

Using the SG Procedures to create and enhance scatter plots

Peter L. Flom

Peter Flom Consulting, LLC

NYASUG, June, 2011

Outline

Introduction

Basic scatter plots and enhancements

- Basic scatter plots with PROC SGPLOT

- Enhancing the scatter plot with PROC SGPLOT

- Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Outline

Introduction

Basic scatter plots and enhancements

- Basic scatter plots with PROC SGPLOT

- Enhancing the scatter plot with PROC SGPLOT

- Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Introduction

In this paper, I discuss scatter plots. Then to illustrate these, I start with a very basic example, and then illustrate some enhancements. Next, I show some problems that can occur, and illustrate some solutions. Finally, I give some more general advice on statistical graphics.

The SG procedures

- ▶ SG PROCs introduced in SAS 9.2, and enhanced in phase 2 of that release. Need SAS Graph.
- ▶ Basic scatter plots with SGPLOT - easy, but lots of options
- ▶ Fancy stuff with SGRENDER - very flexible, not entirely straightforward to use, different syntax
- ▶ VAST improvement on old SAS Graph methods.

Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Introduction to SGPLOT

- ▶ The PROC for basic scatter plots is PROC SGPLOT
- ▶ SGPLOT has many options, see the documentation. I will just show some examples.

A starting example

I found data on unemployment rate and infant mortality for each of the 50 states plus the District of Columbia. The data look like:

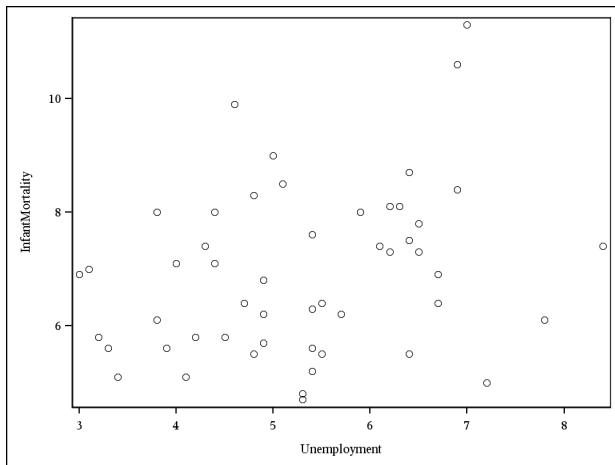
State	Infant Mortality	Unemployment
Alabama	9	5
Alaska	6.9	6.7
Arizona	6.4	5.5
Arkansas	8.5	5.1
California	5	7.2
Colorado	5.7	4.9
Connecticut	6.2	5.7

A starting example - the plot

A simple scatter plot can be done with the following code (assuming the data have been read in).

```
proc sgplot data = UnempIM;  *STARTS THE PROC;  
  scatter x = Unemployment y = InfantMortality;  
  *CREATES A PLOT, NOTE THE USE OF X = AND Y =;  
run;
```

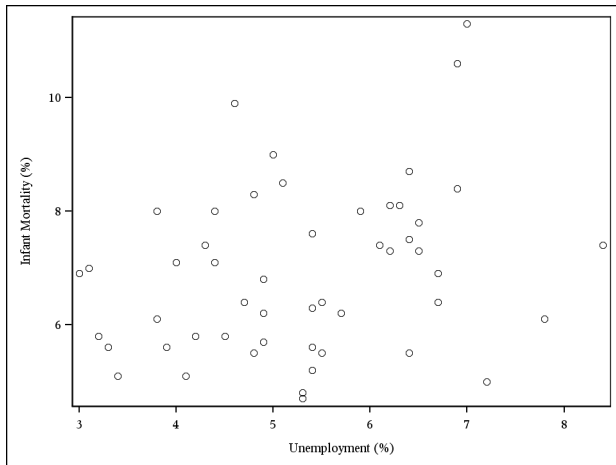
First scatter plot



Clarifying axes

```
proc sgplot data = UnempIM;  
  axis label = "Unemployment (%)";  
  *THIS SHOULD BE SELF EXPLANATORY,  
  THERE ARE OTHER AXIS OPTIONS AS WELL;  
  yaxis label = "Infant Mortality (%)";  
  scatter x = Unemployment y = InfantMortality;  
run;
```

