# THE SAS® SYSTEM FOR STATISTICAL GRAPHICS - A PREVIEW

Michael Friendly, York University


## Abstract

This talk presents an overview of methods for statistical graphics covered in my forthcoming book, *The SAS® System for Statistical Graphics*. In particular, I illustrate the design and implementation of custom graphic displays for:

- symmetry transformation plots
- scatterplots, enhanced with marginal boxplots and concentration ellipses.
- display of additive fits and residuals for two-way tables.
- scatterplot matrix for multivariate data.


## Introduction

The past 15 years have seen dramatic progress in statistical graphics. A wide variety of new graphic methods have been developed to help detect patterns in data and diagnose potential problems or violations of assumptions (e.g., Chambers et al., 1983; Cleveland. 1985).

However, there is often a big gap between theory and practice—between the articulation of a graphical technique and its implementation in a widely accessible form. The SAS System and SAS/GRAPH software, for example, provide the basic tools for statistical analysis and graphical data display. However, many modern methods of statistical graphics (e.g., diagnostic plots, enhanced scatter plots, plots for multivariate data) are not provided directly in SAS procedures.

*The SAS System for Statistical Graphics* is a forthcoming book in the SAS Application Series designed to fill this gap. The primary goals of the book are to survey the kinds of graphic displays that are most useful for different questions and data, and to show how can these displays be done with the SAS System.

The book contains almost 250 exhibits, of which over 150 are done with SAS/GRAPH, SAS/IML, and SAS/QC. The SAS programs for almost all the figures are explained. A number of graphical methods are implemented as general SAS macro programs which can be used with any set of data.

## Overview of *The SAS System for Statistical Graphics*

The book is organized according to the type of data to be analyzed and the data analysis tasks and statistical methods associated with that type of data. The progression is from graphical displays for univariate data to bivariate data, multiple regression, experimental design, multivariate methods and categorical data.

Chapter 1 provides an introduction to statistical graphics. Topics discussed include "Why plot your data?", the various roles of graphics in data analysis, and strategies for what to plot and how to plot it. Chapters 2 and 3 discuss graphical methods for univariate data. Chapter 2 describes methods for portraying the distribution of a single variable, including histograms, stem and leaf displays, boxplots and dot charts. Chapter 3 focuses on methods for plotting theoretical distributions and for comparing an empirical data distribution to a theoretical one, including quantile-quantile plots, density estimation, and diagnostic plots for assessing symmetry of a distribution.

Exploratory methods and graphical techniques for bivariate data are described in Chapter 4. These include simple scatterplots, labelling observations, enhancing a scatterplot with marginal boxplots or confidence ellipses, and various smoothing techniques including locally weighted robust scatterplot smoothing (lowess). Chapter 5 describes techniques for plotting the data and fitted values for linear, polynomial, and multiple predictor regression models. A variety of special-purpose plots for diagnosing violations of assumptions, detecting unduly influential observations, and choosing variables to be included in the model is also presented.

Chapters 6 and 7 are concerned with exploratory and confirmatory graphical methods for comparing groups. Chapter 6 discusses quantile comparison plots and comparative boxplots as well as graphical methods for transforming data to equalize variance. Chapter 7 discusses graphical techniques for experimental design (ANOVA) data, diagnostic plots for transforming data to an additive structure, and plotting power curves for ANOVA designs.

Techniques for multivariate data are presented in Chapters 8 and 9. Exploratory methods for displaying three or more variables simultaneously are discussed in Chapter 8. These include glyph plots, scatterplot matrices, star plots, profiles, and the biplot. Chapter 9 surveys plotting techniques related to standard statistical methods for multivariate data. These methods include $\chi^2$ probability plots for multivariate normality and for detecting multivariate outliers, and graphical techniques related to principal components analysis, canonical correlation, and discriminant analysis (MANOVA).

Chapter 10 covers exploratory graphical methods for categorical data—contingency tables. These techniques are designed to help show *how* categorical variables are related. The association plot and mosaic display both represent the pattern of deviation from independence by rectangles, one for each cell. Correspondence analysis summarizes the pattern of association between the row and column variables by distances in a two-dimensional display.

