Updated July 2018

Recoding and Labeling Variables

This set of notes describes how to use the computer program Stata to recode variables and save them as new variables as well as how to label variables. It assumes that you have set Stata up on your computer (see the "Getting Started with Stata" handout), and that you have read in the set of data that you want to analyze (see the "Reading in Stata Format (.dta) Data Files" handout).

In Stata, most tasks can be performed either by issuing commands within the "Stata command" window, **or** by using the menus. These notes illustrate both approaches, using the data file "GSS2016.DTA" (this data file is posted here: https://canvas.harvard.edu/courses/53958).

Recoding variables

Recoding categorical or quantitative variables can be useful in a number of circumstances. For example, you might want to use fewer, more aggregated categories than those used in collecting the data, change the ordering of a variable's categories for some reason, or recode a quantitative variable as a categorical variable.

Recoding categorical variables

Before you recode a categorical variable, you first need to figure out which numerical values correspond to each of its categories in the data set. To do so, you can use the "codebook" command. For example, if you wanted to figure out the numerical values that correspond to the nine different regions in the "region" (region of interview) variable in GSS2016.DTA, you would type:

```
codebook region
```

and get the following output:

```
            type:  numeric (byte)
           label:  REGION

           range:  [1,9]                        units:  1
   unique values:  9                        missing .:  0/2,867

      tabulation:  Freq.   Numeric  Label
                     175         1  new england
                     313         2  middle atlantic
                     502         3  e. nor. central
                     193         4  w. nor. central
                     550         5  south atlantic
                     205         6  e. sou. central
                     297         7  w. sou. central
                     235         8  mountain
                     397         9  pacific
```

From this we learn, for example, that the value "New England" is coded as 1, whereas the value "East South Central" is coded as 6.

After inspecting the coding of a variable with the "codebook" command, you use the "recode" command to change it. Using the command line window, issue the command:

```
recode varname (rule1 "valuelabel1") (rule2 "valuelabel2"). . .,
generate(newvarname)
```

(Note: enter the command in Stata on one line. "Generate…" is displayed on the second line above due to Microsoft Word formatting.).

"Varname" is the name of the variable you want to recode. After you select the variable, in parentheses you generate "rules" ("rule1", "rule2", etc.) that specify which values you would like to group together when recoding (more on this below) and how you want to label each recoded group ("valuelabel1", "valuelabel2", etc.). After specifying these recoding rules, you type the option "generate" after the comma to generate a new variable that will store your recoded values. "Newvarname" is the name of this new variable.

There are three main types of rules you can use to recode a variable. They are listed in the table below, which is a copy of the one included in the online Stata manual.[1]

rule	Example	Meaning
# = #	3 = 1	3 recoded to 1
# # = #	2 . = 9	2 and . recoded to 9
#/# = #	1/5 = 4	1 through 5 recoded to 4

The first rule changes one old value to a different new one. The second one combines two (or more) old values into a single new one. The third combines a range of consecutive old values into a single new one. Parentheses enclose each distinct recoding rule.

For example, to recode the nine regions in the "region" variable into four regions and to save the recoded values as a new variable ("region_4cat") in the GSS2016.DTA, you would type the command (all on one line):

```
recode region (1/2=1 "Northeast") (3/4=2 "Midwest") (5/7=3 "South")
(8/9=4 "West"), gen(region_4cat)
```

1 https://www.stata.com/manuals13/drecode.pdf

To inspect the frequencies of the new variable, use the tabulate command:

```
. tab region_4cat

  RECODE of |
     region |
 (region of |
 interview) |      Freq.     Percent        Cum.
------------+-----------------------------------
  Northeast |        488       17.02       17.02
    Midwest |        695       24.24       41.26
      South |      1,052       36.69       77.96
       West |        632       22.04      100.00
------------+-----------------------------------
      Total |      2,867      100.00
```

To confirm that you recoded the values as you intended, you can crosstabulate the region and region_4cat variables to compare frequencies (see the "Crosstabulations of Two Categorical Variables" handout for more information):