Scatter plot with axes fixed



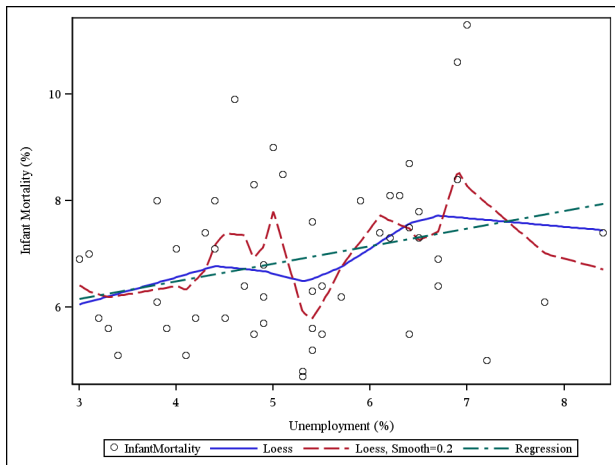
Adding information

- ▶ Not bad for 'out of the box'
- ▶ But we can easily include more information (make Tufte happy)
- ▶ First, let's add loess lines with different amounts of smooth, and a regression line

Code for fixing axes

```
proc sgplot data = UnempIM;
  xaxis label = "Unemployment (%)";
  yaxis label = "Infant Mortality (%)";
  scatter x = Unemployment y = InfantMortality;
  loess x = Unemployment y = InfantMortality
    /nomarkers;
  loess x = Unemployment y = InfantMortality
    /smooth = .2 nomarkers;
  *LOESS WORKS ON THE SAME DATA AS SCATTER,
  SMOOTH CAN BE ADJUSTED.
  NOMARKERS PREVENTS SAS FROM PLOTTING
  EACH POINT 3 TIMES;
  reg x = Unemployment y = InfantMortality;
run;
```

Scatter plot with smooths



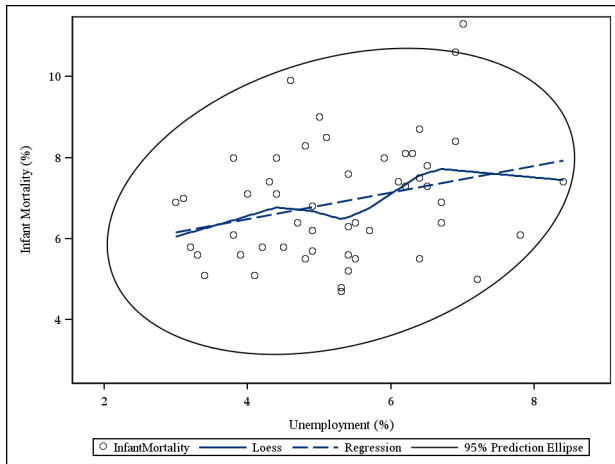
Adding an ellipse

And we might also want to add a confidence ellipse around the points:

```
proc sgplot data = UnempIM;
  axis label = "Unemployment (%)";
  yaxis label = "Infant Mortality (%)";
  scatter x = Unemployment y = InfantMortality;
  loess x = Unemployment y = InfantMortality
    /nomarkers;
  reg x = Unemployment y = InfantMortality;
  ellipse x = Unemployment y = InfantMortality;

*ELLIPSE, LIKE LOESS, OPERATES ON THE SAME DATA;
run;
```