One appendix presents the SAS source code for all the macro programs in the book, and another appendix contains all the data sets analyzed.

The remainder of this paper describes the design and implementation of some of these graphical displays in the SAS System. These methods are illustrated with data on the price, weight, gas mileage and other measures of size and performance on 74 makes of automobiles (Chambers et al., 1983).

### Symmetry Transformation Plots

A particularly useful family of transformations is what Tukey (1977) calls the ***ladder of powers***, where a variable $x$ is transformed to $t_p(x)$ by a power $p$ according to

$$t_p(x) = \begin{cases} x^p, & p > 0 \\ \log_{10} x, & p = 0 \\ -x^p, & p < 0 \end{cases} \qquad (1)$$

For transforming data to become more symmetric, the simplest plots are based on the idea that in a symmetric distribution, the ordered observations, $x_{(i)}$, at depth $i$ from the extremes should be equally distant from the median, $M$. In the ***Upper vs. Lower plot***, we plot

$$M - x_{(i)} \quad \text{vs.} \quad x_{(n+1-i)} - M \quad , \qquad (2)$$

for $i = 1$ to $[n/2]$. In a symmetric distribution, these points should plot as a straight line with slope = 1. For skewed distributions, the points will tend to rise above the line (positive skew) or fall below (negative skew). The upper vs. lower plot for the automobiles data shown in Figure 1 indicates substantial positive skewness.

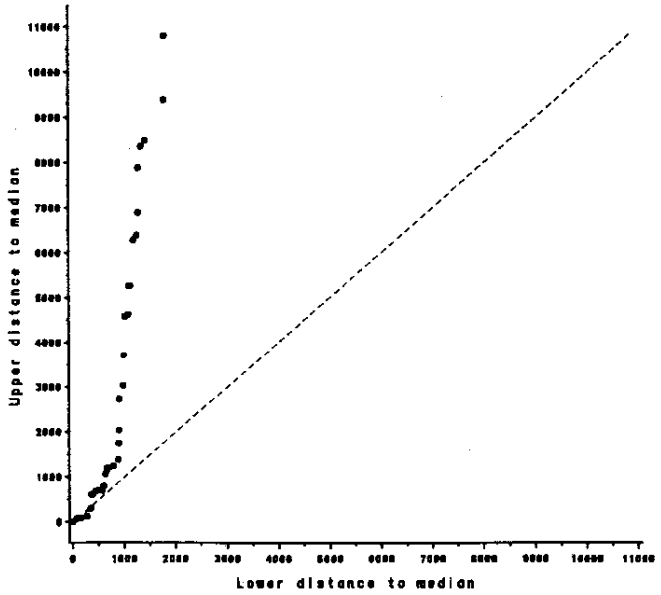


*Figure 1*: Upper vs. Lower plot for Price of automobiles

### Untilting: Mid vs. Spread Plots

A simple modification of this plot changes the coordinates so that the reference line for symmetry becomes horizontal. It is much easier to judge departure from a flat line than a tilted one. In this display, called the ***mid vs. spread plot***, we plot

$$\text{mid} \equiv (x_{(n+1-i)} + x_{(i)})/2 \quad \text{vs.} \quad x_{(n+1-i)} - x_{(i)} \equiv \text{spread}$$

In a symmetric distribution, each mid value should equal the median. Because the plot has slope = 0 when the distribution is symmetric, expansion of the vertical scale allows us to see systematic departures from flatness far more clearly.

