A Sas Syntax For Bootstrap Multivariate Normality Assessment: A Technical Approach

Wan Muhamad Amir W Ahmad, Rabiatul Adawiyah Abdul Rohim, Farah Muna Ghazali, Nor Azlida Aleng, Mohamad Arif Awang Nawi

Abstract: SAS stands for the Statistical Analysis System, a software system for data analysis and report writing. SAS is a group of computer programs that work together to store data values and retrieve them, modify data, compute simple and complex statistical analyses and create reports. This research paper gives special attention to the multivariate normality assessment using SAS syntax procedure through graphical assessment of multivariate normality. This special syntax is created by considering numbers of dependent variable at a time. This provided a clear view about normality assessment for the studied data and allows a parametric analysis for the further data analysis. Through the SAS syntax, an applied researcher can easily use the syntax which provided in this paper by changing variables of interest and run the analysis. The graphical plot will be available after running the syntax. Therefore researcher is able to assess normality of multiple dependent data. The assessment of the normality is based on graphical plot which based on Chi-Square versus Mahalanobis distance plot. This step provided a very basic platform before we are proceed multivariate analysis. As a conclusion, it useful to researchers to check the normality assumption of multiple dependent variables.

Index Terms: multivariate normality, SAS syntax, mahalobis, Chi-Square, parametric analysis, multiple dependent data, assumption

1. INTRODUCTION

A normality test is used to determine whether sample data has been drawn from a normally distributed population or not. A number of statistical tests especially parametric analysis require a normally distributed sample population [1]. Testing for normality is a common procedure in much-applied work and many tests have been proposed to test the normality of certain data distribution [1]. The first formal test of multinormality has been proposed by [2] through multivariate measures of skewness and kurtosis. The chapter discusses the strict multivariate procedures, radius and angles and graphical techniques, and nearest distance test [2-4]. A multivariate analysis refers to analysis in which there are multiple dependent variables [6]. Among commonly used multivariate analyses are exploratory factor analysis and multivariate analysis of variance (MANOVA). The analyses are dependent on the assumption of multivariate normality of relevant variables [6-8]. The importance of normal distribution is undeniable since it is an underlying assumption of many statistical procedures. It is also the most frequently used distribution in statistical theory and applications. Therefore, when carrying out statistical analysis using parametric methods, validating the assumption of normality is of fundamental concern for the analyst. An analyst often concludes that the distribution of the data 'is normal' or 'not normal' based on the graphical exploration (Q–Q plot, histogram or box plot) and formal test of normality [5]. Statistical assessment of multivariate normality is available in a number of statistical packages, for examples in SPSS Amos (Mardia's multivariate kurtosis) and R (Mardia's, Royston's and Henze-Zirkler's multivariate normality tests via MVN package) [8]. Addition, The aim of this research paper is to present the steps to construct chi-square versus Mahalanobis distance plot in SAS using SAS syntax. At first, researchers need to provide data on multiple dependent variables.

A. SAS Syntax Procedure

In this secion, we are going to explain the procedure on how to run the SAS syntax with step-by step procedure. Therefore, through this syntax procedure, reader are able to substitute their data, changing the parameter and straightly run the syntax successfully. Below is the detailed of the guideline to researcher for running the multivariate normality using SAS syntax with their own research data.

A SAS Syntax

Step 1: At /*MACRO BOOTSTRAP*/

```
%MACRO bootstrap(data=Score, booted=booted, boots=2,
seed=1234);
DATA &booted;
```

- Insert datafile name. Let say the data file name is Score. Then write data= Score.
- Let say the bootstrap procedure is running at two times. Therefore, at boots insert 2, so we can write boots=2 . See in Macro Bootstrap.

Step 2: At /*MACRO FOR CALCULATE MULTIVARIATE NORMALITY*/

```
%macroNormultivariate
(data= booted, var= S_Biology S_Chemistry S_Math,
plot=both);
```

- Insert the variables name. Let say the name variable in this study is S_Biology S_Chemistry and S_Math. Then, we can write var = S_Biology S_Chemistry S_Math. See the above macro statement.

Step 3 : At /*INPUT DATA*/

```
Data Score;
Input S_Biology S_Chemistry S_Math;
Datalines;
```

- Insert data name, let say our data name is Score. Therefore Data Score. The name variables in this study are S_Biology S_Chemistry and S_Math. Then, we can write Input S_Biology S_Chemistry S_Math.

