Subject: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Sat, 13 Apr 2024 12:17:04 GMT
View Forum Message <> Reply to Message

Hi all, I hope this message finds everyone in good health. I am currently a Master's student in my final semester. I am using DHS data for my thesis. In the interest of simplicity, I will break down the multifaceted problem I am having below. I hope someone might be able to find the time to help me with this strange issue.

Objective: I created an account with DHS and downloaded the data. The goal with the data downloaded is to disaggregate countries into survey "round" (year of survey) and a handful of variables from that round so that I can then merge each respective round with its' shape file (since these change across time for each country). This is important because I will need to merge data via geographic coordinates for my thesis, which is exploring the impact of environmental variables (ex. precipitation rate) on the propensity of marriage under the age of 18 across Sub-Saharan Africa.

Problem: I bulk downloaded survey and geographic data for every African countries where this was available. I decided to start working with one country only so that I could clear any issues with the code before replicating the process for the rest of the countries. To simplify the process, I grouped the bulk downloaded data into its' respective countries and tried to import batches to STATA to work with. The problem begins with the first country I attempted to work with, Tanzania. While I was able to unzip all the files in STATA, this was the furthest I was able to get because what ensued was a bizarre game of smoke and mirrors with the files. For efficiency, I have listed the most major problems below:
1. In the expanded and unzipped files, sometimes I would see a file that does not have a .dta listed, yet, when I would manually go into this file through my Finder just to double check, there would be a .dta file.

2. There are also situations where an expanded/unzipped file would list its' contents as including a dofile, and when I would go through my Finder to manually ensure that this was there, there would be nothing within the contents of the file.

3. Perhaps the largest issue is that it is impossible to run the do file importing the datasets of .dta files because every single path is different inside those files (not possible to write an extraction loop). I made a list of some of the different paths of the .dta files so anyone reading can better understand the issue. This means that I can't get variable lists into STATA.

Below is the code I have used in STATA:

```
cd "/Users/tbear/Desktop/M2 Thesis/DHSDATA/Tanzania"
  capture log close
```

```
  log using "D:\Niveen Wrking Files\Feps files\FEPS Teaching Files\Year
23-24\MDE\teddi\unzipfiles.log", replace


  ** [1] Unrar/Unzip all files under the main "DHSDATA" folder

  * You need first to run this two lines to make STATA able to extract rar files
  shell set path="C:\Program Files\WinRAR"; %path% & unrar e "*"

  ** some errors resulted while extracting the zip files:
  * Zip files under which also contains another zip files - 7 files:
/*
            "SNBR70FL"
            "SNCR7IDT"
            "SNCR7IFL"
            "SNCR70DT"
            "SNCR70FL"
            "SNBR7IFL"
            "SNBR70DT"
*/
  *  they can be extracted manually, then copy their contents zip files back into the main folder
"DHSDATA"
  * Now unzipping command will work
  local path "/Users/tbear/Desktop/M2 Thesis/DHSDATA/Tanzania"
  local filelist : dir "`path'" files "*.zip", respectcase
  foreach file of local filelist {
   unzipfile `file', replace
  }

  ** [2] Extract all the "dta" files in each subfolder under "DHSDATA" folder


  * make new folder in which all "dta" files will be saved
  global usefile "/Users/tbear/Desktop/M2 Thesis/DHSDATA/Tanzania"
  capture mkdir "/Users/tbear/Desktop/M2 Thesis/DHSDATA/Tanzania/Tanzania_dta"
  clear
  capture set maxvar 100000

  local filelist : dir "$usefile" files "*.DTA", respectcase
 foreach file of local filelist {
  quietly use "`file'", clear
  * save each "data" files into the new folder that we made in the first step
  save "Tanzania_dta/`file'", replace
 }

  local filelist : dir "$usefile" files "*.dta", respectcase
 foreach file of local filelist {
  quietly use "`file'", clear
```

```
  * save each "data" files into the new folder that we made in the first step
  save "Tanzania_dta/`file'", replace
 }


  capture log close


clear
**********
```

Thank you so, so much to anyone that might be able to help!!

