

ThunderCell Plot Interpretation of Correlation Matrices Using the SAS® System

Hurley, G.J., Procter and Gamble Pharmaceuticals, Inc., Mason, OH

ABSTRACT

Correlation matrices are quite useful in exploring relationships among groups of variables. Although generation of a correlation matrix is quite simple using SAS® software, the interpretation in the case of large matrices can be challenging. In *The American Statistician*, Vol. 56 No. 4, p. 316, November 2002, Michael Friendly proposes the use of corrgrams, various graphical displays that allow these correlations to be visually interpreted. In practice, these are of particular use when the number of exploratory variables is large. This idea of the corrgram has been modified into the ThunderCell plot, which is based on the same principles. In *Products Research Testing*, these plots can conveniently illustrate correlations between variables of interest in a fashion that allows relationships to be quickly inferred. Furthermore, it is of more interest and easier to explain to non-statisticians than providing an unordered or alphabetically ordered matrix of correlation values. To automate this process, a SAS macro has been written to generate these plots. The macro is quite simple in structure and leverages both SAS and Visual Basic in one SAS macro call to create such plots. An overview of this macro's structure will be presented, followed by a real data example.

INTRODUCTION

It is quite common to generate a correlation matrix to determine the relationship between a number of variables. In the SAS System, this is most commonly done using `proc corr`. It is often the case that the result is presented blandly, often as a table of numbers. Friendly (Friendly, 2002) states, "there are few techniques for exploratory visual display, and for depicting the patterns of relations among variables in such matrices directly." He proposes the use of corrgrams, graphical displays to optimize interpretation of the correlation between several variables. Based on the concept of the corrgram, the ThunderCell plot has been created to generate an alternative visual method of presenting these data. Named for its similarity to weather maps on the evening news, this plot is easily interpretable by even those who are not statistically inclined and is quite useful in presenting an overview of the relationship among a number of variables.

The concept of the ThunderCell plot is simply to arrange the correlation matrix into a table such that those variables that are most highly correlated are placed next in such a fashion that their rows and columns intersect in a region to form a ThunderCell (or multiple ThunderCells). When color is then applied based on the level of correlation, these regions are easy to identify and those interpreting the plot can focus on these regions.

As a fictional example, consider five variables: height, weight, age, caloric intake, and number of pets. It would be reasonable to assume height and weight would be moderately positively correlated with each other. In addition, it is likely that caloric intake would be highly correlated with weight, moderately correlated with height, and slightly correlated with age. Furthermore, age may be slightly positively correlated to weight and not correlated to height (assume we are discussing 18-40 years of age). Finally, assume that number of pets is slightly correlated with age and that it is not correlated with any of the other three variables. Figure 1 illustrates the concept of a ThunderCell plot for these variables in gray scales.

Figure 1 The ThunderCell Concept.

	Number of Pets	Age	Height	Weight	Caloric Intake
Number of Pets					
Age					
Height					
Weight					
Caloric Intake					

What is to be noted here is that the ThunderCell occurring in the lower right hand corner illustrates a high amount of correlation between these variables. To the investigator, this would indicate a strong relationship between Caloric Intake, Weight, and somewhat so with height.

It is noted here that this example is overly simplistic; there is one ThunderCell with high correlations and decreasing correlations as the distance from the ThunderCell increases. However, it may be possible to imagine two other exploratory variable, perhaps amount of pet food purchased and total weight of all pets. These variables should be highly correlated with number of pets, while not being correlated to the other variables. Hence, a gray-scale ThunderCell plot that looks like Figure 2 is reasonable. What is seen here is that there are two distinct ThunderCells. One ThunderCell describes the correlation between the pet related questions, while the other describes the correlation among the person's height, weight, and age.

Likewise, one may consider groups of variables with three or more ThunderCells.

Although the concept of the ThunderCell plot is really quite simple, an algorithm is necessary to determine how the rows and columns should be arranged if the process of graph generation is to be automated using SAS macro language.

ROW AND COLUMN ARRANGEMENT:

Algorithms exist and are described by Friendly (2002) for arranging rows and columns for one and two ThunderCell Plots. These methods have been adopted for this macro. Furthermore, a method is suggested by Falissard (1996), which may be utilized to create plots with three ThunderCells.

Figure 2 The Two ThunderCell Concept.