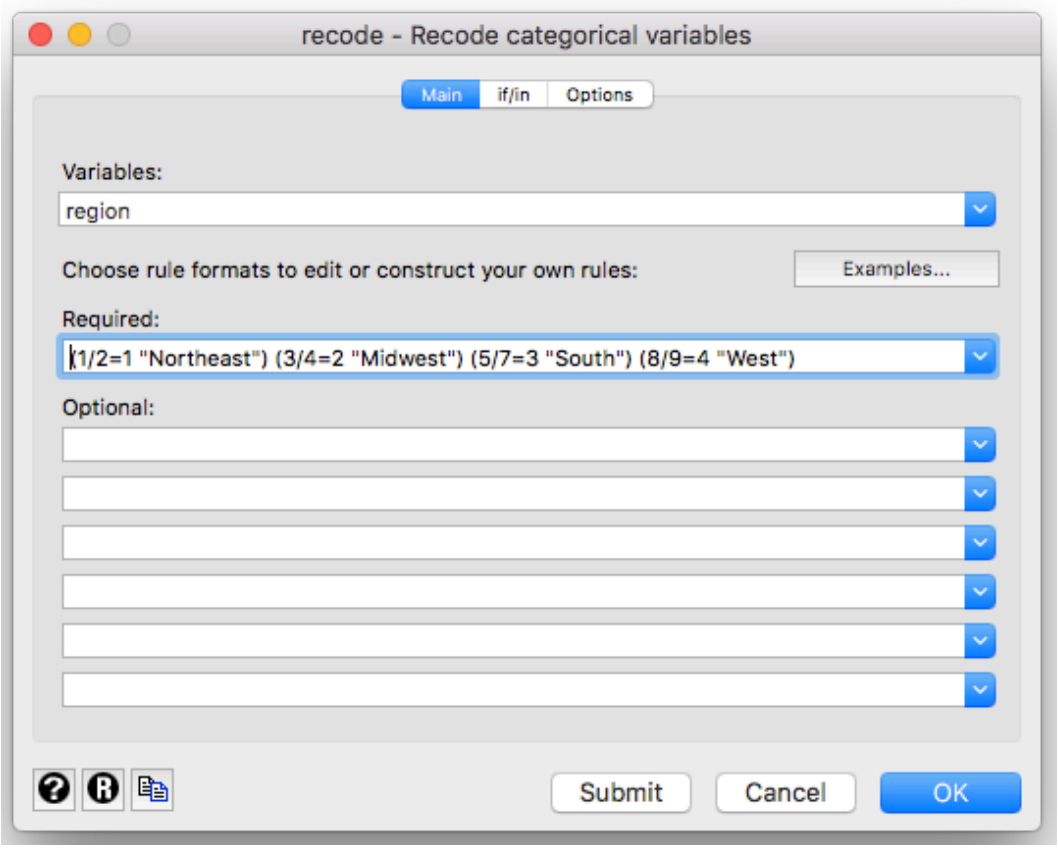
```
. tab region region_4cat

      region of |    RECODE of region (region of interview)
      interview | Northeast    Midwest      South       West |     Total
----------------+--------------------------------------------+----------
    new england |       175          0          0          0 |       175
middle atlantic |       313          0          0          0 |       313
e. nor. central |         0        502          0          0 |       502
w. nor. central |         0        193          0          0 |       193
 south atlantic |         0          0        550          0 |       550
e. sou. central |         0          0        205          0 |       205
w. sou. central |         0          0        297          0 |       297
       mountain |         0          0          0        235 |       235
        pacific |         0          0          0        397 |       397
----------------+--------------------------------------------+----------
          Total |       488        695      1,052        632 |     2,867
```