File Attachments

1) WhatsApp Image 2024-04-03 at 8.08.05 PM.jpeg, downloaded 160 times
2) WhatsApp Image 2024-04-03 at 8.08.27 PM.jpeg, downloaded 159 times
3) WhatsApp Image 2024-04-03 at 8.11.56 PM.jpeg, downloaded 162 times
4) WhatsApp Image 2024-04-03 at 8.12.15 PM.jpeg, downloaded 160 times
5) filenames_dct_paths.pdf, downloaded 30 times

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Bridgette-DHS on Mon, 15 Apr 2024 15:02:02 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

We believe your problem is with the unzipping procedure, and/or the use of a Mac, and not with the DHS files. My personal strategy in this situation would be to find or construct another dta file, zip it, and then try to unzip it, to learn more about the unzipping steps. Hope you can quickly figure this out and proceed with your research. Perhaps the IT staff at your university can help.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Thu, 18 Apr 2024 14:07:05 GMT
View Forum Message <> Reply to Message

Hi, thank you so much for responding. There are now multiple professors from my University across several departments trying to assist with this issue, but the issue is not with respect to

unzipping the files- rather the issue is that the .dta files are not being extracted from the files when unzipping. We have attempted to make a loop using the dofile to extract the .dta but every .dta has a different type of file path- this is why we can't run everything at once... It goes without saying that extracting every .dta file for thousands of files would be nearly impossible in the month and a half I have left until my submission deadline. I'm truly beginning to panic because even professors who have worked with DHS data here have been puzzled by this challenge for several weeks now. Any resources or guidance you might be able to provide would be greatly appreciated.. Thank you so much in advance.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Fri, 19 Apr 2024 14:56:08 GMT
View Forum Message <> Reply to Message

Hi, I would like to help you resolve this issue. A couple of first steps:
1) If you are working in Stata, you only need the DT files (e.g. TZxxvvDT.zip) and not the FL zip files (e.g. TZxxvvFL.zip), with the exception of the geospatial (GE) files which are only in one format. All of the DT files contain the .dta data files that you need. There is no need to use the .do and .dct files that are found in the FL zip files as the same data are in the DT files.zip
2) Once you have just the DT zip files, you will find the .dta files inside of those zip files and should be able to unzip them automatically to the location of your choice.
Can you test this out and see if it works?

I'm a little confused about the images you have shared of the file contents. I think you have cut sections of the output into the images you shared, but I think you have cut them in the wrong place. For example in your image that starts "successfully extracted TZKR41DT.zip ...", you are mixing the output from extracting two different zip files. You are showing the end of the process for extracting TZKR41DT.zip, with the list of files being shown before "successfully extracted TZKR41DT.zip ..." (but not included in the image you sent) and then telling you that 4 files were processed. Then in the same image you are showing the first part of the extraction of TZKR63FL.zip and listing the 8 files extracted from that, but then not showing the success message.

Let us know if this helps clarify your issues.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Mon, 29 Apr 2024 12:04:39 GMT
View Forum Message <> Reply to Message

Hi Trevor, THANK YOU SO MUCH. I have been mostly able to solve the problem thanks to the help you have given me. I have one remaining challenge in order to move forward and that is the merging of the different survey data sets for each survey round. So, to be clear, I am interested in controlling for wealth in my proportional hazard. This means that I will have to combine household data and individual data, as indicators for wealth do not exist in the individual recode for Tanzania 1999 (the year I'm starting this data work with). I have read from Tom Pullum in another part of this forum that "it would be virtually impossible to merge them [individual survey] with the HR file,

which has households as units. You should use the PR file, which has individual household members as units, rather than the HR file." However, it's not clear to me that wealth is included in the PR file, either. To make matters a bit more confusing, I have seen that this type of merge is possible elsewhere on the internet, for example: https://www.researchgate.net/post/How_can_I_merge_Household_database_to_Women_data_base_in_the_DHS_data_using_stata