Step 4: At /**GENERATE BOOTSTRAP SAMPLE**/

```
%bootstrap(data= Score, boots=2);
run;
```

- Insert the number of bootstrap according to user need. For example in this case the boostrap is set at 2 (boots=2).

Step 5: At %Normultivariate (data=booted,var=S_Biology S_Chemistry S_Math, plot=mult)

- Insert studied variable at var. For example in this case the variables are S_Biology S_Chemistry and S_Math. So write var=S_Biology S_Chemistry S_Math.See the above statement.

Step 6: Run the syntax.

2 METHODS

The secondary data was used in this study. Data of this study consist of four variables which namely as in Table I. The detail of data description as shown in Table I.

TABLE I
DATA DESCRIPTION

Num.	Variables	Explanation of user variables
1.	S_Biology	A score of biology test
2.	S_Chemistry	A score of chemistry test
3.	S_ Math	A score of mathematics test

Section A give the build procedure of SAS syntax, which can be used to determine the normality of multiple dependent variables.

B. SAS Syntax

C.

- This section creating a macro bootstrap for data generating. This syntax provided a new set of data by case resampling procedure. Researcher can determine the number of procedure need before launching a procedure.

```
/* MACRO BOOTSTRAP */

%MACRO bootstrap(data=Score, booted=booted, boots=2, seed=1234);
DATA &booted;
** randomly picks an integer from 1 to n;
pickobs = INT(RANUNI(&seed)*n)+1;
** POINT tells SAS to read value pickobs
** NOBS sets n to number of obs in &Data;
** when the point option is used SAS will loop through the data step forever;
SET &data POINT = pickobs NOBS = n;
** saves number of current bootstrap;
REPLICATE=int(i/n)+1;
i+1;
** stop will leave data set when n*&boots obs have been created;
IF i > n*&boots THEN STOP;
RUN;
%MEND bootstrap;
```

- This section calculate a multivariate normality, remove observations with missing values and determine variables, covariance matrix, compute and create values for chi square plot, input data, generate bootstrap sample and print data.

```
/*MACRO FOR CALCULATE MULTIVARIATE NORMALITY */

%macro Normultivariate
(data= booted, var= S_Biology S_Chemistry S_Math , plot=both);
 /* Data Set */
 /* List of Variables */
 /* Create normal multivariat plot */

/* REMOVE OBSERVATIONS WITH MISSING VALUES AND DETERMINE VARIABLES */

Data no_missing;
Set&data;
if nmiss(of &var)=0;
 run;

%let i=1; %let k=0;
%let dsid=%sysfunc(open(&data));
%if &dsid %then %do;
%let token=%scan(&var,&i);
 %do %while (&token ne %str() );
 %if %sysfunc(varnum(&dsid,&token)) ne 0 %then %do;
 %let k=%eval(&k+1);
 %let token=v&k; %end;
 %let i=%eval (&i+1);
 %let token=%scan(&var,&i);%end;
 %let rc=%sysfunc(close(&dsid)); %end;
 %let nvar=&k;


/*COVARIANCE MATRIX*/

Proc princomp Data=no_missing out=print(keep=prin:)
std vardef=n;
run;

/*COMPUTE AND CREATE VALUES FOR CHI-SQUARE PLOT*/

Data Chisquare_Plot;
set print;
Mahalanobis_Dist=uss(of prin1-prin&&k);
keep  Mahalanobis_Dist;
run;

Proc rank Data=Chisquare_Plot out=Chisquare_Plot;
var Mahalanobis_Dist;
ranks rDist;
run;

Data Chisquare_Plot;
set Chisquare_Plot nobs=n;
Chisq=cinv((rDist-0.5)/n,&k);
keep  Mahalanobis_Dist Chisq;
run;
```

```
Proc sgplot data=Chisquare_Plot;
scatter x= Mahalanobis_Dist y=Chisq;
quit;
run;

%exit:
%mend;

/*INPUT DATA */

/*DATA DESCRIPTION
  'S_Biology'='Biology Score';
  'S_Chemistry'='Chemistry Score';
  'S_Math'='Math Score'*/

Data Score;
Input S_Biology S_Chemistry S_Math ;
Datalines;

75.97     71.70     58.48
63.29     66.58     60.63
48.91     61.48     62.97
69.09     75.86     63.39
71.63     66.14     65.51
76.49     68.96     65.67
79.63     73.33     66.04
73.34     68.10     66.21
65.99     67.91     66.64
  ⋮         ⋮         ⋮
84.14     78.26     97.93
80.02     90.79     98.33
90.13     94.45     98.58
72.39     80.99     99.46
;

/**GENERATE BOOTSTRAP SAMPLE**/

%bootstrap(data= Score, boots=2);
run;
;

/**PRINT DATA **/

proc print data=booted;
run;
%Normultivariate (data=booted,
var=S_Biology S_Chemistry S_Math, plot=mult);
run;
```