To recode using the menus, you would proceed as follows

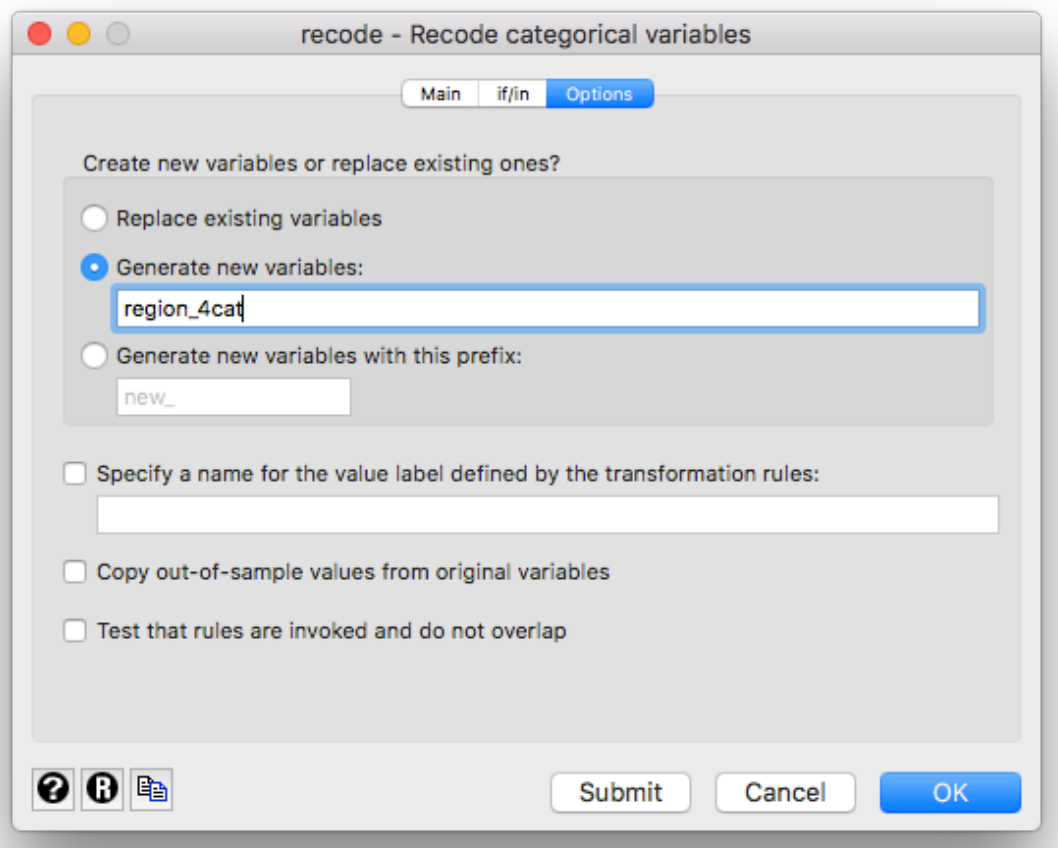
click on “Data”
click on “Create or Change Data”
click on “Other variable-transformation commands”
click on “Recode categorical variable”

A window like this will then open up:



You fill in the variable you want to recode in the “Variables:” box either by typing the variable name or selecting it from the drop-down menu. You then enter the recoding rules and value labels for categories in the “Required:” box; don’t click “OK” quite yet!

Then click on the “Options” tab to open the following window:



Click the “Generate new variables” box and then fill in the new name for your recoded categorical variable in the text box. Then click “OK.”

Recoding quantitative variables

You might want to recode a quantitative variable as a categorical variable. Before you recode the quantitative variable, it is a good idea to inspect its coding. To do so, you can use the "codebook" and "summarize" commands. For example, if you wanted to inspect the coding of the variable "educ" (highest year of school completed) in GSS2016.DTA,

you would type:

```
codebook educ
```

and get the following output:

```
                type:  numeric (byte)
               label:  LABAF, but 21 nonmissing values are not labeled

               range:  [0,20]                       units:  1
        unique values:  21                   missing .:  0/2,867
    unique mv codes:  2                   missing .*:  9/2,867

            examples:  12
                       12
                       14
                       16
```

and you would type:

```
sum educ, detail
```

and get the following output:

```
. sum educ, detail

                highest year of school completed
-------------------------------------------------------------
      Percentiles      Smallest
 1%             6              0
 5%             9              0
10%            11              1       Obs                2,858
25%            12              1       Sum of Wgt.        2,858

50%            13                      Mean           13.73723
                       Largest          Std. Dev.      2.963886
75%            16             20
90%            18             20       Variance        8.78462
95%            19             20       Skewness       -.1855541
99%            20             20       Kurtosis        3.820384
```

From this we learn, for example, that the variable contains integer values only and the values range from 0 to 20 years.