The process of choosing a transformation from the ladder of powers can be made even easier, with a variation of the Mid - Spread plot suggested by Emerson & Stoto (1982). In this display, called the ***Mid vs. $z^2$ plot***, we plot the centered mid value, $(x_{(i)} + x_{(n+1-i)})/2 - M$ as the vertical coordinate, against a squared measure of spread,

$$\frac{\text{Lower}^2 + \text{Upper}^2}{4M} = \frac{(M - x_{(i)})^2 + (x_{(n+1-i)} - M)^2}{4M}$$

as the vertical coordinate. If this graph is approximately linear, with a slope, $b$, then $p = 1 - b$ is the indicated power for a transformation to approximate symmetry. For the automobile price data, this plot is shown in Figure 2. The slope of the line is 0.98, which indicates that log(PRICE) will have a nearly symmetric distribution.

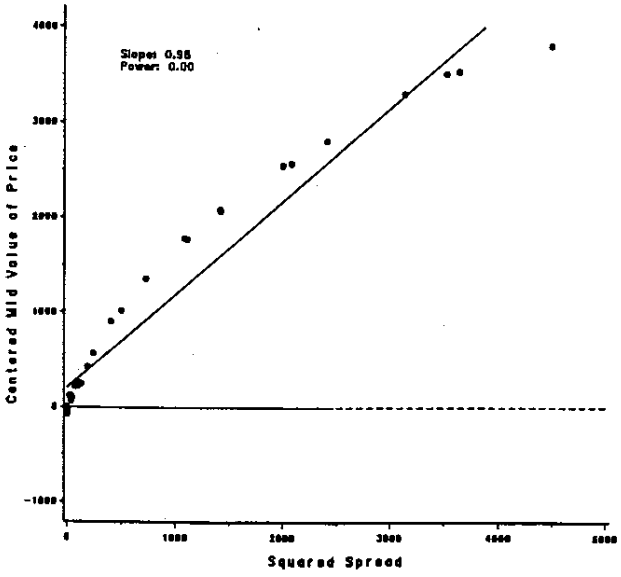


*Figure 2*: Mid vs. $z^2$ plot for PRICE

***Constructing the symmetry plots.*** The essential idea for all these plots is to fold the distribution of scores around the middle value. A simple way to do this is to sort the scores twice: once in ascending order, then in descending order, and then merge the two sets. This operation pairs $(x_{(1)}$ with $x_{(n)})$, $(x_{(2)}$ with $x_{(n-1)})$ and so on. These plots are all implemented in the SYMPLOT macro, which takes the

following parameters:

```
%macro SYMPLOT(
       data=_LAST_,  /* data to be analyzed      */
       var=,         /* variable to be plotted   */
       plot=MIDSPR,  /* Type of plot(s): any of  */
                UPLO, MIDSPR, MIDZSQ, or POWER   */
       out=symplot); /* Output data set          */
```

### *Enhanced Scatterplots*

In statistical graphics the scatterplot is the basic tool on which most methods are based. Scatterplots can be enhanced by adding information to help you interpret the display or understand aspects of the data that are not shown directly. Two such enhancements are illustrated:

- The first adds boxplots for the X and Y variable to the plot margins. This helps show the shape of the distributions of the individual variables and detect univariate outliers.

- The second adds an elliptical confidence region around the mean to highlight the joint relationship. With data for several groups this helps show the extent to which the relationship is the same for all groups.

### *Marginal Boxplots for X and Y*

A boxplot shows the interquartile range (IQR) by a box; "whisker" lines show range of 1.5 IQR beyond the upper and lower quartiles. Observations beyond this range are potential outliers and are usually plotted individually. A number of authors have described how to construct univariate boxplots with the ANNOTATE facility of SAS/GRAPH (Benoit, 1985; Olmstead, 1985).

The same techniques can be used to enhance a scatterplot with a boxplot for each of the X and Y variables with PROC GPLOT or for any of X, Y and Z with PROC G3D. In a two-variable display, the boxplot for X (or Y) is drawn parallel to the X-axis (Y-axis), at one extreme of the other axis, so as not to obscure the points.

Such a display is shown in Figure 3 for the WEIGHT and PRICE of automobiles. The boxplot for PRICE at the top of the figure shows clearly that the distribution of PRICE is positively skewed and that most of the car models are in the under-$9000 price range, with a handful over that. In a regression application, this dense packing of the observations toward the low end of PRICE might suggest the need for a transformation of PRICE to log(PRICE).

