Paper 057-2010

Creating High-Quality Scatter Plots: An Old Story Told by the New SGSCATTER PROCEDURE

Xiangxiang Meng, University of Cincinnati, Cincinnati, OH

ABSTRACT

Scatter plot is a useful exploratory tool for multivariate data analysis and is one of the most commonly used statistical graphics. In traditional SAS/GRAPH®, it needs the cooperation of the GPLOT procedure, several SYMBOL statements and GOPTION statements to create a high-quality scatter plot. New with SAS® 9.2, the SGSCATTER procedure can produce a variety of scatter plots and put them into panels with different layouts within just a few lines of code. This paper will introduce how to create different types of scatter plot with PROC SGSCATTER and how to use ODS GRAPHICS and ODS styles to enhance the graph.

INTRODUCTION

Scatter plot is a useful exploratory tool for multivariate data analysis and is one of the most commonly used statistical graphics. In traditional SAS/GRAPH®, it needs the cooperation of the GPLOT procedure, several SYMBOL statements and GOPTION statements to create a high-quality scatter plot. New with SAS® 9.2, the SGSCATTER procedure can produce a variety of scatter plots and put them into panels with different layouts with just a few lines of code. The illustrative examples in this paper use data from the built-in data set SASHELP.CARS. ODS style HARVEST is used throughout the paper:

```
ods html style=harvest;
data cars;
  set sashelp.cars;
  where make in ('Jeep' 'Chevrolet' 'Ford' 'Chrysler');
run;
```

Variables	Description
Make	Car Manufacturer
MSRP	Manufacturer's suggested retail price
Invoice	Invoice price
MPG_city	Mileage per gallon in city
MPG_highway	Mileage per gallon on highway
Weight	The weight of car
Length	The length of car

Table 1. List of Variables used in examples

GETTING TO KNOW PROC SGSCATTER

PROC SGSCATTER creates various scatter plots with three distinct statements: PLOT, COMPARE and MATRIX statements. **PLOT** statement creates scatter plots that are paneled with independent horizontal and vertical axes.

```
* Example 1;
proc sgscatter data=cars;
  plot invoice*(weight length);
run;
```

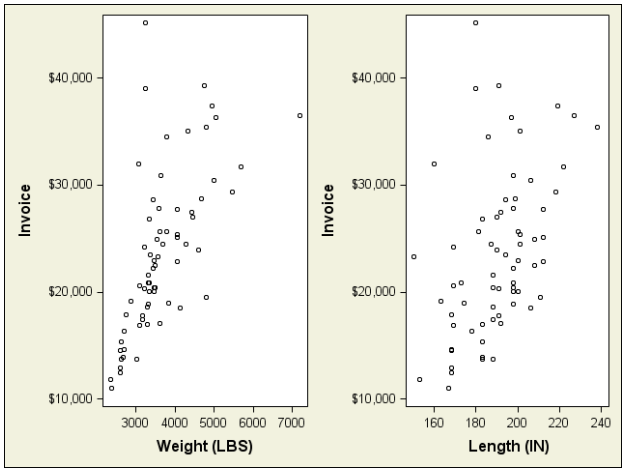


Figure 1. Scatter Plots Created by the PLOT Statement (Example 1)

COMPARE statement creates an M by N panel of scatter plots for M **Y**-variables and N **X**-variables. Scatter plots with shared axes are produced for each (**X**, **Y**) combination.

```
* Example 2;
* M=1, N=2;
proc sgscatter data=cars;
  compare y=invoice x=(weight length);
run;
```

MATRIX statement produces scatter plot matrix with shared axes. The matrix is M by M if M variables are specified in the statement.

```
* Example 3;
* matrix of 3 variable;
proc sgscatter data=cars;
  matrix invoice weight length;
run;
```

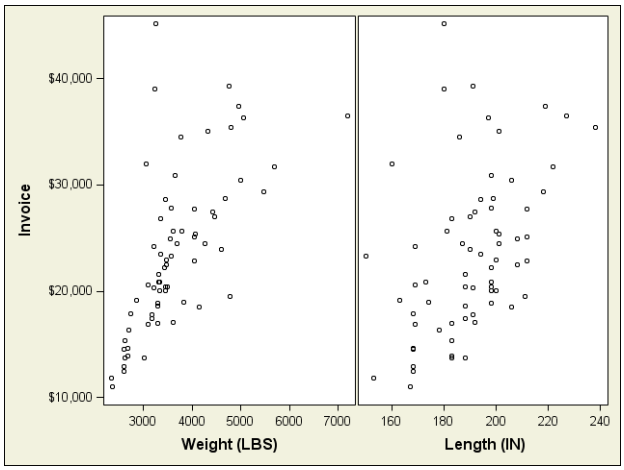


Figure 2. Scatter Plots Created by the COMPARE Statement (Example 2)

