Multivariate data analysis and data reduction techniques

V. Geethalakshmi, Principal Scientist
ICAR-CIFT, Cochin
geethasankar@gmail.com

SAS stands for Statistical Analysis System. It is a software suite developed by SAS Institute. The applications of SAS are data management, advanced analytics, predictive analytics, multivariate analysis and business intelligence and reporting with perfect graphics. More than 200 components are available in SAS. The important SAS components are:

- Base SAS: It is the most widely used component. It has data management facility. You can do data analysis using Base SAS.
- SAS/GRAPH: With the use SAS/Graph you can represent data as graphs. This makes data visualization easy.
- SAS/STAT: It lets you perform Statistical analysis, such as Variance, Regression, Multivariate, Survival and Psychometric analysis.
- SAS/ETS: It is suited for Time Series Analysis.

DATA step and PROC step form the basic building blocks of a SAS program. We start a program with a DATA step to create a SAS data set and then pass the data onto a PROC step. The PROC step processes the data.

Creating dataset

The dataset can be created and variables named by typing directly in the code. Another method is to import files from the computer into SAS workspace. To create a dataset say 'Game' use the following code :

```
DATA Game;              #Name the data set.
INPUT x,y,z;            #Define the variables in this data set
DATALINES;              #In the following lines type the data
58 65 70
23 45 89
11 25 32
;run;
```

The data file can also be read from Excel using 'import' command.

Importing Excel file

```
proc import out=work.myfile datafile="<pathname>" dbms=xlsx replace;
run;
```

Procedures in SAS

The PROC means command is used to extract the descriptive statistics from the dataset. Usage of PROC means and the various options are given below.

```
1)Descriptive statistics
proc means;
variables tpc tc fs;
run;
2)Descriptive statistics specified output
proc means data=work.myfile n mean max min range std fw=8;
run;
3)Descriptive statistics crosstabulation
proc means data=work.geetha maxdec=3;
variables TPC TC FS;
class Lake sp;
types () Lake*sp;
title Average microbial load;
run;
4) The By statement
proc means data=work.geetha;
by season;
variables TPC TC FS;
class Lake sp;
run;
5)Confidence limits for mean
proc means data=work.geetha fw=8 alpha=0.1 clm mean std;
variables TPC TC FS;
run;
```

Correlation and regression

PROC corr and PROC reg are used to compute correlations from the x,y data. The various options are described below :

Finding correlation

```
proc corr data=work.myfile;
variables salinity ph rainfall;
run;
```

Finding regression

```
proc reg data=work.geetha;
model TPC = salinity pH Temp;
run;
```

The selection option of PROC reg will specify the method of variable to be included in the model. The following example gives the data on fitness and code to fitting regression model using various selection methods.