Do you mind clarifying if it is possible to merge household and individual level data? This is the only way I will be able to control for wealth in my proportional hazards analysis. Thank you so much in advance for your time.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Mon, 29 Apr 2024 15:54:34 GMT
View Forum Message <> Reply to Message

It's actually come to my attention that the reason why I was having trouble finding the wealth data in the Household Recode or Household Member Recode for the Tanzania 1999 round is because the wealth index is separated in an entirely different file. For this earlier survey rounds, such as the one I am working on- is it possible to merge the wealth index with the Individual (IR) and Household Member (PR) survey data sets on Stata? Thank you so much in advance.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Mon, 29 Apr 2024 19:49:11 GMT
View Forum Message <> Reply to Message

Hi
You need the TZWI41DT.zip file for the wealth index for the wealth index for this survey. Yes, you can merge the wealth index data to the IR or the PR data. You can find information about merging datasets in the Guide to DHS Statistics in Chapter 1) Introduction and Description of Datasets, Analyzing DHS Data, Matching and Merging Datasets. This doesn't provide specific information for merging the wealth index data, but does provide several examples of how to merge data. The wealth index data is based on households, so you can link the data to either the IR (individuals who live in the household) or PR (persons in the household), using the household ID information.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Wed, 01 May 2024 16:02:18 GMT
View Forum Message <> Reply to Message

Hi, thanks so much for the response. I've been a bit stumped by this merging process because there are three different datasets (IR, PR, and Wealth Index) that I have to combine and I'm a bit confused as to which one should be my base for merging. I was going to make the IR my base for merging because I've seen code that allows for the renaming of the cluster number, household number, and respondent's line number such as this:

use "/Users/tbear/Desktop/THESIS DATA/Tanzania_1999/Tanzania_1999_dta/TZIR41FL.dta", clear

* keep the variables you want
keep v0*
sort v001 v002 v003
save e:/Users/tbear/Desktop/THESISDATA/Tanzania_1999/Tanzania_199 9_dta/TZIR41FL.dta, replace

* Prepare PR file and merge
use "/Users/tbear/Desktop/THESIS DATA/Tanzania_1999/Tanzania_1999_dta/TZPR41FL.dta", clear
* reduce to women who are eligible for the IR file
keep if hv117==1

* keep the variables you want
keep hv0* sa33 sh*

rename hv001 v001
rename hv002 v002
rename hv003 v003

sort v001 v002 v003
merge v001 v002 v003 using ***Not entirely clear what I should be using here
tab _merge
******

BUT the problem is the wealth index only has the hhid variable that I can use to merge- and the IR file does not have this, only the PR file does. Should I be using the PR file as my base, merging the IR file, and then appending the wealth index?

Thank you so much for all of your help.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Wed, 01 May 2024 17:01:21 GMT
View Forum Message <> Reply to Message

Hi
A few notes:
1) It looks like you are opening the IR file, then keeping just a few variables and sorting the file, and then overwriting the original file. This is generally not considered good practice as you are modifying the original file. Generally, you should start with your original file, but save to an interim file with a different name or in a different folder (or both).
2) The naming of your THESISDATA folder seems to vary - in two cases it has a space between THESIS and DATA and in one case it doesn't (this may be a display issue in the user forum as it occasionally puts extra blanks into the text).

3) In terms of the order of merging, I would start by merging the wealth index to the PR file and saving your output to an intermediate file. They both should have hhid so you should be able to merge those without problem. Then merge the info from the PR/wealth data onto the IR file. Below is a rough outline of the process (I haven't tested this, so there may be some bugs - this is just to give you the order of operations):

```
use TZWIxxxx.dta
sort hhid
save TZWIxxxx.dta, replace

use TZPRxxxx.dta, clear
* keep the variables you want from the PR file
keep hhid hv0* ...
sort hhid
merge m:1 hhid using TZWIxxxx.dta
clonevar v001 = hv001
clonevar v002 = hv002
clonevar v003 = hvidx
sort v001 v002 v003
save TZPRxxxx_temp.dta

use TZIRxxxx.dta, clear
sort v001 v002 v003
merge 1:1 v001 v002 v003 using TZPRxxxx_temp.dta
```