For generality, the program was designed as a SAS macro which constructs an ANNOTATE data set to draw a boxplot on any *one* axis in a scatterplot. The basic macro is called BOXAXIS. It takes the parameters shown in below:

```
%macro BOXAXIS(
       DATA=_LAST_,  /* Input dataset            */
       BOXOUT=_DATA_,/* Output ANNOTATE dataset  */
       VAR=,         /* Variable to be annotated */
       BAXIS=X,      /* Boxplot axis- X,Y or Z   */
       OAXIS=Y,      /* Other axis in the plot   */
```

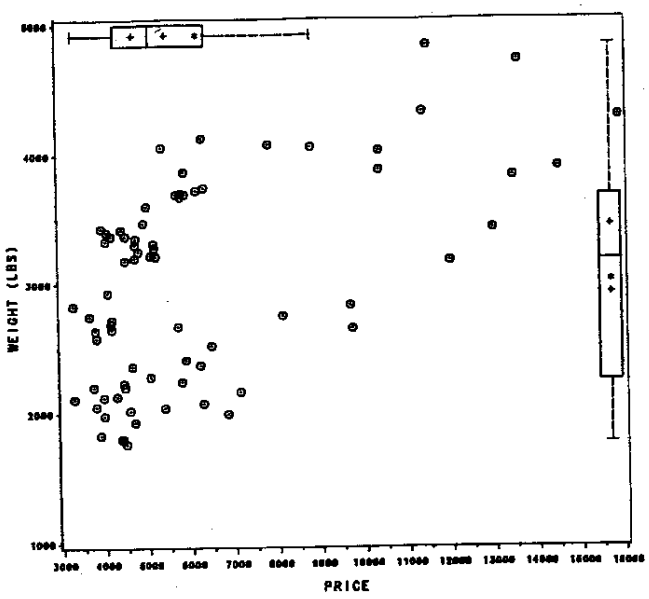


*Figure 3*: Scatterplot of WEIGHT vs. PRICE annotated by BOXANNO macro.

```
       PAXIS=Z,      /* The 3rd axis (for G3D)   */
       POS=99);    /* position on OAXIS 0<POS<100 */
```

To draw boxplots on both axes in a scatterplot such as Figure 3, BOXAXIS is called twice, once with BAXIS=X, OAXIS=Y, and once with BAXIS=Y, OAXIS=X. Each call produces an ANNOTATE data set (named by the BOXOUT parameter). These two data sets can be concatenated and used as the ANNOTATE= input to PROC GPLOT. These two steps are packaged together in another macro, BOXANNO, which produces one ANNOTATE data set containing the instructions to draw the boxplots for both axes.

```
/*-------------------------------------------*
 | BOXANNO macro - annotate boxplots for X & Y |
 *-------------------------------------------*/
%MACRO BOXANNO(
     DATA=_LAST_,  /* Data set to be plotted  */
     XVAR=,        /* Horizontal variable     */
     YVAR=,        /* Vertical variable       */
     OUT=BOXANNO   /* Output annotate dataset */
     );

   %BOXAXIS(DATA=&DATA, VAR=&XVAR,
     BAXIS=X,    OAXIS=Y,    BOXOUT=XANNO);
   %BOXAXIS(DATA=&DATA, VAR=&YVAR,
     BAXIS=Y,    OAXIS=X,    BOXOUT=YANNO);
* Concatenate the two annotate datasets ;
Data &OUT;
     Set XANNO YANNO;
%mend BOXANNO;
```

These macros do *not* plot the graph; it is up to the user to call PROC GPLOT after using BOXANNO. The program fragment below shows how BOXANNO was used to plot Figure 3.

```
%BOXANNO(data=auto,
     xvar=price, yvar=weight, out=boxanno );

title  h=1.5 'Weight vs. Price of Automobiles';
title2 h=1.2 'with Boxplots for Weight and Price';
proc gplot data=auto;
     symbol1 v='-' h=1.5;
     axis1 label=(a=90 r=0);
     plot weight * price /
          frame vaxis=axis1 vm=1 hm=1
          annotate = boxanno;
```

The BOXAXIS macro can also be used to annotate each axis of a PROC G3D scatterplot. To do this, the macro is called once for each axis, and the ANNOTATE data sets are concatenated before calling PROC G3D.