After inspecting the coding of a variable, you use the "recode" command to change it. Using the command line window, issue the command:

```
recode varname (rule1 "valuelabel1") (rule2 "valuelabel2"). . .,
generate(newvarname)
```

(Note: enter the command in Stata on one line. "Generate…" is displayed on the second line above due to Microsoft Word formatting.).

"Varname" is the name of the variable you want to recode. After you select the variable, in parentheses you generate "rules" ("rule1", "rule2", etc.) that specify which values you would like to group together when recoding (more on this below) and how you want to label each recoded group ("valuelabel1", "valuelabel2", etc.). After specifying these recoding rules, you type the option "generate" after the comma to generate a new variable that will store your recoded values. "Newvarname" is the name of this new variable.

There are three main types of rules you can use to recode a variable. They are listed in the table below, which is a copy of the one included in the online Stata manual.[2]

rule	Example	Meaning
# = #	3 = 1	3 recoded to 1
# # = #	2 . = 9	2 and . recoded to 9
#/# = #	1/5 = 4	1 through 5 recoded to 4

The first rule changes one old value to a different new one. The second one combines two (or more) old values into a single new one. The third combines a range of consecutive old values into a single new one. Parentheses enclose each distinct recoding rule.

For example, to recode the "educ" variable into three categories (less than 12 years of schooling, 12 years of schooling, and more than 12 years of schooling) and to save the recoded values as a new variable ("educ_3cat") in the GSS2016.DTA, you would type the command (all on one line):

```
recode educ (0/11=1 "Less than 12 years") (12=2 "12 years") (13/20=3
"More than 12 years"), gen(educ_3cat)
```

2 https://www.stata.com/manuals13/drecode.pdf

To inspect the frequencies of the new variable, use the tabulate command:

```
. tab educ_3cat

    RECODE of educ |
  (highest year of |
 school completed) |      Freq.     Percent        Cum.
-------------------+-----------------------------------
Less than 12 years |        381       13.33       13.33
          12 years |        824       28.83       42.16
More than 12 years |      1,653       57.84      100.00
-------------------+-----------------------------------
             Total |      2,858      100.00
```

To confirm that you recoded the values as you intended, you can crosstabulate the educ and educ_3cat variables to compare frequencies (though this is not recommended if your quantitative variable has many unique values, such as income):

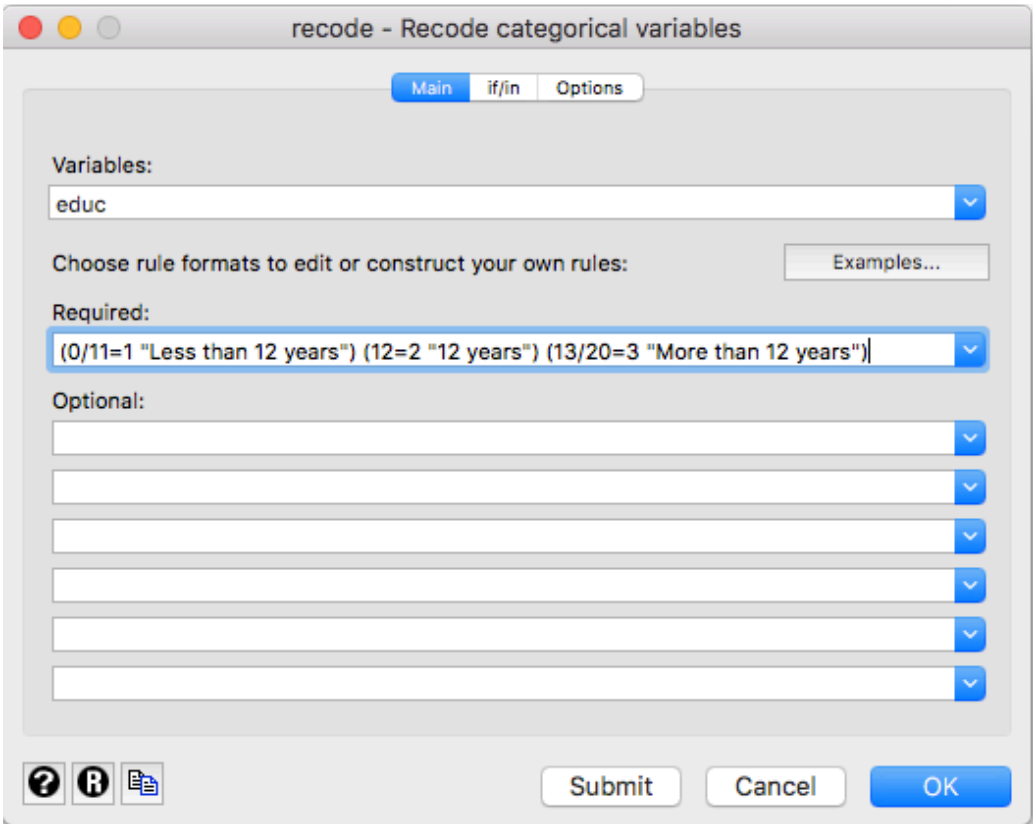
```
. tab educ educ_3cat

   highest |
   year of | RECODE of educ (highest year of
    school |        school completed)
 completed | Less than   12 years  More than |     Total
-----------+---------------------------------+----------
         0 |         2           0          0 |         2
         1 |         3           0          0 |         3
         2 |         3           0          0 |         3
         3 |         3           0          0 |         3
         4 |         2           0          0 |         2
         5 |         4           0          0 |         4
         6 |        31           0          0 |        31
         7 |        18           0          0 |        18
         8 |        48           0          0 |        48
         9 |        59           0          0 |        59
        10 |        90           0          0 |        90
        11 |       118           0          0 |       118
        12 |         0         824          0 |       824
        13 |         0           0        242 |       242
        14 |         0           0        359 |       359
        15 |         0           0        137 |       137
        16 |         0           0        485 |       485
        17 |         0           0        108 |       108
        18 |         0           0        149 |       149
        19 |         0           0         63 |        63
        20 |         0           0        110 |       110
-----------+---------------------------------+----------
     Total |       381         824      1,653 |     2,858
```

