Practicum 2: Basic data manipulations

Harvard Catalyst Certificate in Applied Biostatistics

Goal: This practicum will go over how to do basic data manipulations in STATA such as creating new variables. For the first eight practicums, the commands typed into the command window will be provided, and the appropriate drop down menus will also be shown.

Worked example

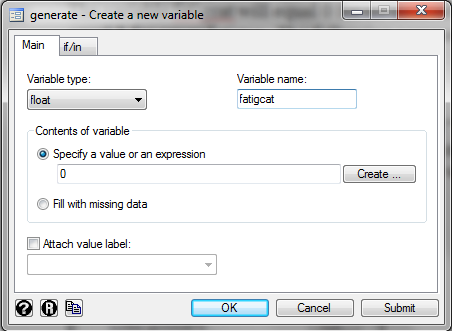
**Generating new variables**

Please load the dataset into STATA by clicking the dataset link for Practicum 1 on the webpage. STATA should open automatically and the dataset should be loaded.

Often, we would like to tabulate the number of people who have no/mild, moderate and severe fatigue. For our example, mild fatigue is a fatigue score less than 30, moderate fatigue is a score between 30 and 40, and severe fatigue is a score greater than 40. To calculate the number and proportion of people who fit in these categories, we must create a new variable called “fatigcat”. The variable fatigcat will equal 0 for no/mild fatigue, will equal 1 for moderate fatigue and will equal 2 for sever fatigue. The following commands are used to create the categories of interest:

*generate fatigcat=0*

This command generates a new variable called “fatigcat” and assigns it a value of zero for all people. To complete this using the drop-down menus, you use *Data/Create or change data/Create new variable*. The window below will open.

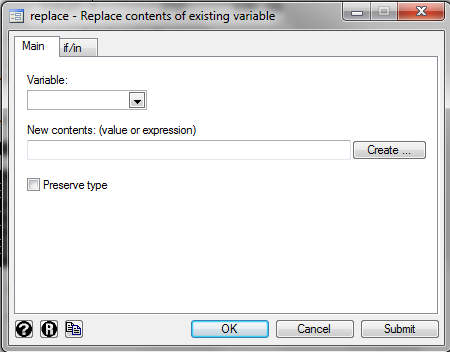


In this simple scenario, you just enter “fatigcat” as the Variable name box and 0 as the Contents of variable.

Next, we use the following command to change the value of “fatigcat” to 1 for patients with moderate fatigue:

*replace fatigcat=1 if fatig>=30 & fatig<=40*

For this command, we tell STATA to change the value of fatigcat if a specific condition is met. Specifically, “fatig” must be greater than or equal 30 AND less than or equal 40. The “&” symbol shows that both conditions must be met. Note when you type this that STATA tells you a specific number of changes were made. To complete this using the drop down menus, you use *Data/Create or change data/Change contents of variable*. The window below will open.



In this menu, you will need to type “fatigcat” in the Variable box and “1” in the New contents box. Then in the “if/in” tab, you will need to type the condition listed above: “fatig>=30 & fatig<=40”.

Finally, to change the value of “fatigcat” to 2 for patients with severe fatigue:

*replace fatigcat=2 if fatig>40*

Again, note that STATA tells you a specific number of changes were made.

Now to calculate the number and proportion of people with each level of fatigue, we can use the *tabulate* command as we did in the previous practicum:

*tabulate fatigcat*

. tabulate fatigcat

fatigcat | Freq. Percent Cum.

------------+-----------------------------------

0 | 167 66.27 66.27

1 | 42 16.67 82.94

2 | 43 17.06 100.00

------------+-----------------------------------

Total | 252 100.00

Note that the table above shows the number of people in each category in the “Freq” column, the percent (or proportion) of total in the “Percent” column and the cumulative percent in the “Cum” column. 66.27% of our sample had no/mild fatigue

Try on your own

Using the commands listed above, generate a new variable to represent no, mild, and moderate/severe depression, “depcat”. For this exercise we will define no depression to be <25, mild depression to be between 25 and 36, and moderate/severe depression to be >36.

Please note that the depression variable is missing one value. Using the approach we present above, how is the patient with missing depression treated? How could we adjust the code from above so that patients with missing depression score also have missing depression category?

Hint: *replace depcat=. if dep==.*

After properly accounting for the patient with missing data, how many people have moderate to severe depression? What proportion of the total does this represent?