So You Want to Analyze Household Surveys?

Stephen Matteo Miller

Household surveys and other large data sets are challenging to analyse, but skill in doing so is highly prized. What follows is a brief introduction to analysing household surveys using the STATA statistical package. A much more thorough introduction to the subject is given by Deaton (1997).

Each data set is unique, but what follows is a primer, documenting a process that you may wish to follow to analyse a household survey. There are many types of surveys, but among the most complicated are the Living Standard Measurement Survey (LSMS) databases.[1] LSMS is a World Bank funded project that aims to help developing country governments compile the information from surveys for use in forming policy. The trouble with these data sets is that they require that you learn to "think big," because often the data sets have more observations that you could ever imagine counting or examining yourself (such data sets are often called "k-large" data sets). The amount of data is often such that it can only be managed if it is partitioned into many smaller data files (you might see twenty or thirty for one survey); sometimes an individual file can even be too big for your computer to manage. Once you figure out how to get the data into the computer's memory, a challenge in and of itself, now the hard part begins.

Broadly speaking, the process should be as follows: 1) read the documentation, including the questionnaire itself as well as the data dictionary, 2) make a list of the data series you wish to pull from the database, 3) compile the subset of data that you wish to use by merging individual data files, 4) analyse the data. A discussion of each step with examples is given below.

Step 1: Read the Documentation

When it comes to documentation for LSMS surveys, the three most important ones are: 1) the report about the household survey data, 2) the questionnaire, and 3) what's often called the data dictionary. You should read through the household survey report from start to finish to figure out how the survey was conducted (you'd be surprised, but some details in these reports, as well as local knowledge about the country, have the potential to help you figure out data puzzles). From the household survey report you'll find out important things like

1. was it a (rotating-) panel, or a one-shot survey?
2. when was the survey conducted?
3. how many people were surveyed?
4. was it nationally representative?
5. was it stratified by urban & rural, and by region?)

[1] See http://www.worldbank.org/lsms.

You don't necessarily have to read everything, but spend between a few hours and a couple of days skimming through and processing it.

Next, you should also have a glance at the questionnaire. It's not necessarily a lot of fun, but before you get to the data, you may wish to spend anywhere from a few hours to two days reading through this. The questions asked will help you determine what the boundaries are that will confine your study (i.e., if there's a section in the survey about how long it takes for a family to get to the market, then you might be able to empirically investigate how distance to market, as a form of transaction cost, affects choices).

You can begin using the third document, the data dictionary, once you start working with the actual data. However, a word of warning is in order, because, especially in many low income countries, you may not actually have any recorded data for some questions. This can be because people are not willing to divulge information. For instance, for a variety of reasons, many people around the world underreport household income when surveyed. As a result, in many low income countries reported household consumption is a better indicator of household income than is reported income. Now that you've read and have some idea about what's available to work with, you can start to see the actual data (at this point you may be in for a great or awful surprise).

Step 2: Make a list of the data series you wish to pull from the database

This is where the fun begins. At this point, you will be going back and forth between the data dictionary, which explains what each variable in the dataset is and how it is labelled, and the data itself. It's in the process of flipping back and forth that you realize that you may have to figure out things like: "am I interested in total household consumption (cash expenditures plus non-cash consumption), or just cash expenditures?" Sometimes you'll learn that the data you have can't answer the question you were initially asking, but that's okay, because often the data you do have will also inspire you to ask questions you had not thought of before. The worst case is when you discover that the data is really poor, in which case you can't do anything you planned to do.

The other tricky part about this is that, depending on how large is the country you may be pulling together variables from many different files. What's worse, some files may be individual-specific, while others may be household data (though often the latter are simply derived from the former)? At this point you may also start to ask: "am I interested in individual behaviour or household behaviour?" So, you'll find at this point that if you haven't already sharpened your research questions, you will now have to. At the least, this step will turn out to be an ongoing process. As you get deeper into your analysis, new questions may emerge, at which point you'll dig around to find the data series to add to your current data set to answer those questions, which brings up the next issue: merging and managing data sets.

So, let's do this. Imagine that the intended data project (in this case a fairly simple one) is to analyse a few "determinants" of household consumption levels. Economic theory suggests that consumption is a function of income and wealth. Therefore, you might include current income, and age and education as a proxy for human capital, in the regression. You may also believe that households in certain regions and in urban areas are wealthier, and accordingly include these variables in the specification.