### Concentration ellipse

When you have $(x, y)$ data for several groups you may want to examine how the means, variances and correlations differ from group to group, and how these relate to the data for the total sample. Adding a concentration ellipse for each group to the scatterplot helps to show these relations.

The idea of a confidence interval for a single variable generalizes to an elliptical joint confidence region for two variables. For observations, $\mathbf{x}_i = (x_i, y_i)$ from a bivariate normal distribution, the elliptical region, called the ***concentration ellipse*** or ***data ellipse***, containing $(1 - \alpha)$ of the data is given by the values $\mathbf{x}$ satisfying

$$(\mathbf{x} - \bar{\mathbf{x}})' \; \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \le \chi^2_2(1-\alpha) \quad . \tag{3}$$

where $\bar{\mathbf{x}} = (\bar{x}, \bar{y})$ are the sample means, S (2×2) is the covariance matrix of $(x, y)$, and $\chi^2_2(1-\alpha)$ is the $(1 - \alpha)$ percentage point of the $\chi^2$ distribution with two degrees of freedom. The 50% data ellipse is analogous to the central box in the boxplot. Points on the boundary of the ellipse (where equality holds in Equation (3)) are calculated with PROC IML and output to a data set, CONTOUR, which can be drawn with the POLY and POLYCONT functions of the ANNOTATE facility.

```
/*-----------------------------------*
 |  Plot the contours using ANNOTATE |
 *-----------------------------------*/
data contour;
   set contour;
   by gp;
   length function $8;
   xsys='2'; ysys='2';
   if first.gp then function='POLY';
               else function='POLYCONT';
   line = gp+1;
```

The observations in the AUTO data are classified by region of origin. To help see how the relationship between WEIGHT and PRICE of an automobile is moderated by region of origin, the data shown in Figure 3 are redrawn in Figure 4, with a 50% data ellipse for each region.

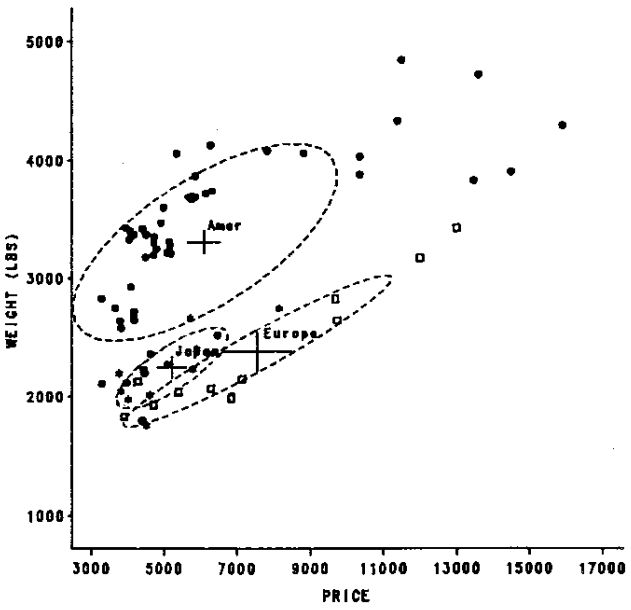


*Figure 4*: WEIGHT vs. PRICE of automobiles with data ellipse for each region of origin

### Lowess Smoothing

A useful and general technique for scatterplot smoothing is robust locally weighted regression smoothing, or ***lowess*** (Chambers et al., 1983). The procedure finds a smoothed fitted value, $\hat{y}_i$, for each $x_i$ by fitting a weighted regression to the points in the neighborhood of $x_i$. The points closest to $x_i$ receive the greatest weight. This is the locally weighted regression part of the procedure, and the weights are called ***neighborhood weights.***

The robust part works as follows. Once the fitted values, $\hat{y}_i$, have been found, the residuals, $r_i = y_i - \hat{y}_i$ are used to determine a new set of weights (***robustness weights***) so that points which have large residuals are down-weighted, and the locally weighted regression is repeated.

