Subject: Inquiry regarding DHS 2018 analysis using R
Posted by woojae1995 on Wed, 18 Jan 2023 11:27:46 GMT
View Forum Message <> Reply to Message

I am currently doing a secondary analysis project using the 2018 DHS dataset of Nigeria.

Currently, I am using R and I have several technical/coding questions. I am currently using the 2018 DHS individual & children dataset.

1) How do I get a list of the column labels in R?
- I want to know the labels for the column (ex. b19 = current age of child in months) and the labels for the answer choices (ex. for the question asking the sex of the respondent; b4, 1= male, 2=female)
- I did find the 'STANDARD RECODE MANUAL for DHS-7' published by the USAID, but it still does not have the full response labels.
- Is this a problem inherent to using R? I heard that labels are easily visible when using STATA. However, since I have been using R till now, I wonder if there is a way to create a list of all the questions & labels for the dataset I am using.

2) How do I merge two dataset in R?
Referencing from this site 'https://dhsprogram.com/data/Merging-datasets.cfm', I merged the children dataset & individual dataset using the following code in R

```
NigeriaIR <- read_dta('NGIR7BFL.DTA')
NigeriaChildrenKR <- read_dta('NGKR7BFL.DTA')
NigeriaKRIR <- merge(NigeriaChildrenKR, NigeriaIR, by = c('v001','v002'))
```
*IR = individual dataset, KR = children dataset

Is this the correct way to merge it? I am concerned because the children dataset itself has 33924 observations, individual dataset has 41821 observations but when I merge them by v001 (cluster number) v002 (household number), I get 52982 observations.
From my crude understanding, I cannot understand how the merged dataset has more observations than the number of observations for the individual dataset. Could anyone explain why this is happening or what I am doing wrong?

3) Is this the correct way to account for the weighted-survey?
```
NigeriaKRIRsvy <- svydesign(id = NigeriaKRIR$v021.x, strata=NigeriaKRIR$v022.x, weights = NigeriaKRIR$v005.x/1000000, data=NigeriaKRIR)
```
*NigeriaKRIR is the merged dataset name
*for some reason, after I merged the dataset, the vXXX variables (ex. v001, v002) change to vXXX.x (ex. v001.x, v002.x)

Thank you all in advance

Subject: Re: Inquiry regarding DHS 2018 analysis using R

Hello,

Thank you for reaching out. At the DHS Program we are beginning to prepare more resources for users to learn DHS data analysis in R but these are not completed yet and please keep an eye on our social media accounts for updates on this. Please also check our code share library on GitHub which provides code for constructing DHS indicators in R:
https://github.com/DHSProgram/DHS-Indicators-R

In the meantime I can answer your questions.

1. To view labels in R you just can just use the command print_labels from the haven library. See below an example:

```
library(haven)
print_labels(NigeriaIR$b4)
```

2. The merge code you have is correct but there is no reason to merge the IR and the KR file. The IR file is the women's file and the unit of analysis is the woman. The KR file is the file for children under five and the unit of analysis is the child. This file also contains all the information about each child's mother, that is why there is no need to merge these files. The reason why you see v001.x etc is because there can be more than one child for each mother. You can learn about the different data files here: https://www.youtube.com/watch?v=fzLNQkkvDeI&t=105s

Let's say instead you want to merge the HR file (household file where the unit of analysis is the household) with the IR file (woman's file). Perhaps there is some information that is only in the household file that you need for your analysis of women in the IR file. Then you would do the following:

```
IRdata <- merge(IRdata,HRdata,by=c("v001", "v002"))
```

Here you don't need the v003 since there can be several woman in the same household. So this is a many to one merge.

Another example:
If you want to merge the PR file (person's file) with the MR file (men's file). This is a one to one merge. You would do the following:

```
MRdata <- merge(MRdata,PRdata,by=c("mv001", "mv002", "mv003"))
```

3. the survey design code is also correct except for using .x for the variables. This was a result of your incorrect merge of the IR and KR files. For instance in an IR or KR file you would do the following.

```
# creating the sampling weight variable.
IRdata$wt <- IRdata$v005/1000000
```

```
mysurvey<-svydesign(id=IRdata$v021, data=IRdata, strata=IRdata$v022,  weight=IRdata$wt,
nest=T)
options(survey.lonely.psu="adjust")
```

Hope this helps.

Best,
Shireen Assaf
The DHS Program