3 RESULTS

The resulting plot is shown in Figure 1. From the plot below, it is clearly shown that the variables form a clear straight line and noticeably there is no outlier at upper right part of the plot. We may conclude that the variables form a multivariate normal distribution.

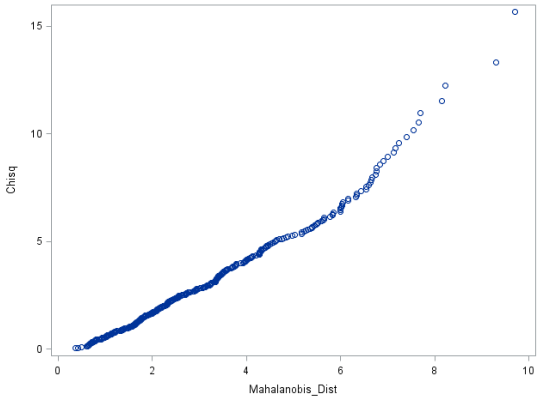

Fig I *Chi-Square versus Mahalanobis distance plot*

4 CONCLUSION

In this research paper, we have shown the steps to obtain a chi square versus Mahalobis distance plot for graphical assessment of multivariate normality assumption in SAS. Then, we have provided a brief and comprehensive SAS syntax for the researcher to check the normality of the multiple dependent variable. Through the SAS syntax, researcher are allow to assess of the normality of a multiple dependent variable by changing the data and variables through the build syntax provided in this article. This step provided a very basic platform before we are proceed multivariate analysis. As a conclusion, it allow researchers to check the normality assumption of multiple dependent variables.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Universiti Sains Malaysia (USM) for providing the research funding (Grant no.1001/PPSG/8012278. School of Dental Sciences)

REFERENCES

[1] Arifin, W N., 2015. “The Graphical Assessment of Multivariate Normality Using SPSS”, Education in Medicine Journal, Vol. 7 (2), pp. 71-705.

[2] Mardia, K V., 1980. ‘Tests of univariate and multivariate normality’, in Krishnaiah P. R. (ed.),Hand-book of Statistics, Volume 1, Chapter 9. North-Holland, Amsterdam, pp. 279–320.

[3] Agostino, R B D., 1982. “Departures from normality, testing for”, in S., N. L., Kotz Johnson and C. B. Read (eds), Encyclopedia of Statistical Sciences, Vol. 2 North-Holland, Amsterdam, pp. 315–324.

[4] Small, N J H., 1985. “Multivariate normality testing for”, in Kotz S., Johnson N. L. and Read C. B.(eds),Encyclopedia of Statistical Sciences, Vol. 6, North-Holland, Amsterdam, pp. 95–100.

[5] Yap, B W., and Sim, C.H., 2011.“Comparisons of various types of normality tests”, Journal of Statistical Computation and Simulation, Vol. 81(12), 2141-2155.

[6] Tabachnick, B G., and L.S. Fidell, 2007 “ Using Multivariate Statistics”, 5th ed. Boston: Pearson Education.

[7] Hair, J.F., Black, W.C., Babin, B J., and Anderson R E., 2009 “ Multivariate Data Analysis, 7th ed. Upper Saddle River, NJ: Pearson Prenticehall.

[8] Burdenski, T. 2000. “ Evaluating univariate, bivariate and multivariate normality using graphical and statistical

procedures. Multiple Linear Regression Viewpoints 26(2):15-28.
[9] Hair J.F. Jr, Black, W.C., Babin, B J., and Anderson R E, 2009.“ Multivariate Data Analysis, 7th ed. Upper Saddle River, NJ: Pearson Prenticehall
[10] Burdenski, T.,2000 “ Evaluating univariate, bivariate and multivariate normality using graphical and statistical procedures. Multiple Linear Regression Viewpoints 26(2):15-28.