### LOWESS macro

The LOWESS macro reads the input data set into PROC IML, calculates the smoothed $\hat{y}$ values, and creates an output data set containing the original and smoothed data. The macro takes the following parameters:

```
%macro LOWESS(
     data=_LAST_,  /* input data set       */
     out=SMOOTH,   /* output data set      */
     x = X,        /* independent variable */
     y = Y,        /* variable to smooth   */
     id= ,         /* row ID variable      */
     f = .50,      /* lowess window width  */
     iter=2 );     /* number of iterations */
```

Figure 5 shows the plot of gas mileage vs. weight for the automobile data with the lowess smoothed curve. There are quite a few points in Figure 5 which are far from the curve, particularly near WEIGHT of 2000 and 4000. In the robust estimation iteration, these points receive robustness weights of

zero, and have no influence in determining the fitted curve. Without the smoothed curve, the relationship appears to be more curved than it actually is.

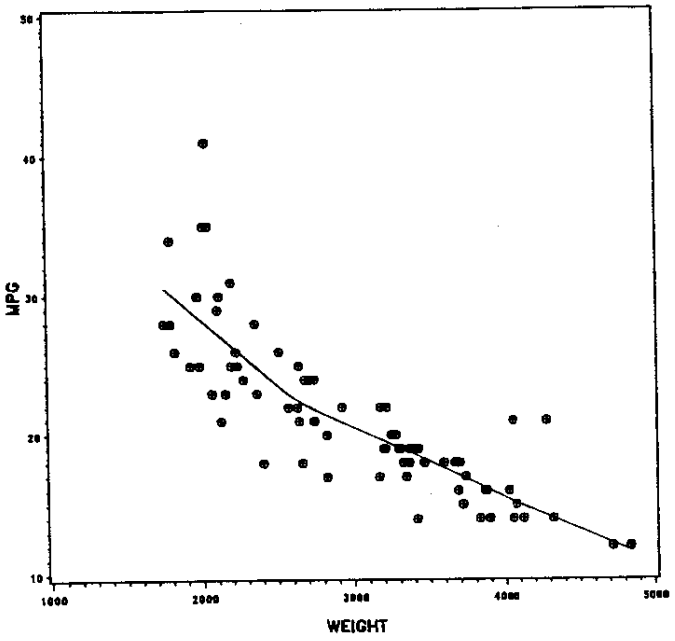


*Figure 5*: Gas mileage vs. weight with robust lowess smooth

### Two Way Tables

For two factor designs with one observation per cell, some displays suggested by Tukey (1977) are useful for viewing the fit and residuals for the additive model,

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \tag{4}$$

If the additive model does not fit, a diagnostic plot of the residuals from this model, $e_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$ against the "comparison values", $c_{ij} = \hat{\alpha}_i \hat{\beta}_j / \bar{y}_{..}$ can be used to find a power transformation of the data which will be more nearly additive. If a plot of $e_{ij}$ vs. $c_{ij}$ is approximately linear with a slope $b$, then the transformation $y \rightarrow y^{1-b}$ will reduce the degree of interaction.

The data set below contains the reaction times for three subjects to make true/false judgments for three types of sentences.

```
data RT;
     input Subject $ Sent1-Sent3;
cards;
SUBJ1  1.7  1.9  2.0
SUBJ2  4.4  4.5  5.7
SUBJ3  6.6  7.4 10.5
```

The plot of the two-way fit and residuals of the additive model, (4), for these data is shown in Figure 6. The ordinate is the response variable, reaction time. The intersections of the grid lines show the fitted value for the additive model, and the vertical lines show residuals which are larger in absolute value than $\sqrt{MSPE}$, the "Pure Error" mean square. The opposite corner pattern of the residuals in this figure is indicative of a non-additive relation.