Scatter plot with ellipse

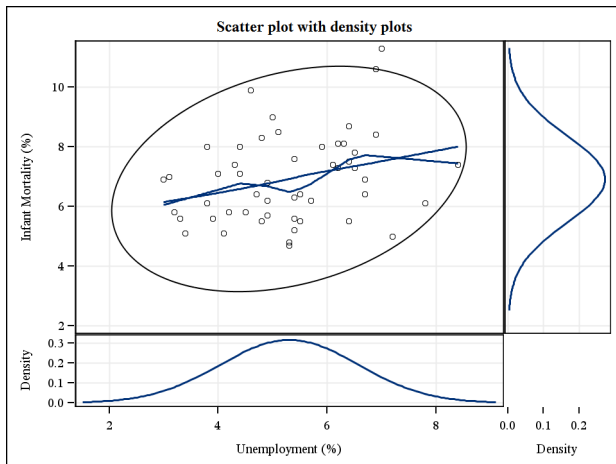


Getting fancy

That's all simple enough, and certainly adds information. But we can add more; we can look at the distribution of each variable separately and plot these in the margins. This requires use of the graph template language, or GTL. You can produce something like this;

A fancy scatter plot

The SAS System



Introduction to graph template language

- ▶ GTL allows very fine control over every aspect of a graph.
- ▶ SAS uses GTL to write graphical PROCs
- ▶ Warren Kuhfeld, of SAS, wrote a book that gives starting templates for many graphs
- ▶ Also see GTL User Guide and GTL Reference Manual
- ▶ MANY options

PROC TEMPLATE for a fancy graph (part 1)

```
proc template;
  define statgraph scatdens2;
  beginnograph;    *BEGIN DEFINING THE GRAPH;
    entrytitle "Scatter plot with density plots";
    *CREATE A TITLE;
  layout lattice/columns = 2 rows = 2
    columnweights = (.8 .2) rowweights = (.8 .2)
      columndatarange = union rowdatarange = union;
  *LAYOUT LATTICE...SETS UP A GRID OF GRAPHS;
  *COLUMNWEIGHTS AND ROWWEIGHTS SETS
    THE RELATIVE SIZE OF THE INDIVIDUAL
      COLUMNS AND ROWS;
```

Picking column and row weights

- ▶ Picking row and column weights affects the whole graph.
- ▶ You want the main point of the graph to occupy most of the space
- ▶ You also want everything to be visible

PROC TEMPLATE for a fancy graph (part 2)

```
columnaxes;  
  columnaxis /label = 'Unemployment (%)'  
    griddisplay = on;  
  columnaxis /label = '' griddisplay = on;  
endcolumnaxes;  
*COLUMNAXES SETS PARTICULAR  
  CHARACTERISTICS OF COLUMNS;  
*THE SECOND ONE HAS NO LABEL (NONE WOULD FIT)  
rowaxes;  
  rowaxis /label = 'Infant Mortality (%)'  
    griddisplay = on;  
  rowaxis /label = '' griddisplay = on;  
endrowaxes;
```

Row and column axes

The row axes labels show up in the left hand margin, and column axes in bottom margin, because they label the rows and columns, respectively.

PROC TEMPLATE for a fancy graph (part 3)

```
layout overlay; *STARTS THE ACTUAL GRAPHING OF DOTS
  scatterplot x = unemployment y = infantmortality
    *GRAPHS THE DOTS;
  loessplot x = unemployment
    y = infantmortality/nomarkers;
  loessplot x = unemployment
    y = infantmortality/smooth = 1;
ellipse x = unemployment y = infantmortality
  /type = predicted;
endlayout;
  densityplot infantmortality/orient = horizontal;
  densityplot unemployment;
endlayout;
endgraph;
end;
run;
```

PROC SGRENDER for the template

```
proc sgrender data = UnempIM template = scatdens2;  
  *NOW WE RENDER THE TEMPLATE WE CREATED;  
run;
```

Rick Wicklin of SAS has pointed out that, in his words, ‘For people who do not like to program, the %sgdesign macro brings up a GUI interface that allows you to create the second image using drag-and-drop and menus. For details and examples, see

`support.sas.com/documentation/cdl/en/
grstatdesignug/62589/HTML/default/viewer.htm'`

but I have not used this feature.