To recode using the menus, you would proceed as follows

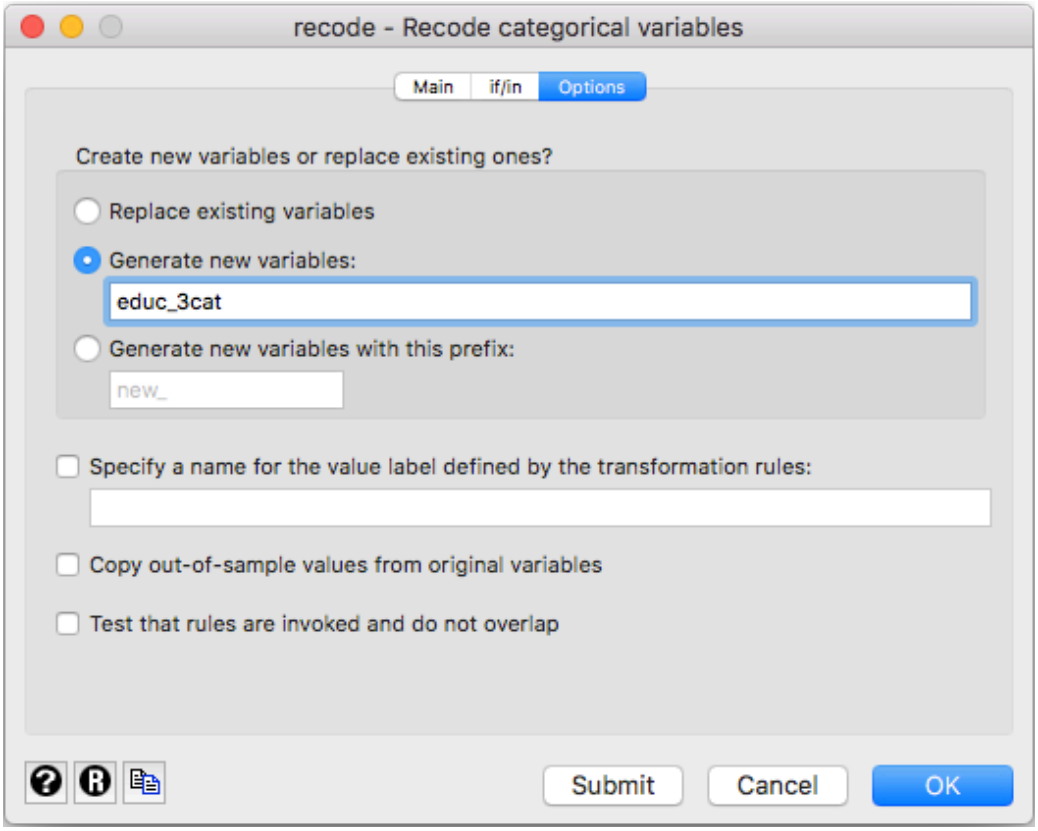
click on “Data”
click on “Create or Change Data”
click on “Other variable-transformation commands”
click on “Recode categorical variable”