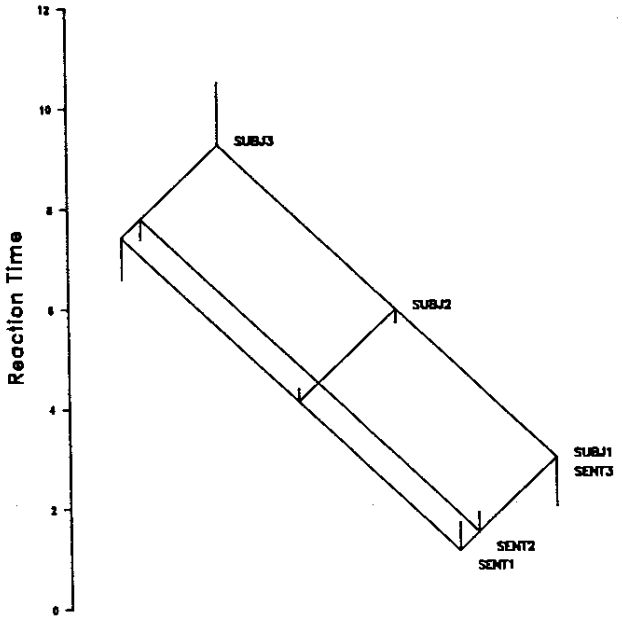


*Figure 6*: Two-way plot of fit and residuals

The diagnostic plot of the residuals and comparison values is shown in Figure 7, together with a least squares line. The slope of this line is $b = 1.57$, so $p = -0.57$, which would lead to the transformation $y_{ij} \rightarrow -1 / \sqrt{y_{ij}}$ which for these data would be interpreted as a measure of $\sqrt{\text{speed}}$.

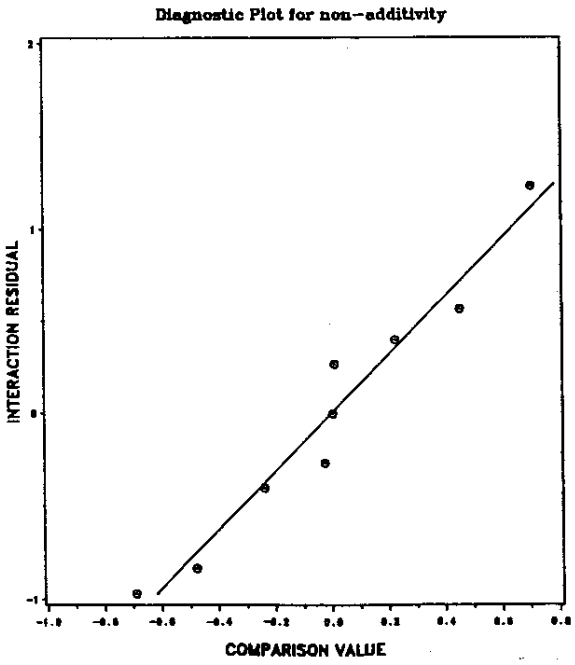


*Figure 7*: Diagnostic plot for transformation to additivity

### Scatterplot Matrix

Graphical methods can be particularly effective for dealing with multivariate datasets because a well-designed display can portray far more quantitative information in a comprehensible form than can be shown by other means. Figure 8 is an example of a technique for multivariate data called a ***scatterplot matrix***. It shows the relations among the variables PRICE, WEIGHT, MPG, and REPAIR (repair records) in the AUTO data, with the region of origin determining the plotting symbol. In this plot we can see:

- moderately strong (negative) correlations between MPG and both PRICE and WEIGHT.
- High mileage cars also tend to have better REPAIR records and are mostly Japanese.
- A positive relation between PRICE and WEIGHT for all three regions of origin, with US models generally heavier.
- The relationship between PRICE and REPAIR record is complex, and possibly nonlinear.