A note on scales

- ▶ Scales change in the graphs, even with same data
- ▶ Accommodates the added features such as density plots and ellipses
- ▶ In 'Real life' you would only show one of these graphs

Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Introduction to overplotting

- ▶ Scatter plots are very useful but can have problems
- ▶ One common one is overplotting, more than one observation has the same values, or almost the same values
- ▶ Solution depends on type and amount of overplotting
- ▶ Sometimes jittering is enough
- ▶ With more data, change the plotting symbol
- ▶ With even more, consider parallel box plot

A data set

Here I create a data set with

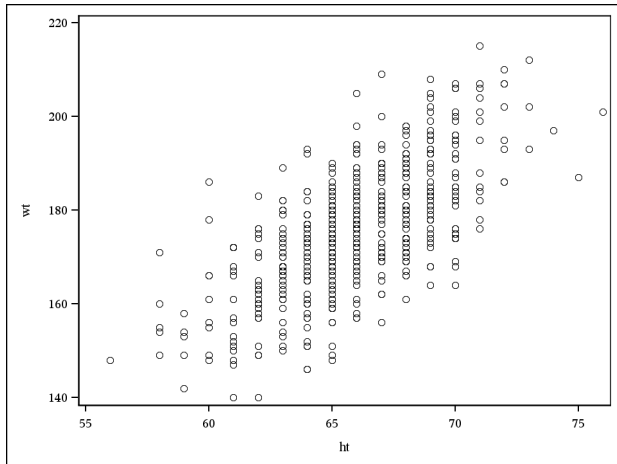
1. Real valued simulated heights and weights (realht and realwt)
2. The same heights and weights rounded to the nearest inch and pound (ht and wt)
3. A jittered version of these (jitht and jitwt)
4. SAS doesn't support jittering in SG, you have to do it in data
5. Judgement needed to pick amount of jitter

A data set

```
data htwt;  
  do i = 1 to 10000;  
    realht = rannor(1828282)*3 + 66;  
    realwt = realht*2 + realht**2*.01  
      + 10*rannor(12802194);  
    ht = round(realht,1);  
    wt = round(realwt,1);  
    jitht = ht+rannor(1818282);  
    jitwt = wt+rannor(199328282);  
    output;  
  end;  
run;
```

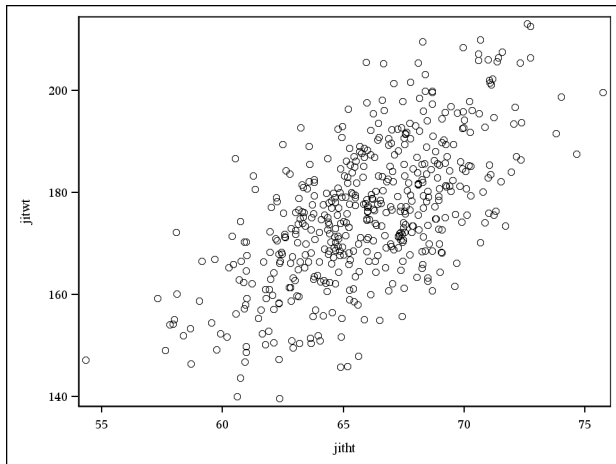
Moderate overplotting due to discretization

If we have a data set of 500 people with rounded height and weight, the plot will not show all the points clearly