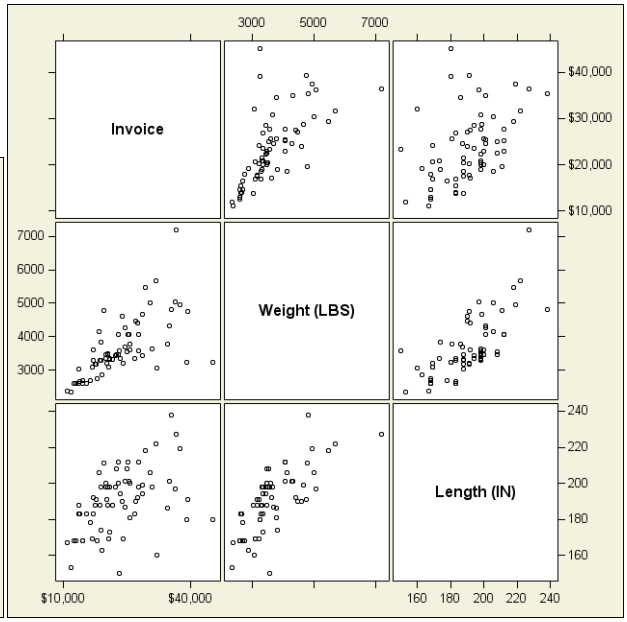


Figure 3. Scatter Plots Created by the MATRIX Statement (Example 3)

ADJUSTING THE LAYOUT

Similar to the PLOT statement in PROC GPLOT, the syntax of the PLOT statement in PROC SGSCATTER is flexible:

Statement	Number of Plots Created
`plot invoice*length;`	1
`plot invoice*length invoice*weight;`	2
`plot invoice*(length weight);`	2, (same layout as above)
`plot (MSRP Invoice)*(length weight);`	4
`plot (MSRP Invoice)*(length weight) MSRP*MPG_highway;`	5

Table 2. Syntax of the PLOT Statement

The layout of scatter plots is decided by the order of their appearance in the statement. SAS® optimizes the numbers of rows and lines used in a panel. This can be adjusted by two options **ROWS=** and **COLUMNS=**

```
* Example 4.1;
proc sgscatter data=cars;
  plot invoice*(weight length) / rows=2 columns=1;
run;
```

```
* Example 4.2;
proc sgscatter data=cars;
  plot invoice*(weight length) / rows=1 columns=2;
run;
```

