Power Transformations Using SAS/IML® Software

Ioannis C. Dimakos, Syracuse University, Computing & Media Services

Abstract

Sample SAS/IML® code is presented here that will assist users in performing power transformations to non-normally distributed variables. The code estimates a range of values of λ, selects the appropriate value, and assesses the normality of the transformed variable. The macrocode has been tested using SAS® software, release 6.09 for Unix platforms, and SAS 6.11 for Windows 95, but should work with other SAS releases, as well.

Power Transformations

When dealing with parametric statistics, one of the assumptions that must hold true is that the examined variables are normally distributed (Kirk, 1982). SAS users can use the SAS/BASE® UNIVARIATE procedure, with the optional statements NORMAL and PLOT, to test for normality and obtain stem-and-leaf and box-and-whisker plots of the variables' distributions. Other available options inlude the SAS/QC® CAPABILITY procedure with the QQPLOT statement for quantile-quantile plots (qq-plots) of the examined variables or SAS/INSIGHT® to obtain similar descriptive statistics and graphs of high resolution.

If the variables under examination are not normally distributed, several transformations are suggested based on the nature of the data. Counts can be transformed to near normality by taking their square roots. Correlation coefficients are transformed to Fisher's Zs. However, when the choice of transformation is not obvious, power transformations may be suitable for the job (Johnson & Wichern, 1993).

According to Box & Cox (1964) who coined the term, power transformations shrink large values of a variable X, while increasing small values of that variable. In general, the family of power transformations is defined as:

$$x^{\lambda} = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ lnx & \lambda = 0 \end{cases} \qquad (1)$$

which is continuous in λ for $x > 0$. One selects a value of λ that maximizes the function:

$$l(\lambda) = -\frac{n}{2} ln\left[\frac{1}{n}\sum_{j=1}^{n}(x_j^{\lambda} - \overline{x^{\lambda}})^2\right] + (\lambda - 1)\sum_{j=1}^{n} lnx_j \qquad (2)$$

with $\overline{x^{\lambda}}$ being defined as the arithmetic average of the transformed observations:

$$\overline{x^{\lambda}} = \frac{1}{n}\sum_{j=1}^{n} x_j^{\lambda} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_j^{\lambda}-1}{\lambda}\right). \qquad (3)$$

The Code

A macro `bctrans` for power transformations according to Box & Cox is outlined below. Using the IML procedure (SAS Institute, Inc., 1989), the macro estimates a range of λ values and their corresponding $l(\lambda)$ values based on formulæ 2 and 3. It further provides a scatterplot of the $l(\lambda)$ and λ values. The macro then selects the highest λ value, transforms the original variable based on this λ value and assesses the normality of the distribution of the transformed variable using the PROC UNIVARIATE procedure.

```
%macro bctrans(data=,out=,var=,r=,
               min=,max=,step=);

proc iml;
 use &data;
 read all var {&var} into x;
 n=nrow(x); one=j(n,1,1); mat=j(&r,2,0);
 lnx=log(x); sumlog=sum(lnx);
 start;
 i=0;
 do lam=&min to &max by &step;
  i=i+1;
  lambda=round(lam,.01);
  if lambda = 0 then xl=log(x);
   else xl=((x##lambda) - one)/lambda;
   mean=xl[:];
   d=xl-mean;
   ss=ssq(d)/n;
   l=-.5*n*log(ss)+((lambda-1)*sumlog);
   mat[i,1] = lambda;
   mat[i,2] = l;
 end;
 finish;
 run;
 print "Lambdas and their l(lambda) values",
        mat[format=8.3];
 create lambdas from mat;
 append from mat;
quit;

data lambdas;
 set lambdas;
 rename col1=lambda col2=l;
run;

proc plot data=lambdas nolegend;
 plot l*lambda;
 title 'lambda vs. l(lambda) values';
run;
quit;

proc sort data=lambdas;
 by descending l;
run;

data &out;
 set lambdas;
 if _n_>1 then delete;
run;

proc print data=&out;
 title 'Highest lambda and l(lambda) value';
run;

proc iml;
 use &data;
 read all var {&var} into old;
 use &out;
 read all var {lambda l} into power;
 if lambda=0 then new=log(old);
 else new=old##power[1];
 create final from new;
 append from new;
quit;

data final;
 set final;
 rename col1=&var;
run;

proc univariate normal plot data=final;
 title 'Normality Assessment for';
 title2 'Power-Transformed Variable';
run;
%mend bctrans;
```

How the Macro Works

The macro takes the following arguments:

data: the original data set where the variable to be transformed can be found;

out: temporary data set for storing the highest λ and $l(\lambda)$ values;

var: the name of the variable to be transformed;

r: the number of λ values requested during the PROC IML iteration;

min: the minimum λ value;

max: the maximum λ value; and

step: the increment in the λ values from min. to max.