A window like this will then open up:



You fill in the variable you want to recode in the “Variables:” box either by typing the variable name or selecting it from the drop-down menu. You then enter the recoding rules and value labels for categories in the “Required:” box; don’t click “OK” quite yet!

Then click on the “Options” tab to open the following window:



Click the “Generate new variables” box and then fill in the new name for your recoded categorical variable in the text box. Then click “OK.”

Labeling variables

Labeling *variables* is different from labeling *values*, which was demonstrated in the recoding command (e.g. value 1 is labeled "Northeast" in the variable region_4cat).

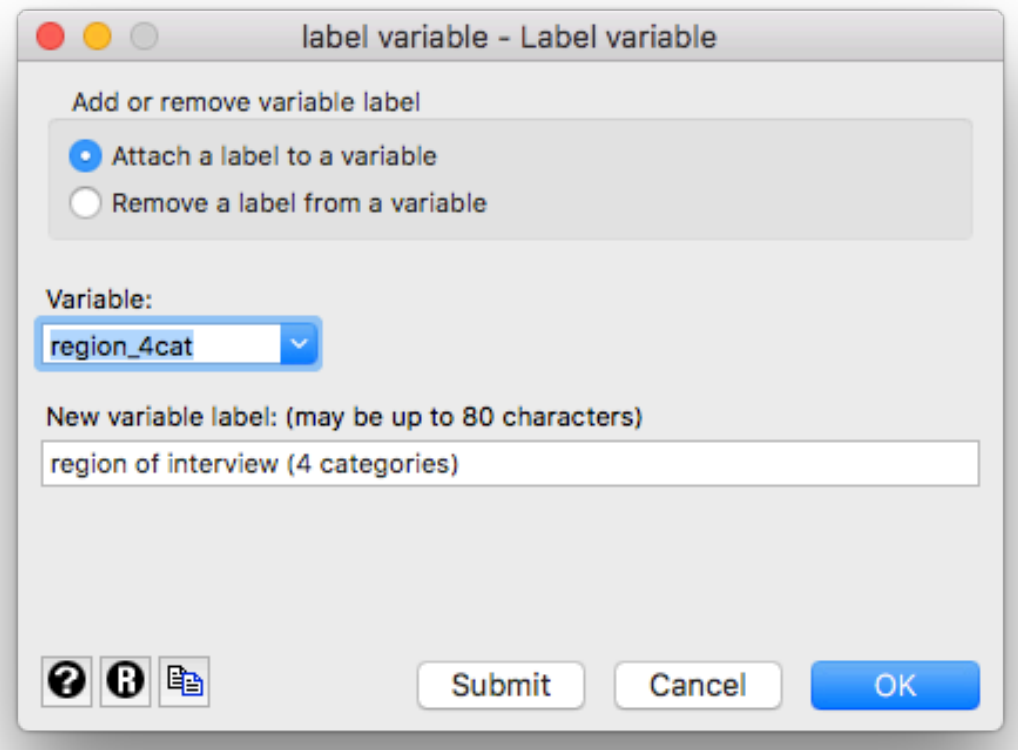
To label the variable region_4cat you would type:

```
label var region_4cat "region of interview (4 categories)"
```

It is probably easier to label the new variable using the command line as illustrated above, but to do it using the menus, you would proceed as follows

click on "Data"
click on "Data Utilities"
click on "Label Utilities"
click on "Label Variable"

A window like this will then open up:



Click the "Attach a label to a variable" box and then fill in the name of your categorical variable in the text box. Next, under "New variable label:" enter the variable label in the text box. Then click "OK."