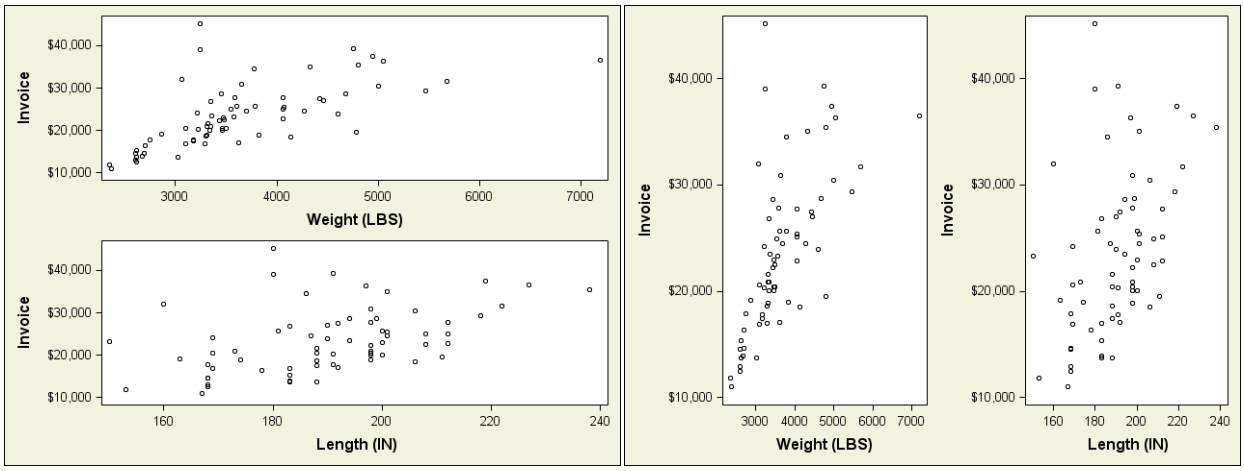


Figure 4. Use ROWS= and COLUMNS= for Layout Control

Compared with PLOT statement, COMPARE and MATRIX statement create panels with more standard layouts. In COMPARE statement, the layout is decided by the number of X-variable and Y-variable. MATRIX statement always creates a square matrix of plots.

SEVERAL WAYS TO ENHANCE THE GRAPH

USING GROUP= TO CREATE COMPARATIVE SCATTER PLOTS

It is always helpful to mark points with different colors and symbols for data from multiple groups or cohorts. In traditional SAS/GRAPH®, it takes a long code using PROC GPLOT and several SYMBOL statements to customize a comparative scatter plot. However, in the new PROC SGSCATTER, a high-quality comparative graph can be produced by adding the **GROUP=** option to one of its three statements:

```
* Example 5;
proc sgscatter data=cars;
  plot MPG_city*weight / group=make;
  where make in ('Ford' 'Chrysler' 'Chevrolet');
  title 'Scatter Plot by Make';
run;
```

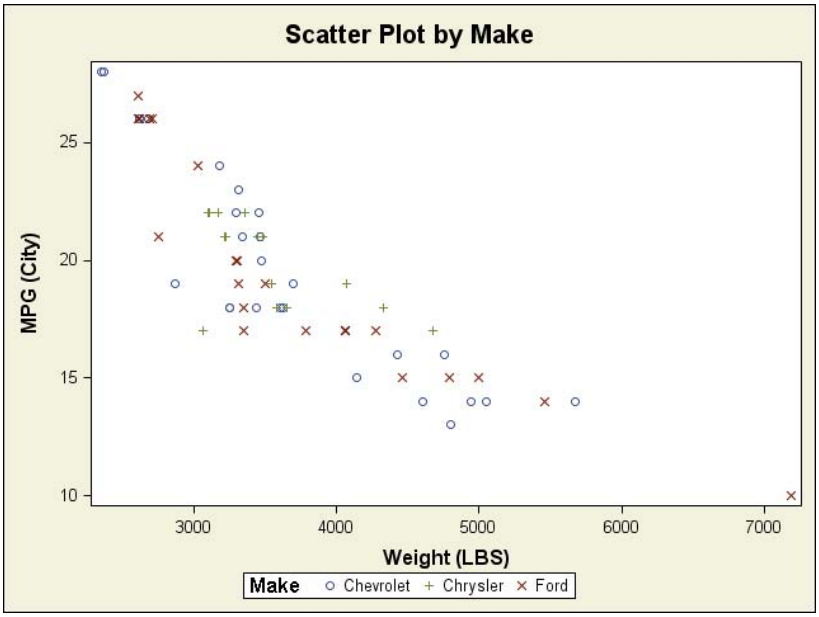


Figure 5. Comparative Scatter Plots Created by GROUP= Option

USING DATALABEL= TO CREATE LABELED GRAPHS

DATALABEL= is another option that saves a lot of code. In SAS® 9.1, no SAS/GRAPH® procedure can easily add labels to the points in a scatter plot, except using the built-in macro %PLOTIT. In PROC SGSCATTER, simply use the DATALABEL= option:

```
* Example 6;
* compute the means of MSRP and MPG_highway for each car maker;
proc sql;
  create table cars2 as
    select origin, make, mean(MSRP) as MSRP,
           mean(MPG_city) as MPG_city,
           mean(MPG_highway) as MPG_highway
    from sashelp.cars
    group by origin, make
    order by origin, make;
quit;

proc sgscatter data=cars2;
  plot MSRP*MPG_highway / datalabel=make group=origin grid;
  title 'Averaged MSRP vs. Highway MPG for Car Makers by Origin';
  format MSRP dollar6.0;
  label MSRP='Manufacturer Suggested Retail Price' MPG_highway='Highway MPG';
run;
```