I found some data derived from the 1994 Armenian Household Survey that I once analysed (in fact, it's alot cleaner than what I had; in other words, often what you get will not be so nicely organized).[2] You'll see four files: 1) the urban household-level file (YHHOLD.DTA), 2) the rural household-level file (AHHOLD.DTA), 3) the urban individual-level file (YIND.DTA), and 4) the rural individual-level file (AIND.DTA). These files were probably complied from at least twice as many raw data files (it's been a while since I worked with the data). So, already, alot of the dirty work that you'll have to learn how to do has been done, but, you can still see how to do the basics with the following exercise.

Let's assume you've downloaded these files to your C:\drive. In fact, you should probably a folder for the files you'll use a location of your preference. The first things you'll want to do is grab the data dictionary (print outs can be helpful if you don't like flipping back and forth between programs on the computer), and have a look to see what data series you're interested in. Let's start with the urban individual file. From these individual-level files we would only like to keep information for the household head's level of education.

```
log using "C:\Armenia.log", replace
clear
set mem 200m
use "C:\YIND.DTA"
tab yrelat, gen(d)
keep if d1==1
*** This removes anyone who is not a household head ***
keep yhid ysex yage yeducat ylfs
*** This keeps only the variables you want ***
sort yhid
*** STATA's very picky about having sorted data ***
save "C:\urbanind.dta", replace
```

Now let's do the same for rural families

[2] The data is available from the following Household Expenditure and Income Data for Transitional Economies (HEIDE) website:
http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:20346891~pagePK:64214825~piPK:64214943~theSitePK:469382,00.html

```
clear
use "C:\AIND.DTA"
tab arelat, gen(d)
keep if d1==1
*** This removes anyone who is not a household head ***
keep ahid asex aage aeducat alfs
*** This keeps only the variables you want ***
sort ahid
*** STATA's very picky about having sorted data ***
save "C:\ruralind.dta", replace
```

We're done with two of the four data files, so now let's get the data from the household level files. Let's start with the urban file. Well extract the variables we'll use from the data set, sort the data, and then save the new urban household-level data set.

```
clear
use "C:\YHHOLD.DTA"
keep yhid ytothhx ytothhy yregion1 yhhsize yweit
sort yhid
save "C:\urbanhhold.dta", replace
```

Finally, we'll do the same thing with the rural file

```
clear
use "C:\AHHOLD.DTA"
keep ahid atothhx atothhy aregion1 ahhsize aweit
sort ahid
save "C:\ruralhhold.dta", replace
```

Now that you've picked the data series, you can merge them all together.

Step 3: Compile the subset of data that you wish to use by merging individual data files

With the individual variables chosen, it is now time to start bringing them together. We can start with the urban file. First, when merging data, STATA will create separate columns for variables with different names, even if the information is the same. So, you will have to spend a little effort renaming variables. We will add the individual-level data to the household level data and create a new file called simply urban.dta

```
clear
use "C:\urbanhhold.dta", replace
merge yhid using "C:\urbanind.dta"
tab _merge
drop _merge
rename yhid hid
rename yregion1 region1
rename yhhsize hhsize
rename ytothhx tothhx
rename ytothhy tothhy
rename yweit weit
rename ysex male
rename yage age
rename ylfs lfs
rename yeducat educat
sort hid
save "C:\urban.dta", replace
```

For rural households the codes area as follows:

```
clear
use "C:\ruralhhold.dta", replace
merge ahid using "C:\ruralind.dta"
tab _merge
drop _merge
rename ahid hid
rename aregion1 region1
rename ahhsize hhsize
rename atothhx tothhx
rename atothhy tothhy
rename aweit weit
rename asex male
rename aage age
rename alfs lfs
rename aeducat educat
sort hid
save "C:\rural.dta", replace
```

Now that we have our urban and rural files, we can merge the two together to get. Before you get to the regression model, one of the first things you might like to try is to see if you can get back information about the country's demographics. For instance, if a the documentation of the household survey says that you have a nationally representative survey

```
clear
use "C:\urban.dta"
merge hid using "C:\rural.dta"
tab _merge
rename _merge urb
replace urb = 0 if urb==2
save "C:\regressiondata.dta", replace
```