After data are read from data set `&data`, they are input in a vector **x**. We also create a vector **one** with equal number of rows to **x** and values of 1, as well as a $r \times 2$ matrix **MAT**. A looping process selects values of λ, checks if the value is zero or not and applies formula 1. The mean of these values is then estimated using formula 3. The value of the function $l(\lambda)$ is then estimated based on formula 2. Results are put in columns 1 and 2 of **MAT** in the last two lines of the loop, and a data set `lambdas` is created. The PLOT procedure provides a visual display of the λ and the $l(\lambda)$ values. These values are then sorted and the highest λ is output to data set `&out`. A second run of PROC IML transforms the variable `&var` and PROC UNIVARIATE is called to test the normality of its distribution.

An Example

We demonstrate the use of the `bctrans` with the following example. Variable X represents school age children's responses to a visual acuity test. Forty eight observations were included in this analysis.

Exploratory data analysis showed that variable X is not normally distributed, but is heavily skewed. The PROC UNIVARIATE with options NORMAL and PLOT confirms this lack of normality. There is a great variation in the distribution, and the Shapiro-Wilk test is highly significant ($W = .685599, p < .00001$). Selected output from PROC UNIVARIATE is as follows:

```
Variable=x
                 Moments

 N                 48  Sum Wgts         48
 Mean        93.42063  Sum         4484.19
 Std Dev     90.63809  Variance   8215.263
 Skewness    2.152888  Kurtosis    3.94885
 USS         805033.2  CSS        386117.4
 CV           97.0215  Std Mean   13.08248
 T:Mean=0    7.140895  Pr>|T|       0.0001
 Num ^= 0          48  Num > 0          48
 M(Sign)           24  Pr>=|M|      0.0001
 Sgn Rank         588  Pr>=|S|      0.0001
 W:Normal    0.685599  Pr<W         0.0001

   Stem Leaf                     #  Boxplot
      3 79                       2     *
      3 3                        1     *
      2 89                       2     0
      2 1                        1     0
      1
      1 00001223                 8  +-----+
      0 555555677777779999      18  *--+--*
      0 2223333344444444        16  +-----+
        ----+----+----+----+
    Multiply Stem.Leaf by 10**+2
```

We request thirty one λ values from a minimum -2 to a maximum +1 by 0.1 and we invoke the macro `bctrans` as follows:

```
%bctrans(data=sugi22.orig,out=boxcox,var=x,
   r=31,min=-2,max=+1,step=.1)
```

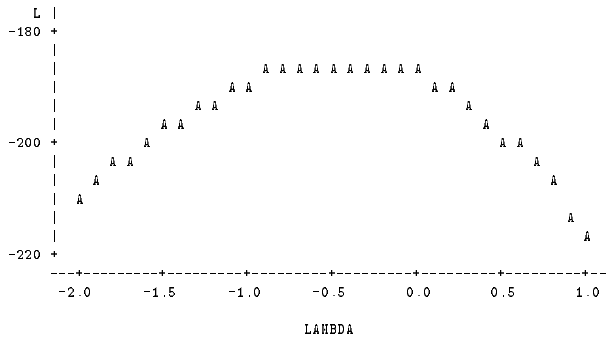
The macro `bctrans` gives the following results. First, it prints the list of λ and $l(\lambda)$ values:

```
Lambdas and their l(lambda) values

     MAT
  -2.000 -209.257
  -1.900 -206.692
  -1.800 -204.227
  -1.700 -201.867
  -1.600 -199.622
  -1.500 -197.498
  -1.400 -195.504
  -1.300 -193.651
  -1.200 -191.949
  -1.100 -190.408
  -1.000 -189.042
  -0.900 -187.863
  -0.800 -186.885
  -0.700 -186.123
  -0.600 -185.592
  -0.500 -185.306
  -0.400 -185.281
  -0.300 -185.529
  -0.200 -186.064
  -0.100 -186.896
   0.000 -188.032
   0.100 -189.480
   0.200 -191.239
   0.300 -193.310
   0.400 -195.687
   0.500 -198.364
   0.600 -201.330
   0.700 -204.573
   0.800 -208.080
   0.900 -211.836
   1.000 -215.825
```

