 ~ V005))
```

```
# Re-weight men due to 15% sampling probability
combined <- transform(combined, adj_weight=ifelse(HV104 == 1 & HV027 == 1, weight*(1/.15),
                                weight))
```

```
combined <- combined %>%
  mutate(resident = case_when(HV104 == 1 ~ MV135,
                HV104 == 2 ~ V135))
```

```
combined <- combined %>%
  mutate(resident = case_when(resident == 1 ~ 1,
                    resident == 2 ~ 0))
```

```
table(combined$resident, combined$HV102)
     0     1
 0 24141     0
 1     0 787667
```

```
all.equal(as.numeric(combined$HV102)[!is.na(combined$V005) | !is.na(combined$MV005)],
combined$resident[!is.na(combined$resident)]
)
```

```
TRUE
```

Still some minor discrepancies for other variables though, such as marital status. More NAs in the PR file. Better to use variables in individual files, when possible?

```
#Add S301/SM213/HV116 to col_select calls for IR/MR/PR datasets to code in previous chunk
```

```
combined <- combined %>%
  mutate(marriage = case_when(HV104 == 1 ~ SM213,
                HV104 == 2 ~ S301))
```

```
combined$marriage
Labels:
 value                 label
    0           Never married
    1         Currently married
    2 Married, gauna not performed
    3              Widowed
    4              Divorced
    5             Separated
    6              Deserted
```

```
combined$HV116
```

Labels:
 value   label
    0   Never married
    1   Currently married
    2 Formerly/ever married

table(combined$marriage, combined$HV116)

      0   1   2
 0 207332 2198   265
 1   1892 566533 1402
 2   1718   499   36
 3   106   1114 20034
 4   113   220   3126
 5   70   634   3406
 6   16   109   938

sum(table(combined$marriage))-sum(table(combined$HV116[!is.na(combined$V005) |
!is.na(combined$MV005)]))
[1] 47