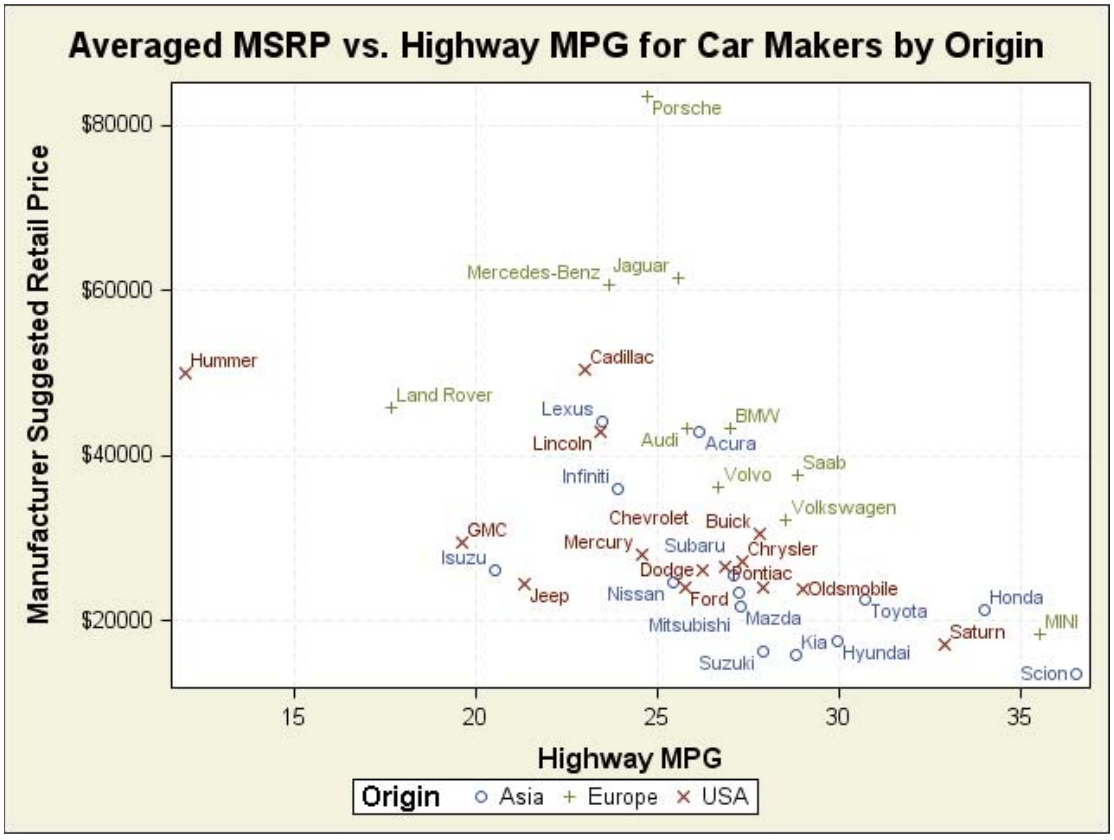


Figure 6. Labeled Scatter Plots Created by DATALABEL= Option

Note:

- IF no label variable is specified in the DATALABEL= option, the values of the Y-variable will be used as labels.
- Option GRID creates gridline according to the ticks on both axes.
- BY, WHERE, LABEL and FORMAT statements can be used in PROC SGSCATTER.

ADDING FITTING CURVES

Several types of fitting curve can be added to the scatter plots created by PROC SGSCATTER:

Option	Description	Available in statements:		
		PLOT	COMPARE	MATRIX
ELLIPSE=<= (*options*)>	Confidence or prediction ellipse.	Y	Y	Y
REG=<= (*options*)>	Linear, quadratic, or cubic regression fit with confidence limits	Y	Y	N
LOESS=<= (*options*)>	LOESS curve with linear or quadratic local fit, and confidence limits	Y	Y	N
PBSLINE=<= (*options*)>	Penalized B-spline curve with confidence limits	Y	Y	N

Table 3. Fitting Curves Produced by PROC SGSCATTER

This feature is shown in the following two examples:

```
* Example 7;
* Fitting a quadratic regression curve with 95% confidence limits;
proc sgscatter data=cars2;
  plot MSRP*MPG_highway / datalabel=make group=origin grid
                          reg=(degree=2 clm nogroup);
  title 'Averaged MSRP vs. Highway MPG for Car Makers by Origin';
  title2 '-- with quadratic regression fitting and conf. intervals --';
  format MSRP dollar6.0;
  label MSRP='Manufacturer Suggested Retail Price' MPG_highway='Highway MPG';
run;
```