	Total Weight of All Pets	Amount of Pet Food Purchased	Number of Pets	Age	Height	Weight	Caloric Intake
Total Weight of All Pets							
Amount of Pet Food Purchased							
Number of Pets							
Age							
Height							
Weight							
Caloric Intake							

ONE THUNDER CELL

Friendly suggests the use of Eigenvalues of the correlation matrix (Principal Components) to determine the order of these values. This methodology works as follows. For a plot requiring one ThunderCell, the components of the first Eigenvector are ranked and the variables are displayed in this order. It is noted that the variables contributing to the ThunderCell having the largest components on the correlation matrix's first Eigenvector will represent the most influential contributions to the first Principal Component, which accounts for the greatest amount of the data's variability as compared to the other Principal Components.

TWO THUNDER CELLS

For a two-cell solution, a vector of angles between the first and second Eigenvectors is ranked. The elements of this vector are determined as follows:

$$\alpha_i = \begin{cases} \tan^{-1}(e_{i2}/e_{i1}), & e_{i1} > 0 \\ \tan^{-1}(e_{i2}/e_{i1}) + \pi, & e_{i1} \leq 0 \end{cases}, \text{ where } e_{ij} \text{ is the } i^{\text{th}} \text{ element of the } j^{\text{th}} \text{ largest of the first two eigenvectors (j=1, 2), and } \alpha_i \text{ is the } i^{\text{th}} \text{ element of the vector of angles.}$$

According to Friendly, "the angles between vectors approximate the correlations between variables."

THREE THUNDER CELLS

Falissard suggests that there is a three-cell solution as well. This can be obtained through a spherical representation of a correlation matrix. However, this method is somewhat complex mathematically, and requires the use of an iterative procedure known as the Newton Raphson Algorithm. Furthermore, a situation where this is necessary has not been seen by this

author. Hence, at this point in time, this feature has not been added to the macro due to the complexity.

MACRO OVERVIEW:

A SAS macro has been written that generates the correlation matrix using Proc Corr and arranges the rows and columns according to the algorithms described above. The program then uses the Output Delivery System to create a Rich Text File. This Rich Text File is modified by a Visual Basic Program, which is executed in SAS. The following is a sample macro call illustrating the five parameters necessary to execute the macro:

```
%thunder (dsn=ddl.MWSUG14, varlist= var1 var2
var3 var4 var5 var6 var7 var8 var9 var10
var11 var12,
output=%str(H:\Files\presentation\MWSUG
2003\MWSUG14.rtf), numcells=1,
corrtype=pearson);
```

The five main parameters are **dsn**, the data set containing the variables to be analyzed, **varlist**, the list of the variables to be analyzed, **output**, the final output location, **numcells**, the ThunderCells arrangement (1 or 2 cells), and **corrtype**, the type of correlation matrix requested (pearson, spearman).

The macro functions by first using Proc Corr to obtain the Pearson or Spearman Correlation Matrix as specified by the user. This is output to a dataset using an ODS statement. Following this, Proc IML is used to read in the dataset, to obtain the Eigenvalues of the matrix, and to reorder the matrix rows, columns, and variable labels appropriately. ODS is then used to write this matrix out to an RTF table located in the output destination. The location of the output is then written to a text file, and the dm statement is used to execute a visual basic program which reads the location of the RTF file from the aforementioned text file and opens the RTF file to perform shading, font, and title

modifications, as well as creating a legend along the bottom of the page that the plot occurs on. This file is then written to the same location in which the RTF file was originally written. In this manner, the user need never execute anything beyond SAS.

A REAL EXAMPLE

The National Health Interview Survey (NHIS) is conducted annually by the National Center of Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC). The NCHS, CDC make these data publicly available on their website, <http://www.cdc.gov/nchs/nhis.htm>. For the purpose of illustrating the ThunderCell plots, the 2000 family dataset, familyxx was used. From this data set, for simplicity, ten variables were selected to analyze. The ten variables selected here were fhospct (the number of family members in the hospital overnight in the past twelve months), fm_educ (education of adult with highest education in the family), fncgrp (total combined family income (grouped)), fm_eldr (the number of family members aged 65 or older), fhicarct (the number of family members with Medicare Health Insurance), fssrct (the number of family members with Social Security or Railroad Retirement), fliadlct (the number of family members needing help with chores or shopping), fhicovct (the number of family members with health insurance coverage), fm_size (size of family), and fm_kids (the number of family members aged seventeen or younger).

