The SAS PCTL9 (Quantile) Macro

Ellen Hertzmark and Donna Spiegelman

December 13, 2016

Abstract

The %PCTL9 macro is intended to make any desired number of quantiles for a list of variables. It can also make quantile indicators and median-score trend variables. A subset of the data can be used to determine the quantile boundaries.

Keywords: SAS, macro, quantiles, deciles, quintiles

Contents

1 Description

%PCTL9 is a SAS macro that makes any desired number of quantiles for a list of variables. The ordinal score (i.e. quantile number from 0 to *NUMQUANT* -1) is automatically included in the output data set. A subset of the data (e.g. the controls in a case-control study) can be used to determine the quantile boundaries.

2 Invocation and Details

In the description below, values given after the '=' are the default values for the parameters.

After you make the SAS data set with the variables for which you want to compute quantiles, type

```
%pctl9(
 data=<input data set>,
        REQUIRED
 varlist=<list of variables for which you want
         quantiles generated>,
         REQUIRED
 numquant=5 <number of quantiles desired>,
            REQUIRED
 quantfor= <a condition to subset the data,
            if you want to base the quantiles on
            a subset of the data, e.g. quantfor=caco eq 2 >
 mscore=T <whether the new data set (outdat= ) should contain
            quantile medians>
            Note:  if a variable is missing, the median score
            will also be missing.
            TRUE or FALSE ,
            OPTIONAL
 quantname=quint <prefix to be used to name the ordinal score
                  for each variable>,
                  For example, if you are making quintiles of
                  X, Y, and Z, the ordinal score variables will
                  be named QUINTX, QUINTY, and QUINTZ, respectively.
```

```
                       OPTIONAL
outdat=  <name of output data set>,
                     REQUIRED
indic=T <whether the new data set should have quantile indicators>
        The indicator variables will be named
        <quantname><variable name><quantile number>, e.g. quintx0,..,quintx
        and labelled
        i<quantname><variable name>_<quantile number>,
        e.g. iquintx_0,...,iquintx_4.
        T or F ,
        OPTIONAL
cutdat= <user-specified SAS data set name, which will have
        the cutpoints and medians>,
        OPTIONAL
indref=<list of reference levels for indicators for
       each variable, IN THE SAME ORDER AS THE VARLIST>,
       If you do not give a list here, the lowest quantile
       will be used for every variable.
       If you give only one value, it will be used for all
       the variables.
       OPTIONAL
usemiss=1  < if INDIC=T, whether to make a missing indicator
             variable when making indicators for the variables>
             OPTIONAL
pcut=F  <whether to print the cutpoints (i.e. what would be in
         cutdat>
         T or F
         OPTIONAL
prnt=F  <whether to print up to 10 lines from the output of proc rank>
         T or F
         OPTIONAL
where= <a conditional clause to subset the data,
       e.g. where= gender eq M or where=caco eq 2 >
       OPTIONAL
```

3 Example

For the example we use the dataset we used for the examples of %LGT-PHCURV9, to make quintiles for AGE and BMI. For purposes of the example, all nonessential variables in the dataset have been dropped.

Below is a list of the variables in the dataset before the %PCTL9 call.

```
--------------------------------------------------------------------------------


#     Variable    Type    Len

3     AGE         Num      8
2     BMI         Num      8
1     CASE        Num      8


--------------------------------------------------------------------------------
```

Here are the quintile statistics for the controls based on proc rank.

```
--------------------------------------------------------------------------------

/udd/stleh/doctn/pctl  Program pctlex   12DEC2016   15:46    stleh           2
data merge0 before calling pctl9
controls only

The MEANS Procedure

                       Analysis Variable : AGE

     Rank for
Variable AGE    N Obs       N         Minimum          Median         Maximum
--------------------------------------------------------------------------------
           0     5710     5710     24.0000000      27.0000000      28.0000000

           1     5432     5432     29.0000000      30.0000000      30.0000000

           2     5300     5300     31.0000000      31.0000000      32.0000000
```

```
       3     4319    4319     33.0000000      33.0000000      34.0000000

       4     5507    5507     35.0000000      36.0000000      46.0000000
-------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------
```

data merge0 before calling pctl9
controls only

The MEANS Procedure

Analysis Variable : BMI

Rank for Variable BMI	N Obs	N	Minimum	Median	Maximum
0	5291	5291	16.0700000	19.3100000	20.1200000
1	5212	5212	20.1400000	20.8000000	21.4600000
2	5196	5196	21.4700000	22.1500000	23.0200000
3	5333	5333	23.0300000	24.0500000	25.6100000
4	5236	5236	25.6300000	28.2900000	39.9900000

Here is the macro call to make quintiles for AGE and BMI using the control observations to determine the quantile boundaries.

```
%pctl9(data=merge0,
 varlist=age bmi, outdat=withquint, quantfor=case eq 0, cutdat=cuts);
```

Note that, except for the name of the *OUTDAT*, the macro call leaves the default values in place.

Here is the distribution of the quintiles made by %PCTL9 (from cutoffs derived from the controls).

quantile distributions in whole dataset

The FREQ Procedure

quintage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5920	21.84	5920	21.84
1	5605	20.68	11525	42.52
2	5449	20.11	16974	62.63
3	4445	16.40	21419	79.03
4	5683	20.97	27102	100.00

quintbmi	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5477	20.21	5477	20.21
1	5341	19.71	10818	39.92
2	5321	19.63	16139	59.55
3	5480	20.22	21619	79.77
4	5483	20.23	27102	100.00

If there are missing data, they will have missing ordinal and median scores.
Also note that for these "continuous" variables, the quantiles are not all the
same size. This relates partly to the fact that only the controls were used
to make the quintiles and partly to the "lumpiness" of the data, especially
for age.