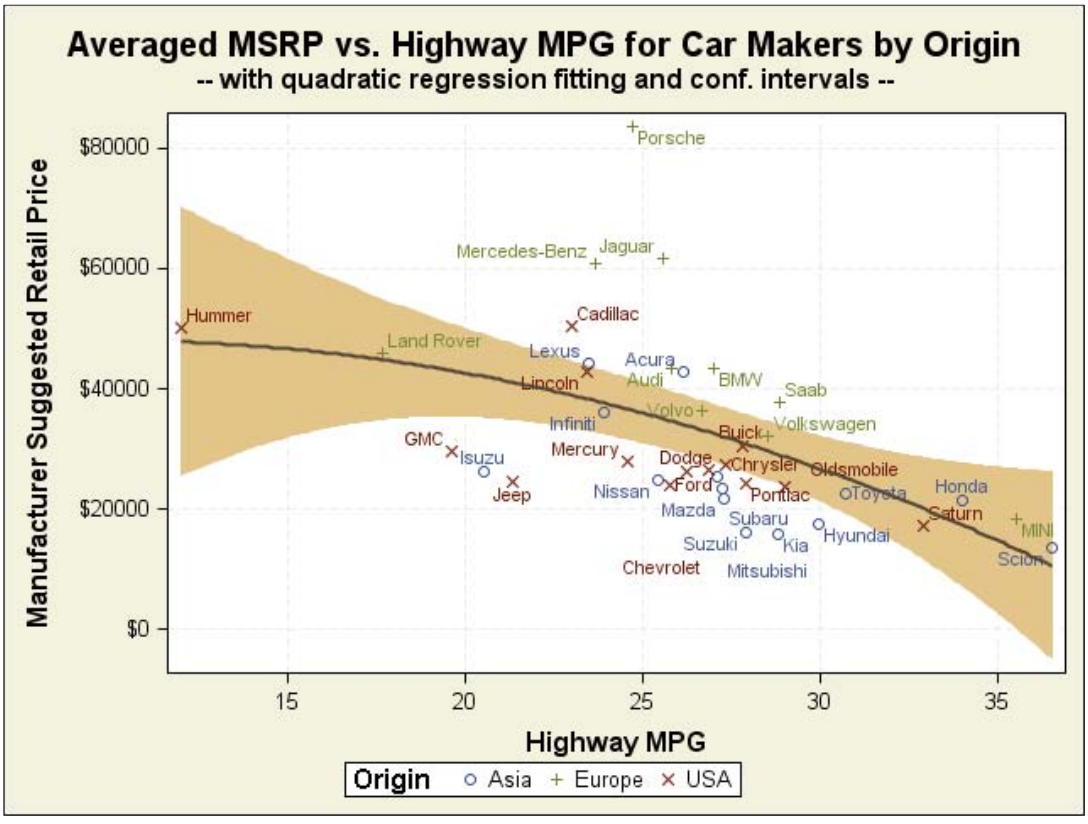


Figure 7. Scatter Plots with Regression Fittings and Confidence Intervals

Note: Use option **DEGREE=**2 to fit a quadratic curve, **CLM** to display the upper and lower bounds of the confidence interval for the mean response, and **NOGROUP** to fit one curve for the whole data.

```
* Example 8;
* Scatter plots with 95% prediction ellipse;
proc sgscatter data=cars2;
  compare y=MSRP x=(MPG_highway MPG_city)
          / group=origin ellipse=(alpha=0.05 type=predicted);
  title 'Averaged MSRP vs. Highway/City MPG for car makers by Origin';
  title2 '-- with 95% prediction ellipse --';
  format MSRP dollar6.0;
  label MSRP='Manufacturer Suggested Retail Price'
        MPG_highway='Highway MPG' MPG_city='CITY MPG';
run;
```