The following code was used to generate Figure 3, a one ThunderCell presentation of this correlation data.

```
%thunder (dsn=NHIS.FAMILYXX, varlist= fm_educ
fm_eldr fm_kids fhicarct fhicovct fhospct
fncgrp fliadlct fm_size fssrct,
output=%str(H:\Files\presentation\MWSUG
2003\data\familypl.rtf), numcells=1,
corrtype=pearson);
```

The numbers in Figure 3 are indicative of a high correlation between the number in the family receiving social security or railroad retirement, the number in the family over 65, and the number with Medicare health insurance. These correlations aren't particularly surprising given that these three attributes are all age dependent. Not as highly correlated, but on the edge of the ThunderCell is the number in the family needing help with chores or shopping. This is an interesting attribute to have join the ThunderCell. One possible explanation for this is that it may be more common for older people to need help with chores or shopping than younger people. Other strong correlations are seen between family size, number of family members covered by health insurance, and the number of family members aged seventeen or younger. This appears to be another cell in formation, perhaps involving total combined family income (grouped). Hence, it seems that a two ThunderCell solution may be more appropriate to explain these data. In general, if variables not contained in the first ThunderCell are demonstrating high correlations, it is ideal to explore a two ThunderCell solution.

A two ThunderCell presentation was considered with the following code:

```
%thunder (dsn=NHIS.FAMILYXX, varlist= fm_educ
fm_eldr fm_kids fhicarct fhicovct fhospct
fncgrp fliadlct fm_size fssrct,
output=%str(H:\Files\presentation\MWSUG
2003\data\family2.rtf), numcells=2,
corrtype=pearson);
```

This is illustrated in Figure 4.

The first ThunderCell seems to remain the same, although it relocates to another section of the chart. The second ThunderCell forms toward the lower right. It can be seen that the aforementioned attributes family size, number of family members covered by health insurance, and the number of family members aged seventeen or younger, and total combined family income (grouped) form this second ThunderCell. Additionally, although the correlation is not particularly strong, the education of the highest adult in the family also seems to be related to these attributes.

The alternative to such a visual presentation may have been an ordering such as what follows in Figure 5. Although the results can be obtained from such a table, it is clear through examination that this task is much more difficult.

CONCLUSION:

The ThunderCell plot is an efficient way to look at correlation data. Strongly influenced by Friendly's paper, this method allows the analyst to quickly assess relations between groups of variables in large, multivariate datasets. The creation of a macro to analyze such data has facilitated the production of these plots and decreased the time for their production. The particular method used to create the plots, specifically a hybrid of SAS language and Visual Basic version 6, allowed for a great combination of power from both an analysis and a formatting level.

REFERENCES

Falissard B. A Spherical Representation of a Correlation Matrix. *Journal of Classification* 1996; 13:267-280.

Friendly M. Corrgrams: Exploratory Displays for Correlation Matrices. *The American Statistician*. 2002; 56:316-324.

ACKNOWLEDGMENTS

This author would like to acknowledge David Hengehold and Rob Baker for their help in proposing and refining the idea of the "ThunderCell" plot.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

George J. Hurley
Procter and Gamble Pharmaceuticals, Inc.
Health Care Research Center, Box 594
8700 Mason Montgomery Road
Mason, OH 45040
Phone: (513) 622-3020
Fax: (513) 622-0987
Email: hurley.gj@pg.com

FIGURE 3 NHIS DATA: 1-THUNDERCELL PRESENTATION.

THUNDERCELL PLOT										
PEARSON CORRELATIONS	FM_KIDS	FM_SIZE	FHICOVCT	FINCGRP	FM_EDUC	FHOSPCT	FLIADLCT	FSSRRCT	FHICARCT	FM_ELDR
# AGED 17 AND YOUNGER			0.56	0.10	-0.04	0.15	-0.16	-0.27	-0.11	-0.25
SIZE OF FAMILY				0.27	-0.03	0.18	-0.06	-0.16	-0.17	-0.12
# WITH HLTH INS COVERAGE	0.66			0.38	-0.06	0.17	-0.03	-0.06	-0.06	-0.08
TOTAL COMBINED FAMILY INCOME (GROUPED)	0.10	0.27	0.38		0.21	-0.03	-0.16	-0.13	-0.19	-0.18
EDUC OF ADULT WITH HIGHEST EDUC IN FAM	-0.04	-0.03	-0.06	0.21		-0.04	-0.06	-0.05	-0.02	0.00
# IN HOSP OVERNIGHT, 12M	0.15	0.18	0.17	-0.03	-0.01		0.14	0.09	0.09	0.08
# NEED HELP W/CHORES, SHOP	-0.10	-0.06	-0.03	-0.15	-0.06	0.14		0.25	0.27	0.24
# REC SOC SEC OR RR RET	-0.22	-0.16	-0.04	-0.10	-0.04	0.09	0.26			
# W/ MEDICARE HLTH INS	-0.24	-0.17	-0.06	-0.19	-0.02	0.09	0.27			
# AGED 65 AND OLDER	-0.25	-0.17	-0.06	-0.18	0.00	0.08	0.24			