Below is the proc contents for the data set WITHQUINT.

#	Variable	Type	Len	Label
3	AGE	Num	8	
2	BMI	Num	8	
1	CASE	Num	8	
5	medquintage	Num	8	median score for age
13	medquintbmi	Num	8	median score for bmi
4	quintage	Num	8	
8	quintage1	Num	3	iquintage_1
9	quintage2	Num	3	iquintage_2
10	quintage3	Num	3	iquintage_3
11	quintage4	Num	3	iquintage_4
7	quintagem	Num	3	missing quintage

Here is a printout of the dataset CUTS, made by %PCTL9.

#	Variable	Type	Len	Label
6	quintager	Num	3	iquintage_0
12	quintbmi	Num	8	
16	quintbmi1	Num	3	iquintbmi_1
17	quintbmi2	Num	3	iquintbmi_2
18	quintbmi3	Num	3	iquintbmi_3
19	quintbmi4	Num	3	iquintbmi_4
15	quintbmim	Num	3	missing quintbmi
14	quintbmir	Num	3	iquintbmi_0

----

cutoffs

```
                        m  m  m  m  m  m  m  m  m  m  m
                        i  i  i  i  i  e  e  e  e  e  a
         n  n  n  n  n  n  n  n  n  n  d  d  d  d  x
         q  q  q  q  q  q  q  q  q  q  q  q  q  q  q
         u  u  u  u  u  u  u  u  u  u  u  u  u  u  u
         i  i  i  i  i  i  i  i  i  i  i  i  i  i  i
         n  n  n  n  n  n  n  n  n  n  n  n  n  n  n
  _  _   t  t  t  t  t  t  t  t  t  t  t
  T  F   a  a  a  a  a  a  a  a  a  a  a  a  a  a  a
```

```
     Y R   g    g    g    g    g    g  g  g  g  g  g  g  g  g  g  g
 O P E   e    e    e    e    e    e  e  e  e  e  e  e  e  e  e  e
 b E Q   _    _    _    _    _    _  _  _  _  _  _  _  _  _  _  _
 s _ _   0    1    2    3    4    0  1  2  3  4  0  1  2  3  4  0

 1 0 5 5710 5432 5300 4319 5507 24 29 31 33 35 27 30 31 33 36 28
```

```
     m  m  m  m                      m  m  m  m  m
     a  a  a  a                      i  i  i  i  i
     x  x  x  x  n  n  n  n  n  n  n  n  n  n
     q  q  q  q  q  q  q  q  q  q  q  q  q  q
     u  u  u  u  u  u  u  u  u  u  u  u  u  u
     i  i  i  i  i  i  i  i  i  i  i  i  i  i
     n  n  n  n  n  n  n  n  n  n  n  n  n  n
     t  t  t  t  t  t  t  t  t  t  t  t  t  t
     a  a  a  a  b  b  b  b  b  b  b  b  b  b
     g  g  g  g  m  m  m  m  m  m  m  m  m  m
 O   e  e  e  e  i  i  i  i  i  i  i  i  i  i
 b   _  _  _  _  _  _  _  _  _  _  _  _  _  _
 s   1  2  3  4  0  1  2  3  4  0  1  2  3  4

 1 30 32 34 46 5291 5212 5196 5333 5236 16.07 20.14 21.47 23.03 25.63
```

```
     m     m     m     m     m     m     m     m     m     m
     e     e     e     e     e     a     a     a     a     a
     d     d     d     d     d     x     x     x     x     x
     q     q     q     q     q     q     q     q     q     q
     u     u     u     u     u     u     u     u     u     u
     i     i     i     i     i     i     i     i     i     i
     n     n     n     n     n     n     n     n     n     n
     t     t     t     t     t     t     t     t     t     t
     b     b     b     b     b     b     b     b     b     b
     m     m     m     m     m     m     m     m     m     m
 O   i     i     i     i     i     i     i     i     i     i
 b   _     _     _     _     _     _     _     _     _     _
 s   0     1     2     3     4     0     1     2     3     4

 1 19.31  20.8  22.15  24.05  28.29  20.12  21.46  23.02  25.61  39.99
```

WARNING: When two or more quantiles collapse into one interval for a

8

variable, a WARNING like the one below (for the variable `badvar`) will be output to the SAS .lst file

```
WARNING in PCTL macro run:  The macro did not make 5 quantiles
        for the variables badvar.
        The nonempty quantiles are shown in the output.
```

4   Computational Methods

The macro uses proc rank to group the observations in the *QUANTFOR* subset (or the whole dataset if *QUANTFOR* is null) into *NUMQUANT* groups. Observations with the same value of the original variable will always be in the same group. The macro then finds the minimum, median, and maximum values for the variable within each quantile.

A quantile for a variable contains all observations whose values are

```
>= the minimum for that quantile
AND
< the minimum for the next quantile
```

The highest quantile contains all observations whose values are greater than or equal to the minimum for the highest quantile. NOTE: Because of the asymmetry of the strict and weak inequalities, the ordinal scores for -X will not necessarily be the reverse of those for X. For truly continuous data this should not present a serious problem, but for "lumpy" data it may.

5   Warnings

6   Credits

Written by Ellen Hertzmark, Mathew Pazaris, and Donna Spiegelman for the Channing Laboratory. Questions can be directed to Ellen Hertzmark, `stleh@channing.harvard.edu`, (617) 432-4597.

7 See Also

See also a macro to make indicator variables from categorical variables, `/usr/local/channing/sasautos/indic3.sas`. This macro is used inside the %PCTL9 macro if you ask for indicator variables.