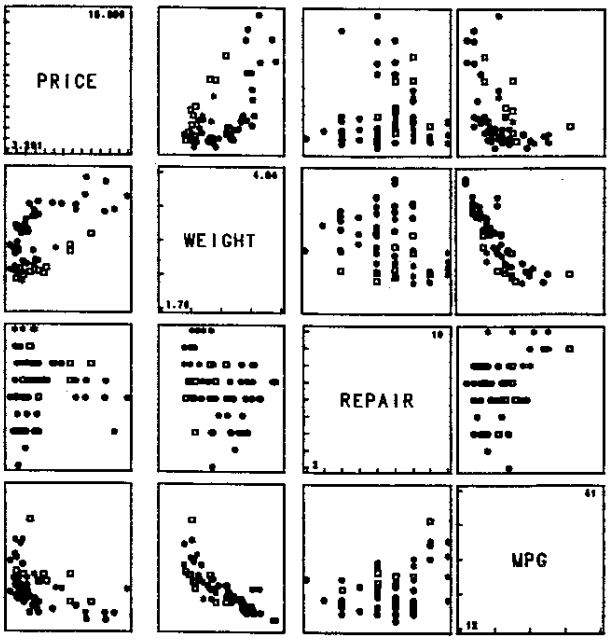


*Figure 8*: Scatterplot matrix for AUTO data. US models: circles, European models: squares, Japanese models: stars.

### SCATMAT macro

The scatterplot matrix is a good candidate for a general SAS macro. Here I'll describe the basic ideas behind the macro SCATMAT which constructs a scatterplot matrix for any number of variables.

For $p$ variables, $x_1, \ldots, x_p$, the scatterplot matrix is a $p \times p$ array in which the cell in row $i$, column $j$ contains the plot of $x_i$ against $x_j$. The diagonal cells are used for the variable names and scale markings. In the SAS macro language, this can be done with two nested %DO loops. In the schematic code fragment below, &VAR is the list of variables to be plotted (e.g., X1 X2 X3) from the data set &DATA, and &NVAR is the number of variables. The set of $p \times p$ plots is then displayed with a PROC GREPLAY step, which is also constructed by the SCATMAT macro.

```
%do i = 1 %to &nvar;                    /* rows */
  %let vi = %scan(&var , &i );
  %do j = 1 %to &nvar;                  /* cols */
    %let vj = %scan(&var , &j );
    %if &i = &j %then %do;    /* diag panel */
      data title;
         length text $8;
         xsys = '1'; ysys = '1';
         x = 50; y = 50;
         text = "&vi";
         size = 2 * &nvar;
         function = 'LABEL';  output;
      proc gplot data = &data;
        plot &vi * &vi / frame
             anno=title vaxis=axisl haxis=axisl;
      axisl label=none value=none major=none
           minor=none offset=(2);
      symbol v=none i=none;
    %end;
    %else %do;              /* off-diag panel */
      proc gplot data = &data;
        plot &vi * &vj / frame
             nolegend vaxis=axisl haxis=axisl;
      axisl label=none value=none major=none
             minor=none offset=(2);
      symbol v=+ i=none h=&nvar;
    %end;
  %end; /* cols */
%end;   /* rows */
```

***Author's Address***. For further information, contact:

Michael Friendly
Psychology Department, Rm 210 BSB
York University
Downsview, ONT, Canada M3J 1P3
BITNET: <FRIENDLY@YORKVM1>

### References

Benoit, P. (1985). Box-and-Whisker plots using the ANNOTATE facility. *SAS Communications, X(3)*, 34-35.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth.

Cleveland, W. S.(1985). *The Elements of Graphing Data.* Monterey, CA: Wadsworth Advanced Books.

Emerson, J. D., & Stoto, M. A. (1982). Exploratory methods for choosing power transformations. *Journal of the American Statistical Association, 77*, 103-108.

Olmstead, A. (1985). Box Plots using SAS/Graph Software. *Proceedings of the SAS User's Group International Conference, 10*, 888-894.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading, MA: Addison Wesley.