Less Correlation More



FIGURE 4 NHIS DATA: 2-THUNDERCELL PRESENTATION.

THUNDERCELL PLOT										
PEARSON CORRELATIONS	FHOSPCT	FM_ELDR	FHICARCT	FSSRRCT	FLIADLCT	FHICOVCT	FM_SIZE	FM_KIDS	FINCGRP	FM_EDUC
# IN HOSP OVERNIGHT, 12M		0.08	0.09	0.09	0.14	0.17	0.18	0.15	0.03	-0.04
# AGED 65 AND OLDER	0.08				0.24	0.08	0.17	0.15	0.10	0.00
# W/ MEDICARE HLTH INS	0.09				0.27	0.06	0.17	0.24	0.18	-0.02
# REC SOC SEC OR RR RET	0.09				0.25	0.06	0.15	0.22	0.19	-0.05
# NEED HELP W/CHORES, SHOP	0.14	0.24	0.27	0.25		0.03	0.05	0.10	0.15	-0.00
# WITH HLTH INS COVERAGE	0.17	0.08	0.09	0.09	0.14			0.66	0.38	0.06
SIZE OF FAMILY	0.18	0.17	0.17	0.15	0.05				0.27	-0.08
# AGED 17 AND YOUNGER	0.15	0.05	0.14	0.23	0.10	0.66			0.10	0.04
TOTAL COMBINED FAMILY INCOME (GROUPED)	0.03	0.15	0.10	0.19	0.15	0.38	0.27	0.10		0.21
EDUC OF ADULT WITH HIGHEST EDUC IN FAM	0.04	0.00	0.02	0.05	0.00	0.00	0.03	0.04	0.21	

Less Correlation More



FIGURE 5 Unorganized Correlation Matrix.

Label	FM_EDUC	FM_ELDR	FM_KIDS	FBICARCT	FHICOVCT	FHOSPCT	FINCGRP	FLIADLCT	FM_SIZE	FSSRRCT
Educ of adult with highest educ in fam	1.00000	0.00011	-0.03564	-0.02227	-0.05970	-0.04105	0.21297	-0.00062	-0.03249	-0.04803
# aged 65 and older	0.00011	1.00000	-0.26163	0.87362	-0.07884	0.07719	-0.16488	0.23536	-0.17488	0.80482
# aged 17 and younger	-0.03564	-0.26163	1.00000	-0.24445	0.66105	0.15398	0.09596	-0.09605	0.85553	-0.21992
# w/ MEDICARE hlth ins	-0.02227	0.87362	-0.24445	1.00000	-0.05838	0.09103	-0.19306	0.27134	-0.16660	0.81116
# with hlth ins coverage	-0.05970	-0.07884	0.66105	-0.05838	1.00000	0.16767	0.38161	-0.03118	0.74046	-0.05904
# in hosp overnight, 12m	-0.04105	0.07719	0.15398	0.09103	0.16767	1.00000	-0.02620	0.13798	0.17834	0.09384
Total combined family income (grouped)	0.21297	-0.16488	0.09596	-0.19306	0.38161	-0.02620	1.00000	-0.14548	0.26666	-0.19446
# need help w/chores, shop	-0.00062	0.23536	-0.09605	0.27134	-0.03118	0.13798	-0.14548	1.00000	-0.06025	0.25219
Size of family	-0.03249	-0.17488	0.85553	-0.16660	0.74046	0.17834	0.26666	-0.06025	1.00000	-0.14611
# rec Soc Sec or RR ret	-0.04803	0.80482	-0.21992	0.81116	-0.05904	0.09384	-0.19446	0.25219	-0.14611	1.00000