Step 4: Analyse the data

Before you get to the regression model, one of the first things you might like to try is to see if you can get back information about the country's demographics. For instance, if a the documentation of the household survey says that you have a nationally representative survey that is stratified by urban-rural and regionally, then you should be able to get estimates of the urban and rural populations, as well as across regions, back using the sampling weights (which are called probability weights in STATA). You will notice that the weights available in this particular data set are not sampling weights, but if they were you would indicate in the syntax that you were using probability weights, "pw".[3]

Although tables are not necessarily the best way to present data, cross tabulation is a very useful tools in survey data analysis, because it can give you a quick way to summarize a particular aspect of the data. While the output is not reported, since the weights are not truly sampling weights, the following codes would be used to get an idea about the distribution of the population across regions, urban-rural, and nationally.

```
gen one = 1
table reg urb [pw=weit], c(sum one) col row
```

This is (possibly) your first cross-tabulation in STATA. By using the options "col" and "row", the output in the last column and the last row give you your totals. If you wanted to get an idea of what is the average consumption per household across regions, and across urban and rural areas, you would use the following codes

```
table reg urb [pw=weit], c(mean tothhx median tothhx) col row
```

Now let's move on to a simple regression model. Consider the following "consumption function"

$$\ln(c_i) = b_0 + \underbrace{b_1 \ln(y_i)}_{\substack{current\\ income}} + \underbrace{b_2 age_i}_{\substack{age\ \ of\\ household\\ head}} + \delta_{male} male + \delta_{urb} urb + \delta_{employ} employ + \delta_{inactive} inactive + e_i$$

where $\ln(c_i)$ is the natural log of household consumption, $\ln(y_i)$ is the natural log of household income, age_i is the age of the household head, $male$ is a dichotomous, or "dummy" variable takes a value of 1 if the household head is male, urb is a dummy variable that takes on a value of 1 if family is located in an urban area, employ is a dummy variable that takes on a value of 1 if the household head is employed, $inactive$ is a dummy variable that takes on a value of 1 if the household head is economically inactive, and e_i is the random disturbance term.

[3] In STATA there are four types of weights that you can choose from. The most common are sampling weights, or "pweights," which reflect the inverted probability that the sampling design "selected" the observation. The other three options in STATA, which are not generally used for household survey data analysis are frequency weights, or "fweights", which tell you how many duplicate observations there are in the data, analytic weights, or "aweights," which divide the observation by the variance, to down-weight any large observations, and finally, importance weights, or "iweights," which tell you in a sense how important an observation is.

```
replace male = 0 if male==2
*** Females were coded as 2 in the data, so the above line recodes ***
*** them as zero, so that the dummy variable reflects males        ***
tab lfs, gen(d)
rename d1 employ
rename d2 unemp
rename d3 inactive

gen lncon = log(tothhx)
gen lninc = log(tothhy)
reg lncon lninc age male urb employ inactive, robust
log close
```

```
Linear regression                    Number of obs = 2294
                                     F(6, 2287) =  161.48
                                     Prob > F = 0.0000
                                     R-squared = 0.2930
                                     Root MSE = 1.5383
Lncon        Coef.        Std. Err.  t      P>t   [95% Conf.   Interval]

lninc        .621761      .029062    21.39 0.000 .5647705     .6787516
age         -0.0050924    .0029072   -1.75 0.080 -.0107934    .0006086
male         0.2178578    .0773759    2.82 0.005  .0661236    .369592
urb          0.3271723    .1042616    3.14 0.002  .122715     .5316296
employ      -0.2142715    .1649641   -1.30 0.194 -.5377664    .1092234
inactive    -0.0634109    .0910623   -0.70 0.486 -.2419842    .1151625
_cons        3.364355     .3537642    9.51 0.000 2.670623     4.058087
```

You see that the elasticity with respect to log of income is 0.62. For every additional year, on average is associated with a half a percent reduction in consumption. Male headed households, households located in urban areas both tend to have higher consumption. While households in which the head of household is either employed or economically inactive tend to have lower consumption. Notice that weights were not used here. They're not too useful in this context, because you can't be sure if a household's characteristics are nationally representative (other than the region they're from or whether they are urban or rural).

References

Deaton, Angus. (1997) *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press: Baltimore, Maryland.