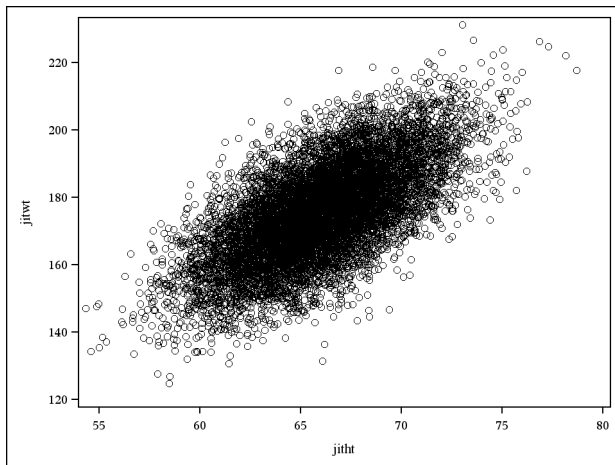
Jittering for moderate overplotting

Here, simply jittering the data works well



More severe overplotting

However, if we have 10,000 points, jittering is not enough

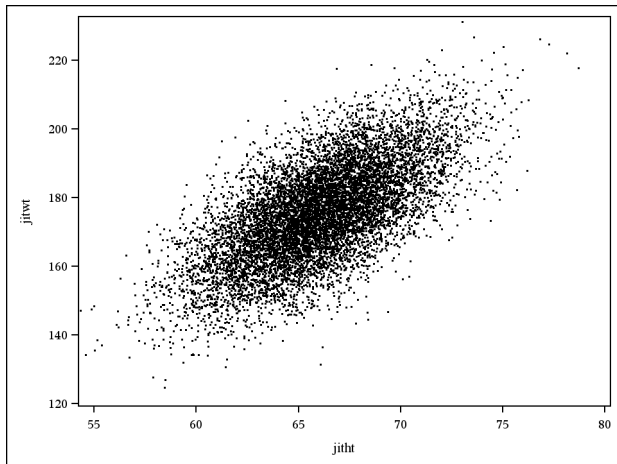


Changing the plotting symbol

We can change the plotting character and its size with the following program:

```
proc sgplot data = htwt;  
  scatter x = j1ht y = j1wt/  
    markerattrs = (size = 2 symbol = circlefilled);  
run;
```

Scatter plot with different symbol

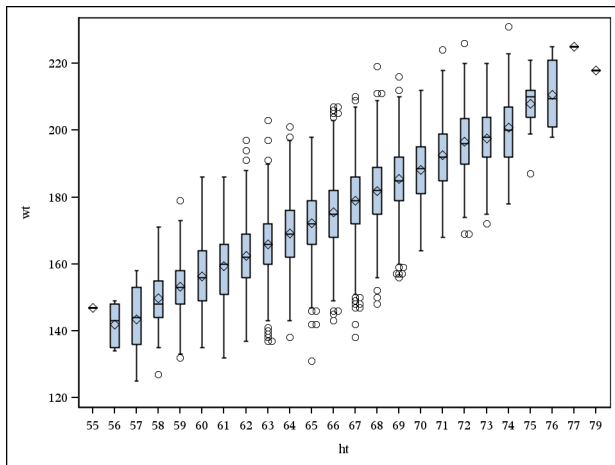


Parallel boxplots as an alternative

An alternative is to abandon the scatter plot and use parallel boxplots:

```
proc sgplot data = htwt;  
  vbox wt/category = ht spread;  
  *THE SPREAD OPTION PREVENTS OVERLAP;  
run;
```

Parallel boxplots



Getting fancy again

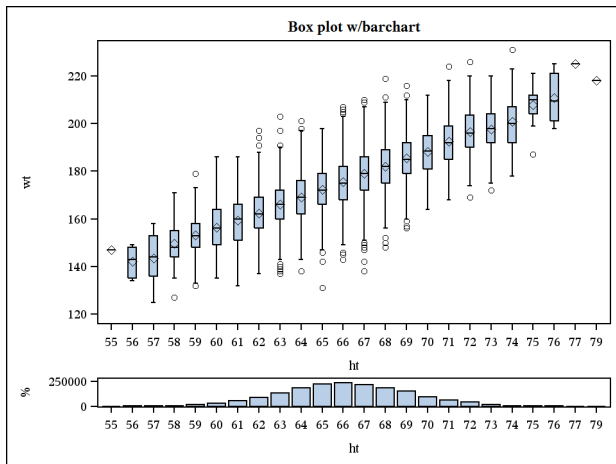
```
proc template;
  define statgraph fancybox;
  begingraph;
    entrytitle "Box plot w/histogram";
    layout lattice/rows = 2 columns = 1
      order = columnmajor rowweights = (.8 .2);
    columnaxes;
      columnaxis /griddisplay = on;
columnaxis /label = '' griddisplay = on;
    endcolumnaxes;
      boxplot x = ht y = wt;
      barchart x = ht y = wt;
    endlayout;
  endgraph;
end;
run;
```

