Tutorial: Creating and labeling new variables

Sometimes you may need to create a new variable that is a function of one or more existing variables. We delve into data management or data cleaning in this module, focusing on new variables.

- **Create a new continuous variable** called `agesq1` that is the square of age at exam 1 using the generate command.

    - `generate agesq1=age1*age1`
    - `drop agesq1`
    - `generate agesq1=age1^2`
    - Note that you could have also selected *Data/Create or change variables/Create new variable* to bring up the “Generate a New Variable” dialog box. Type `agesq1` into the box under “Generate Variable”. Type `age1^2` into the “Contents” box.

Missing data in continuous variables: If a person’s age was not recorded at exam 1 (i.e. `age1 == .`), then by default, Stata defines `agesq1 == .`

- **Create a new categorical variable** called `agecat1` based on participants’ ages at exam 1, with the following three categories: 30-39, 40-49, 50-59, and 60-70 years old.

    -
      ```
      generate agecat1=.

      replace agecat1=1 if age1 < 40

      replace agecat1=2 if age1 >= 40 & age1 < 50

      replace agecat1=3 if age1 >= 50 & age1 < 60

      replace agecat1=4 if age1 > 60 & age1 < .
      ```

Missing data in categorical covariates: Stata treats missing values as equal to positive infinity – Make sure you explicitly take care of missing values when you create a new categorical variable!

- **Errors in Stata**: Examine what happens in the command window when you execute the `generate` and replace command. What happens if you make an error?

- It is always a good idea to **double check your work** after creating new variables!

    -
      ```
      summarize age1 agesq1

      tabulate agecat1, missing
      ```

Labels

For the sake of clarity, you may want to give labels to your data. Labels can be given to the entire **dataset; to the individual variables; or to the values of individual variables**. Labels allow you to keep a detailed explanation of the contents of a dataset.

Variable Labels

The variables that were in the original `fhs.dta` dataset are already labeled. **Looking in the variables window**, we see that there are two columns in Stata – one for the variable name and one for the variable label. But, the new variables that we just constructed do not have labels.

- First, we will **add labels to the new variables** we created, `agesq1` and `agecat1`.

  -
    ```
    label variable agesq1 "Age squared, exam 1"

    label variable agecat1 "Categorical age, exam 1"
    ```
  - Note that you could have also clicked on *Data/Labels/Label Variable* and moved the cursor and click in the box under “Variable”. Type `agecat1` or select `agecat1` from the Variables window and enter an appropriate variable label. These labels will appear in any tables or graphs that you make with this data.

Value Labels

- **Construct a table** of the frequency of males versus females at exam 1.

  -
    ```
    tabulate sex1
    ```

Notice that we have one line that says “Male” and one that says “Female”. In Stata, `sex1` is stored as a numeric integer variable, and someone has manually added the labels “Male” and “Female” to the variable to avoid confusion when constructing tables.

- **Add labels to the values** of `agecat1` to ensure that we remember how the 4 categories of `agecat1` are defined. We will tabulate the variable before and after adding labels, to examine the difference.

  -
    ```
    tabulate agecat1

    label define agecatlabel 1 "30-39" 2 "40-49" 3 "50-59" 4 "60-70"

    label values agecat1 agecatlabel

    tabulate agecat1
    ```