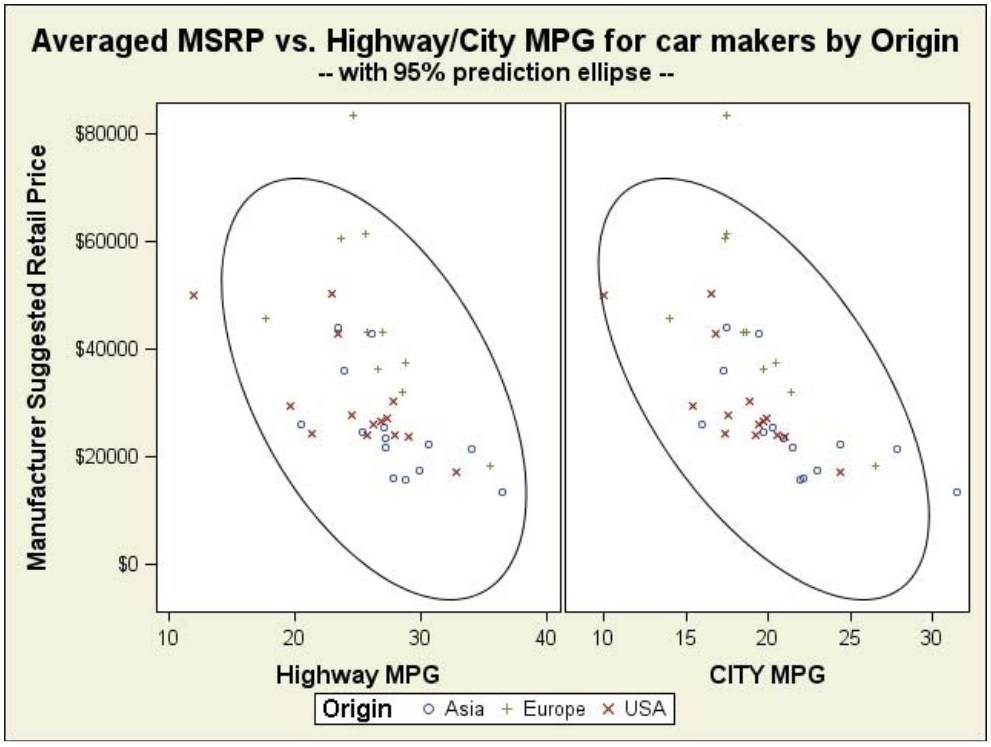


Figure 8. Scatter Plots with 95% Prediction Ellipses

Note: Option TYPE=PREDICTED creates confidence ellipse for a new observation. Using TYPE=MEAN to create confidence ellipse for means.

FILLING THE DIAGONALS OF A MATRIX

By default the MATRIX statement put the labels or names of the specified variables to the diagonal entries of the scatter plot matrix. The diagonals can be embellished with histograms, normal or kernel density fittings by the **DIAGONAL=** option:

```
* Example 9;
title 'Scatter Plot Matrix with
Histograms and Normal Fitting Curves';
proc sgscatter data=cars;
  matrix invoice weight length
  / diagonal=(histogram normal);
run; quit;
```

