How can I recode continuous variables into groups in Stata?

Authored by

stats writer

June 30, 2024

To recode continuous variables into groups in Stata, you can use the "recode" command. This command allows you to specify the ranges or values of the continuous variable and assign them to a new group variable. Alternatively, you can use the "generate" command to create a new categorical variable based on the values of the continuous variable. This process of recoding allows for easier analysis and interpretation of data as it breaks down a large range of values into distinct categories.

How can I recode continuous variables into groups? | Stata FAQ

There may be times that you would like to convert a continuous variable into groups. For example, you might want to convert
a continuous reading score that ranges from 0 to 100 into 3 groups (say low,
medium and high). You can use egen with the cut() function
to do this quickly and easily, as illustrated below. We will illustrate this
with the hsb2 data file with a variable called
write that ranges from 31 to 67.

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

summarize write

Variable | Obs Mean Std. Dev. Min Max

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

write | 200 52.775 9.478586 31 67

We can use egen with the cut() function to make a variable called
writecat that groups the variable
write into the
following 4 categories.

30 up to (but not including) 40
40 up to (but not including) 50
50 up to (but not including) 60
60 up to (but not including) 70

```
egen writecat = cut(write), at(30,40,50,60,70)
```

The tabstat command below is used to verify that the data are grouped as we expected. We can see that when writecat is in the lowest category (30) that write ranges from 31 to
39, and so forth as we expect, e.g., the values when writecat is in category
30 correspond to write having values of 30 up to (but

not including) 40.

tabstat write, by(writecat) stats(min max)

Summary for variables: write
Group variable: writecat

writecat | Min Max
---------+--------------------
30 | 31 39
40 | 40 49
50 | 50 59
60 | 60 67
---------+--------------------
Total | 31 67

Here we use the same command but
our last category is from 50 to 60. As you see, it generates a missing
value because there are a number of values that are 60 or higher and thus outside of the range we specified. This
shows that if there are values outside of the range you provide, those will be

assigned a missing value.

egen writecat2 = cut(write), at(30,40,50,60)
(53 missing value generated)

If we use the icodes
option, cut() will create integer codes 0, 1, 2 and so forth. In
the example below, you can see that it created codes 0, 1, 2 and 3.

egen writecat3 = cut(write), at(30,40,50,60,70) icodes
tabstat write, by(writecat3) stats(min max)

Summary for variables: write
Group variable: writecat3

writecat3 | Min Max
----------+--------------------
0 | 31 39
1 | 40 49
2 | 50 59
3 | 60 67
----------+--------------------
Total | 31 67

If you use the label option (which automatically implies icode), then it will create integer values like above, but it will also create value labels. As you see below, the variable writecat4 is labeled 30- 40- 50- and 60-.

```
egen writecat4 = cut(write), at(30,40,50,60,70) label
tabstat write, by(writecat4) stats(min max)

Summary for variables: write
Group variable: writecat4

writecat4 | Min Max
----------+--------------------
30- | 31 39
40- | 40 49
50- | 50 59
60- | 60 67
----------+--------------------
Total | 31 67
------------------------------
```

We use the nolabel option on the table command to suppress
the display of the value labels and you can see that the variable really is
coded 0, 1, 2 and 3.

```
tabulate writecat4, nolabel

writecat4 | Freq. Percent Cum.
------------+-----------------------------------
0 | 21 10.50 10.50
1 | 51 25.50 36.00
2 | 75 37.50 73.50
3 | 53 26.50 100.00
------------+-----------------------------------
Total | 200 100.00
```

If you prefer, you can ask cut()
to choose the cutoffs to form groups with approximately the same number per
group. Below we request the creation of 4 (roughly) equally sized groups.

```
egen writecat5 = cut(write), group(4) label
```

```
table write writecat5

----------------------------------
writing | writecat5
score | 31- 45.5- 54- 60-
----------+-----------------------
31 | 4
33 | 4
35 | 2
36 | 2
37 | 3
38 | 1
39 | 5
40 | 3
41 | 10
42 | 2
43 | 1
44 | 12
45 | 1
46 | 9
47 | 2
49 | 11
50 | 2
52 | 15
```

53 | 1

54 | 17

55 | 3

57 | 12

59 | 25

60 | 4

61 | 4

62 | 18

63 | 4

65 | 16

67 | 7

For more information, see the help or reference manual about
egen.