Rendering the plot

```
proc sgrender data = htwt template = fancybox;  
run;
```

Parallel boxplot with bar chart

The SAS System



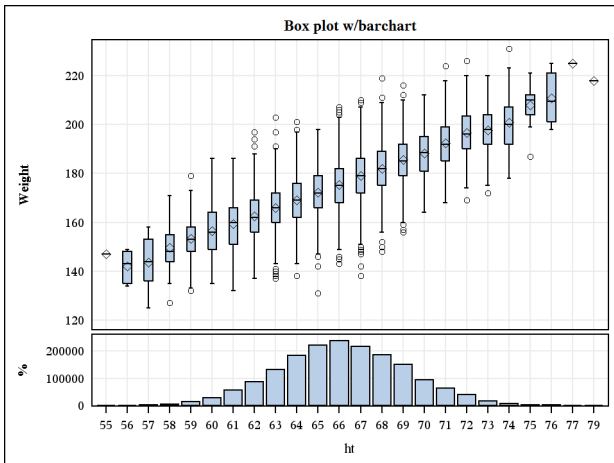
Externalizing the axes

```
proc template;
  define statgraph fancybox2;
    begingraph;
      entrytitle "Box plot w/barchart";
      layout lattice/rows = 2 columns = 1 rowweights = 1
        rowdatarange=union
        rowgutter=3px
        coldatarange = union;

      rowaxes;
        rowaxis / griddisplay=on label="Weight"
        rowaxis / griddisplay=on label="%"
      endrowaxes;
        boxplot x = ht y = wt;
        barchart x = ht y = wt;
    columnaxes;
      columnaxis /griddisplay = on;
    endcolumnaxes;
  endlayout;
end;
```

Parallel boxplot with bar chart, external axes

The SAS System



Outline

Introduction

Basic scatter plots and enhancements

Basic scatter plots with PROC SGPLOT

Enhancing the scatter plot with PROC SGPLOT

Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Some principles

- ▶ Graphics are not for precise lookup, they are not table substitutes
- ▶ Pie is delicious, but not nutritious
- ▶ If a graph is worth thousands of words, it may be worth an edit or two

Questions to ask

- ▶ What to graph
- ▶ Why do you want to graph it?
- ▶ Who will see it?
- ▶ How are they going to look at it?
- ▶ When do you need it done?

Outline

Introduction

Basic scatter plots and enhancements

- Basic scatter plots with PROC SGPLOT

- Enhancing the scatter plot with PROC SGPLOT

- Getting Fancy with SGRENDER and the GTL

Overplotting

Thoughts on graphics

Summary

Summary

Scatter plots are a very valuable graphical tool. The SG PROCs in SAS allow many scatter plots to be produced easily, and the graph template language allows very fine control over all aspects of a graph.

Contact information

Peter L. Flom

515 West End Ave

Apt 8C

New York, NY 10024

peterflomconsulting@mindspring.com

(917) 488 7176

www.statisticalanalysisconsulting.com

SAS stuff

SAS[®] and all other SAS Institute Inc., product or service names are registered trademarks or trademarks of SAS Institute Inc., in the USA and other countries. ® indicates USA registration. Other brand names and product names are registered trademarks or trademarks of their respective companies.