It is also possible to construct hhid from hv001 and hv002 or from v001 and v002, and vie versa, but I don't think you need to.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Wed, 01 May 2024 18:40:35 GMT
View Forum Message <> Reply to Message

Hi Trevor, thank you SO much for the guidance :). The DHS data is amazing but definitely a little tricky to navigate at first. I've adapted the code to meet the needs of what I'm trying to do but I've been a bit stuck since your previous message because I keep receiving an error message on STATA telling me that "variable hhid does not uniquely identify observations in the using data"

I'm not sure if this is because I had to clone whhid and set it equal to hhid at first (because the case identifier in these older wealth indexes doesn't match exactly the case identifier in the PR file) but in any case I know for a fact that hhid uniquely identifies cases in the PR file.. Could this be a situation wherein I have to construct hhid from hv001 and hv002 or from v001 and v002 as you alluded in your previous message (really hope not lol)... Below is my code just in case you might be able to see any problems I haven't picked up on so far.

```
use TZWI41FL.dta
sort whhid
save TZWI41FL.dta, replace

use TZPR41FL.dta, clear
```

```
* keep the variables you want from the PR file
clonevar whhid = hhid
keep hhid hv005 hv007 hv025 hv219
sort hhid
merge m:1 hhid using TZPR41FL.dta
clonevar v001 = hv001
clonevar v002 = hv002
clonevar v003 = hvidx
sort v001 v002 v003
save TZPR41FL_temp.dta

use TZIR41FL.dta, clear
sort v001 v002 v003
merge 1:1 v001 v002 v003 using TZPR41FL_temp.dta
```

As always thank you, thank you for any guidance you might be able to provide!

File Attachments

1) Screen Shot 2024-05-01 at 9.40.04 PM.png, downloaded 12 times

---

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Wed, 01 May 2024 19:50:06 GMT
View Forum Message <> Reply to Message

The following line:
merge m:1 hhid using TZPR41FL.dta
should refer to the WI file, not the PR file.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Thu, 02 May 2024 12:46:09 GMT
View Forum Message <> Reply to Message

Hi Trevor, I fixed this line of code- thank you so much. Unfortunately, there is still an error cropping up with respect to the hhid case identifier. Below is the lines of code concerned:

```
use TZWI41FL.dta
sort whhid
save TZWI41FL.dta, replace

use TZPR41FL.dta, clear
* keep the variables you want from the PR file
clonevar whhid = hhid
keep hhid hv005 hv007 hv025 hv219
```

```
sort hhid
merge m:1 hhid using TZWI41FL.dta
clonevar v001 = hv001
clonevar v002 = hv002
clonevar v003 = hvidx
sort v001 v002 v003
save TZPR41FL_temp.dta

use TZIR41FL.dta, clear
sort v001 v002 v003
merge 1:1 v001 v002 v003 using TZPR41FL_temp.dta
```

Unfortunately, it seems like the wealth index cannot be emrged using either "hhid" or "whhid," as I've received error messages for both iterations of this code (attached in photos). I've been trying to figure out what the issue is and perhaps I shouldn't be using the command "clonevar" and perhaps I should just be renaming the variable entirely? I want to make sure that this is correct.

As always, really grateful for your assistance!

File Attachments

1) Screen Shot 2024-05-02 at 3.43.59 PM.png, downloaded 13 times
2) Screen Shot 2024-05-02 at 3.44.07 PM.png, downloaded 14 times

---

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Thu, 02 May 2024 14:03:54 GMT
View Forum Message <> Reply to Message

Hi
The problem is that hhid does not exist in the WI file, but whhid does. You had the right idea when you created whhid in the PR data using the clonevar statement, but then you dropped it immediately by not including it in the keep statement. So you need to include whhid in the keep statement and then use it in the next few statements as follows:

```
use TZPR41FL.dta, clear
* keep the variables you want from the PR file
clonevar whhid = hhid
keep whhid hhid hv005 hv007 hv025 hv219
sort whhid
merge m:1 whhid using TZWI41FL.dta
```

