Review: Introduction to STATA

1. Introduction

The aim of this session is to familiarize students with the STATA environment. To do this, we use a subset of variables of the Peruvian Household National Survey (Encuesta Nacional de Hogares-ENAHO 2011). We use only a subset of variables. To start, use the following routine:

```
use conglome vivienda hogar dominio estrato codperso p203 p207 p208a p209
using "enaho01-2011-200.dta", clear
```

A. To begin the empirical work, inspect the database by executing the following operations:

   i. Assign the path to the directory where you will work using a global labeled "path". Recall the discussion about globals and locals in the companion guide. Then, create a log file using the name `PS1.log` (hint: define the global "path" by typing `global path="herethefolderaddress"` where the stuff in parentheses is the address of the folder where your files are located).

   ii. Provide a description of the variables in the database using the household identifier to sort them (hint: use commands `use` and `sort`).

B. Now, you must determine whether the unit of analysis is correct. To do this, we need to check whether the identifiers for individuals in the dataset are uniquely associated to each individual. Check for duplicated observations, creating an indicator variable equal to 1 in case an observation is duplicated and 0 otherwise. Then delete these observations from the dataset. (Hint: use the command `duplicates` on `conglome vivienda hogar codperso` using the option `tag`).

C. Usually identifiers are composed of more than one variable associated with different levels of aggregation. To create a single identifier, let's follow these steps:

   i. Replicate the original variables `conglome vivienda hogar codperso` and name them `conglome2 vivienda2 hogar2 codperso2` respectively. Convert them to numerical variables (Hint: use command `destring` with option `replace`).

   ii. Perfom the opposite operation by converting variables `conglome2 vivienda2 hogar2 codperso2` into string variables (Hint: use command `tostring` with option `replace`).

   iii. Convert the numerical variables `conglome vivienda hogar codperso` in string variables; add zeros as applicable for all observations to have the same length. Finally, create a variable called `hhid` as the sum of the `conglome2 vivienda2 hogar2` variables and variable `individ` as the sum of `conglome2 vivienda2 hogar2 codperso2` (hint: use the command `tostring`. See the addendum for more details).

iv. Perform the reverse exercise. From `individ` recover the `conglome2 vivienda2 hogar2 codperso2` variables (Hint: use the `substr` function from the command `generate`). List all these variables for the first 10 observations (hint: use command `list`).

D. Much of the work with databases may be reduced to creating indicators and, in particular, to ensure consistency. To illustrate this, let's create the following variables:

i. From the individual information, generate a variable for family size per household (`famsize`) using the command `egen`. This is an extremely useful command, so take your time exploring the multiple options it offers. Notice that we want a variable that assigns the household size to all the members of a household. (Hint: there are many ways to do this, so feel free to use the way you prefer. A simple way to do it would be to create an auxiliary variable equal to 1 for all the observations and then use the `egen` command with the option `sum( )` applied over the auxiliary variable and the option `by( )` using the household identifier. Please, check the help file for the command `egen`).

ii. From p203 (relationship with the household head) and p207(sex), create the variable `sexhead` (1 = Male, 0 = Female) that recover household head sex. Notice that, as in the previous case, we want to assign the information of the household head to all household members (hint: create an auxiliary variable for the relevant characteristic and use command `egen, by` to assign this information to all household members). Check for missing categories and replace them with `"."` . Recode and add labels when appropriate.

iii. Create a new database with the average at household level for variables `famsize` and `sexhead` using the command `collapse`. List the observations for the variable `house` for values below 10. Save this database with the name `temphh.dta`.

E. Now perform a similar exercise for the `sumaria-2011.dta` :

i. Upload the variables `conglome vivienda hogar pobreza inghog2d`. Sort the dataset by `conglome vivienda hogar`. Save this database with the name `temp_sumaria.dta`.

F. Now we have two databases, both containing relevant information. Using the current dataset as master dataset, add the variables available in `temphh.dta` (hint: use the `merge` command). Save this new dataset under the name `hh_2011.dta`.

G. Using the dataset previously created, we are able to start the analysis of the data.

i. Let's now create a database with the most important statistics by level of household poverty. Then, export it to an excel file (use the following name: `HW1_yourlastnamename.xls`). Discuss your results. (Hint: use the `collapse` command and then `xml_tab`).

ii. Use a t-test to evaluate whether there are statistical differences between the socio-economic characteristics you chose in the previous part according to household level poverty. Which dimensions, if any, seem to be more different between poor and non-poor households?