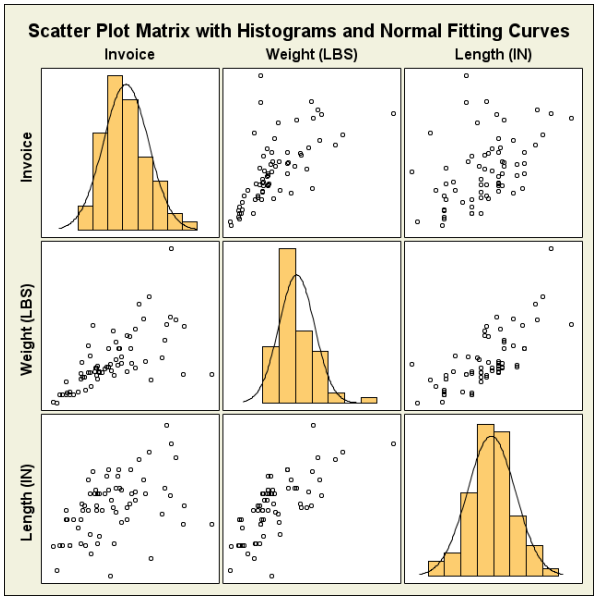


Figure 9. Enhanced Scatter Plot Matrix

COOPERATING WITH ODS STATEMENTS

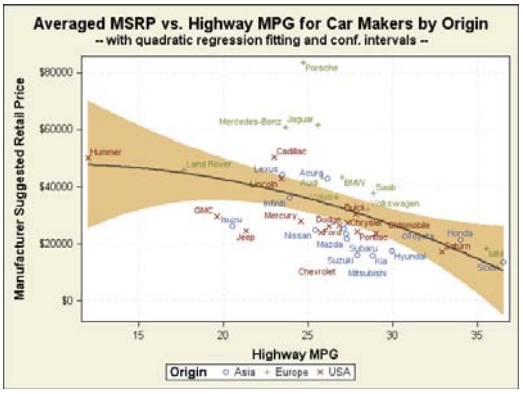
Traditional SAS/GRAPH® procedures create graphs that are saved in SAS® catalogs and can be displayed and edited in the GRAPH window. GOPTIONS, SYMBOL, AXIS and other SAS/GRAPH® statement control the appearance of the graph. On the other hand, the new statistical graphics procedures, such as PROC SGSCATTER, create and display graphs in standard image formats, such as BMP and PNG, by using the Output Delivery System (**ODS**) directly. Instead of GOPTION and SYMBOL statements, ODS statement, especially **ODS graphics**, are used in controlling the appearance of graphs produced by PROC SGSCATTER.

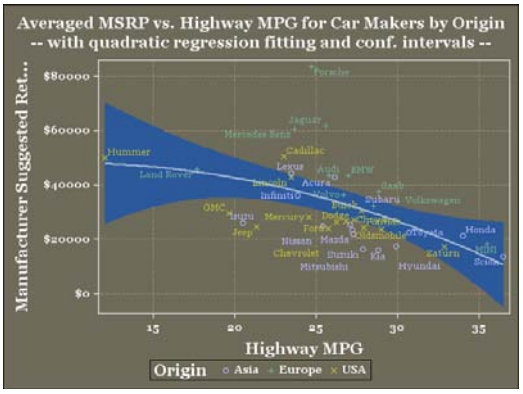
CHANGING ODS STYLE

No matter what ODS destination the graphs are being delivered ti, changing the style of the ODS output affects the appearance of the scatter plots created by PROC SGSCATTER. This is the easy way to get an embellished graph without sophisticated coding.

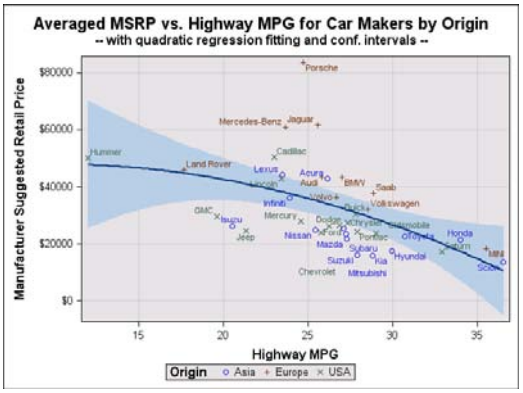
```
* Example 9;
* Apply different style to example 7;

ods html style=harvest;                ods html style=education;
```






```
ods html style=BarrettsBlue;           ods html style=Journal;
```



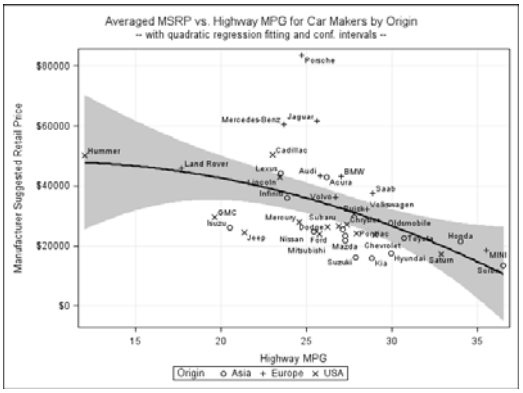


Figure 10. Scatter Plots with different ODS style

USING ODS GRAPHICS STATEMENT

The relationship of ODS GRAPHICS and PROC SGSCATTER is similar to that of GOPTIONS and PROC GPLOT. There are a variety of options in ODS GRAPHICS that control the size, resolution, naming and other properties of the graphs created by PROC SGSCATTER or by any other statistical graphics procedure. Some useful options are:

`WIDTH=, HEIGHT=`	Control the width and height of the graph
`IMAGENAME=, IMAGEFMT=`	Specify the file name and format of the graph
`BORDER=ON|OFF`	Control the display of graph border
`RESET=INDEX`	Reset index postfix used in file names in creating multiple graphs.
`RESET=ALL`	Reset all options