Then, the macro plots the λs against their $l(\lambda)$s:

Lambda vs. l(lambda) values



The macro `bctrans` then selects the highest λ value:

```
Highest lambda and l(lambda) value

       LAMBDA        L

        -0.4     -185.281
```

It then transforms the original variable X based on the selected λ, assesses its normality with PROC UNIVARIATE NORMAL PLOT, and provides box-and-whiskers and stem-and-leaf plots of its distribution.

```
Variable=new

                 Moments

 N                 48  Sum Wgts         48
 Mean        0.192023  Sum        9.217109
 Std Dev     0.051734  Variance   0.002676
```

```
Skewness   -0.14195  Kurtosis   -0.62541
USS        1.895691  CSS        0.125793
CV         26.94175  Std Mean   0.007467
T:Mean=0   25.71549  Pr>|T|       0.0001
Num ^= 0         48  Num > 0          48
M(Sign)          24  Pr>=|M|      0.0001
Sgn Rank        588  Pr>=|S|      0.0001
W:Normal   0.962036  Pr<W         0.2037
```

And the stem-and-leaf and box-and-whiskers plots of the distribution:

```
  Stem Leaf                       #  Boxplot
    28 5                          1     |
    26 5179                       4     |
    24 059                        3     |
    22 22675678                   8  +-----+
    20 444656                     6  |     |
    18 13346785                   8  *--+--*
    16 12568                      5  |     |
    14 4772678                    7  +-----+
    12                                  |
    10 448                        3     |
     8 249                        3     |
       ----+----+----+----+
   Multiply Stem.Leaf by 10**-2
```

```
                   Normal Probability Plot
0.29+                                           +++ *
    |                                      **+*+*
    |                                   ****+
    |                              *****
    |                          ****+
0.19+                      *****
    |                   ****
    |              ******
    |            +++
    |        +++***
0.09+   *+++* *
     +----+----+----+----+----+----+----+----+----+----+
         -2        -1          0        +1        +2
```

Final Comments

The presented macro bctrans provides an easy way for SAS users to transform their data to near normality prior to invoking any parametric statistic. The macro allows users to select the range of λ values to be considered as well as the distance between two consecutive λs. For more information on power transformations, SAS users should consider the paper by Box and Cox (1964) as well as the work of C. Hallahan (1990) and his BOXCOX macro, which he presented at SUGI15.

Users interested in a discussion of transformations in general are urged to also consider the relevant chapter by Emerson and Stoto (1983). In their work they consider transformations as a tool for ‘enhancing the interpretation’ (p. 126) of data and focus on *matching transformations*, which allow transformed data to look similar to the original data.

REFERENCES

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*, 211–252.

Emerson, J. D., & Stoto, M. A. (1983). Transforming data. In D. C. Hoaglin, F. Mosteller, & J. W. Turkey (Eds.), *Understanding robust and exploratory data analysis*, pp. 97–128. New York: Wiley.

Hallahan, C. (1990). A sas macro for the box-cox transformation: Estimation and testing. *SUGI Proceedings, 15*, 1339–1344.

Johnson, R. A., & Wichern, D. W. (1993). *Applied Multivariate Statistical Analysis* (3rd ed). Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R. E. (1982). *Experimental Design: Procedures for the behavioral sciences* (2nd ed). Pacific Grove, CA: Brooks/Cole.

SAS Institute, Inc. (1989). *SAS/IML® Software: Usage and Reference, Version 6, First Edition*. Cary, NC: SAS Institute, Inc.

I would like to thank Dr. Silas Halperin of the Psychology Department for the statistical courses he has taught over the years and the many discussions we had on statistical and SAS related issues. This macro is based on earlier macros he presented in his ”Multivariate Analysis” class.

Author Information:

Ioannis Dimakos, Computing & Media Services, Syracuse University, 120 Hinds Hall, Syracuse, NY 13244-2390. Email address: idimakos@syr.edu