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by tednoel on Thu, 02 May 2024 22:12:57 GMT
View Forum Message <> Reply to Message

Hi Trevor, once more, thank you a million times over for the guidance. The importance of the 'keep' statement had slipped my mind! The WI and PR data set combined perfectly, and I'm almost there, but of course another last minute problem has cropped up when trying to combine everything with the IR dataset. Here I've been stuck on a somewhat bizarre issue for the past couple of hours. STATA is insisting when I try and do the last merge that the "variable _merge is already defined." I'm aware that when STATA performs merges it might save the merge as a variable so I attempted to drop the merge command and then rerun the merge but when I try and drop the "_merge" it tells me that this variable doesn't exist. However, when I try to rerun the merge it then tells me once more that the variable is already defined. I'm attaching a screenshot here so you can see what the problem looks like in STATA. You've been of such incredible help so far so I figured I'd ask because maybe this is a problem others have had as well? It's unusual but I know you've seen so much with respect to the issues that might crop up with DHS data in STATA. As always, thank you so much in advance for any guidance you might be able to provide <3

File Attachments

1) Screen Shot 2024-05-03 at 1.09.24 AM.png, downloaded 16 times

---

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Thu, 02 May 2024 23:18:27 GMT
View Forum Message <> Reply to Message

Hi
_merge is probably in TZPR41FL_temp.dta. You should drop it before you save TZPR41FL_temp.dta.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by yyumeee on Sun, 12 May 2024 17:54:26 GMT
View Forum Message <> Reply to Message

Sorry to bother, I was having some problems with some merging on the WI datasets.
My study requires assigning WI values to certain locations, without really needing to connect the WI dataset to the HR dataset, but in order to connect to the GPS Dataset I need the DHSCLUST variable which is obtainable through the HR dataset (var hv001).
I followed the steps detailed in the GitHub document, but R doesnt merge HR and WI datasets, since the whhid in the HR dataset created to merge with the whhid of the WI dataset don't seem to match. The problem looks to be in the fact that thw whhid are written in a slight different way (HR dataset has a blank space between the two numbers, whereas WI does not).
Is there a way to be able to fix this (since I need to perform this task on multiple countries, this needs to be easily iterable)?
Alternatively, if there can be a way to directly merge the WI and GPS datasets without needing the HR dataset, that would be great (I guess the first number of the whhid corresponds to the DHSCLUST, but the problem is the same in that for double digits clusters + double digits households ids there is no blank space to separate the two, so that the whhid 1028 could very

well mean both cluster 10 household 28 or cluster 102 household 8).
Thank you in advance.

File Attachments

1) Screenshot (248).png, downloaded 13 times
2) Screenshot (249).png, downloaded 8 times

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by Trevor-DHS on Mon, 13 May 2024 14:59:54 GMT
View Forum Message <> Reply to Message

Hi
I'm a little confused by a couple of things in your message"
1) In the example you are using I can see that hv000 is ZW7, which means you are using the Zimbabwe 2015 dataset. However, there is no WI file for this survey as the wealth index is already in variable hv270 (and hv271 for the scores). So, I assume you are incorrectly matching to a different survey's wealth index data. That I think explains why you see the gap in the hhid fields in one file but not the other. WI data files exist for surveys DHS2-4 and a few from phase 5, but the wealth indices are included in the main data files for phases 6-8 and mot of phase 5.
2) You say that you want to merge the GPS and WI data without needing the HR data. But it makes no sense to use just the WI and GPS data - surely you a using other DHS variables too, in which case you can link to that file directly.

Subject: Re: Strange Issues w/ Data Formatting from DHS
Posted by yyumeee on Mon, 13 May 2024 17:12:52 GMT
View Forum Message <> Reply to Message

Hi, thank you for the response.
It seems you were spot on and it was just me not getting the right datasets.
The WI scores have a different scale than the ones on the WI set, but I take they just need to be normalized?