Table 4. Useful Options in ODS Graphics

The last example in this paper illustrates how to use ODS HTML, ODS GRAPHICS, and PROC SGSCATTER together to produce a high quality image saved in the pre-specified directory:

```
* Example 11;
ods html gpath='C:\' style=harvest;
ods graphics / reset=all width=12in height=6in border=off
               imagename='example' imagefmt=png;

proc sgscatter data=cars2;
  plot MSRP*(MPG_highway MPG_city)
       / datalabel=make group=origin
         grid reg=(degree=2 clm nogroup);

  title 'Averaged MSRP vs. Highway/City MPG for Car Makers by Origin';
  title2 '-- with quadratic regression fitting and conf. intervals --';

  format MSRP dollar6.0;
  label MSRP='Manufacturer Suggested Retail Price'
        MPG_highway='Highway MPG'
        MPG_city='City MPG';
run;
ods html close;
```

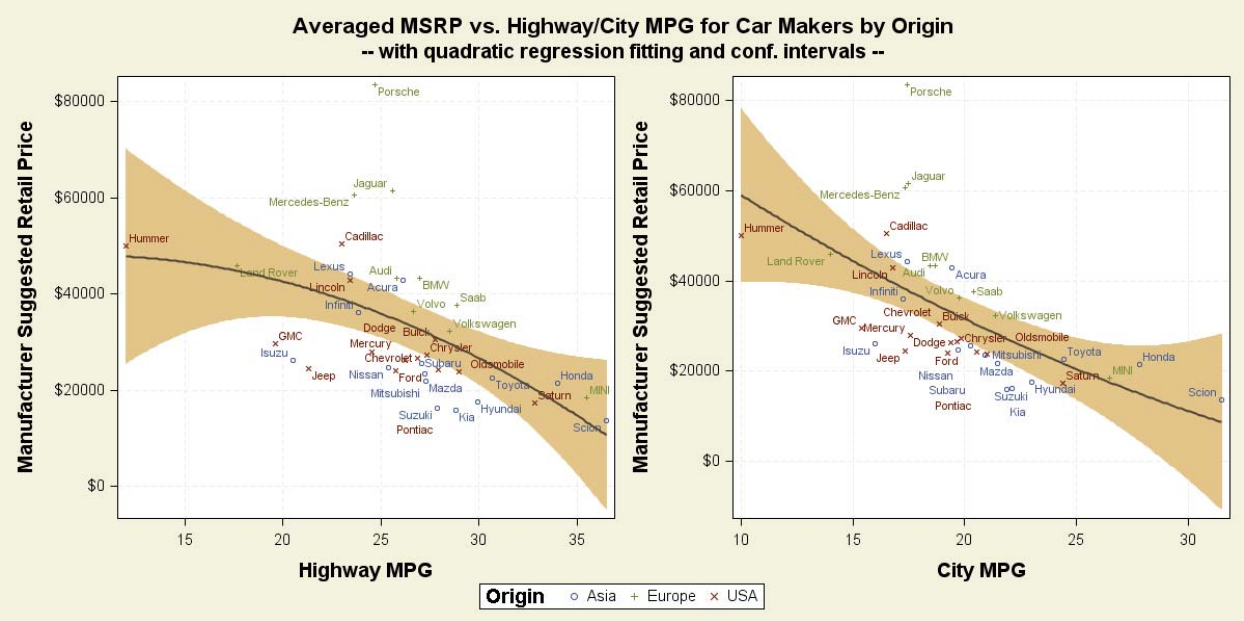


Figure 11. Using RPOC SGSCATTER together with ODS HTML, and ODS Graphics

CONCLUSION

The new SGSCATTER procedure provides an exciting method to produce paneled scatter plots. Its simple and natural syntax and seamless cooperation with ODS GRAPHICS make PROC SGSCATTER a powerful tool for data visualization. Start using SGSCATTER to explore your data!

REFERENCES

- Dan Heath, Secrets of the SG Procedures. SAS Global Forum 2009, Paper 324
- SAS Institute Inc., SAS/GRAPH 9.2: Statistical Graphics Procedures Guide.

CONTACT INFORMATION

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

Xiangxiang Meng
Department of Mathematical Science University of Cincinnati
mengxa@mail.uc.edu