Paper 4989-2020

Add SGSCATTER, SGPLOT, SGPANEL and SGRENDER Procedures to Your SAS® Toolbelt

Alex Chaplin, Bank of America

ABSTRACT

SGSCATTER, SGPLOT, SGPANEL and SGRENDER procedures provide a range of graphical data visualization options to help any data analyst tell the story of the data. If you are a SAS Enterprise Guide® user be aware that SAS Enterprise Guide generates graphs using the GPLOT procedure which is the predecessor to PROC SGPLOT. To really bring your data alive you will need to be familiar with the expanded capabilities in PROC SGPLOT and introduce yourself to PROC SGPANEL.

INTRODUCTION

We will demonstrate the SG procedures to test some of the assumptions for simple linear regression, compare sales and profits across different time periods and plot in 3 dimensions.

This paper is not intended to demonstrate the fundamentals of linear regression but to demonstrate use of the SG procedures in data analysis.

DEMO 1: CAN WE PREDICT IRIS SEPAL LENGTH?

We start with the null hypothesis that we cannot predict sepal length from sepal width.

We will use the famous iris data set handily available in SASHELP.IRIS.

The iris data set is based on the work of Ronald Fisher and Edgar Anderson. The data set contains a sample of 50 each of 3 iris species to create a linear discriminant analysis model to distinguish the species from each other. We will use the measurement of iris sepal widths from each species to see if sepal width is predictive of sepal length. Our null hypothesis is that we cannot use sepal width to predict the value of sepal length. The sepals are the leaves that cover a bud and remain immediately below the flower after it blooms.

The assumptions of linear regression are

- Normal distribution
- Linear relationship between x (input or predictor) variables and y (output or predicted) variable
- Homoscedasticity – equal variance
- No multicollinearity
- Independent errors

Let's run PROC SGPLOT to create a frequency histogram and distribution curve for sepal width for the iris Setosa species. This will help us determine if the data is normally distributed.
Please note the following in the code

- Use of a macro variable and where clause to filter the data.
- Including the macro variable in the title. Note the use of double quotation marks in the title to resolve the value of the macro variable.
- Scale=count signifies we want to look at frequency

```
%let species=Setosa;
proc sgplot data=sashelp.iris(where=(species="&species."));
   title "Sepal Width Frequency Distribution for &species.";
   histogram sepalwidth / scale=count;
   density sepalwidth / scale=count;
run;
title;
```

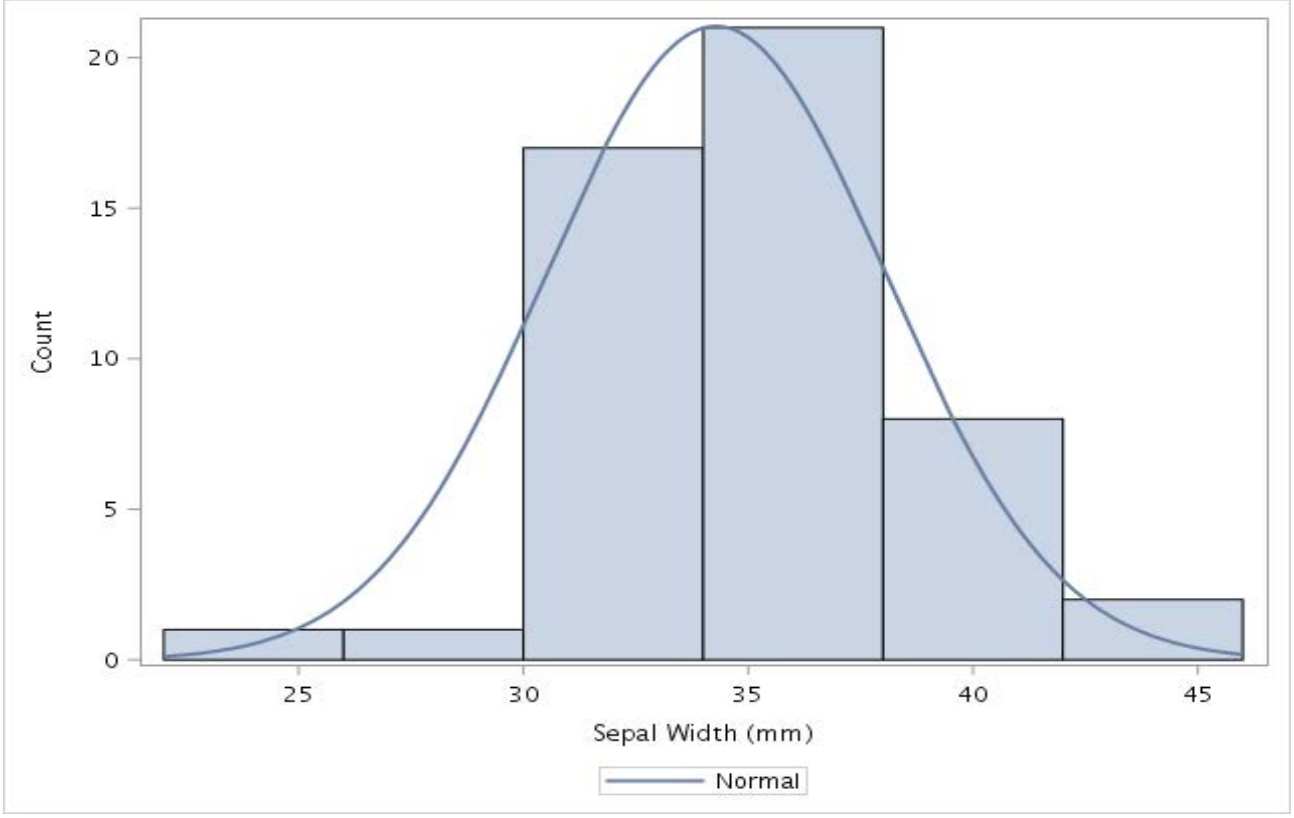

Figure 1: Sepal width frequency distribution for Setosa

The distribution looks normal, so we will test another two assumptions of linear regression.

- Does sepal width have a linear relationship with sepal length?
- Do we see homoscedasticity? Homoscedasticity means variances are consistent across all pairs of x and y. In this case sepal width (x) and sepal length (y).

This time we run PROC SGSCATTER to create a scatter plot of sepal width (x) and sepal length (y).

Notes on the code:

- We shall take the same approach as the previous code in using macro variable, where clause and title.
- Use of / reg to create a regression line. This is the line of best fit for the points.

```
%let species=Setosa;
proc sgscatter data=sashelp.iris(where=(species="&species."));
   plot sepallength*sepalwidth / reg;
   title "Sgscatter scatter plot of Sepal Width (x) against Sepal Length (y) for &species.";
run;
```

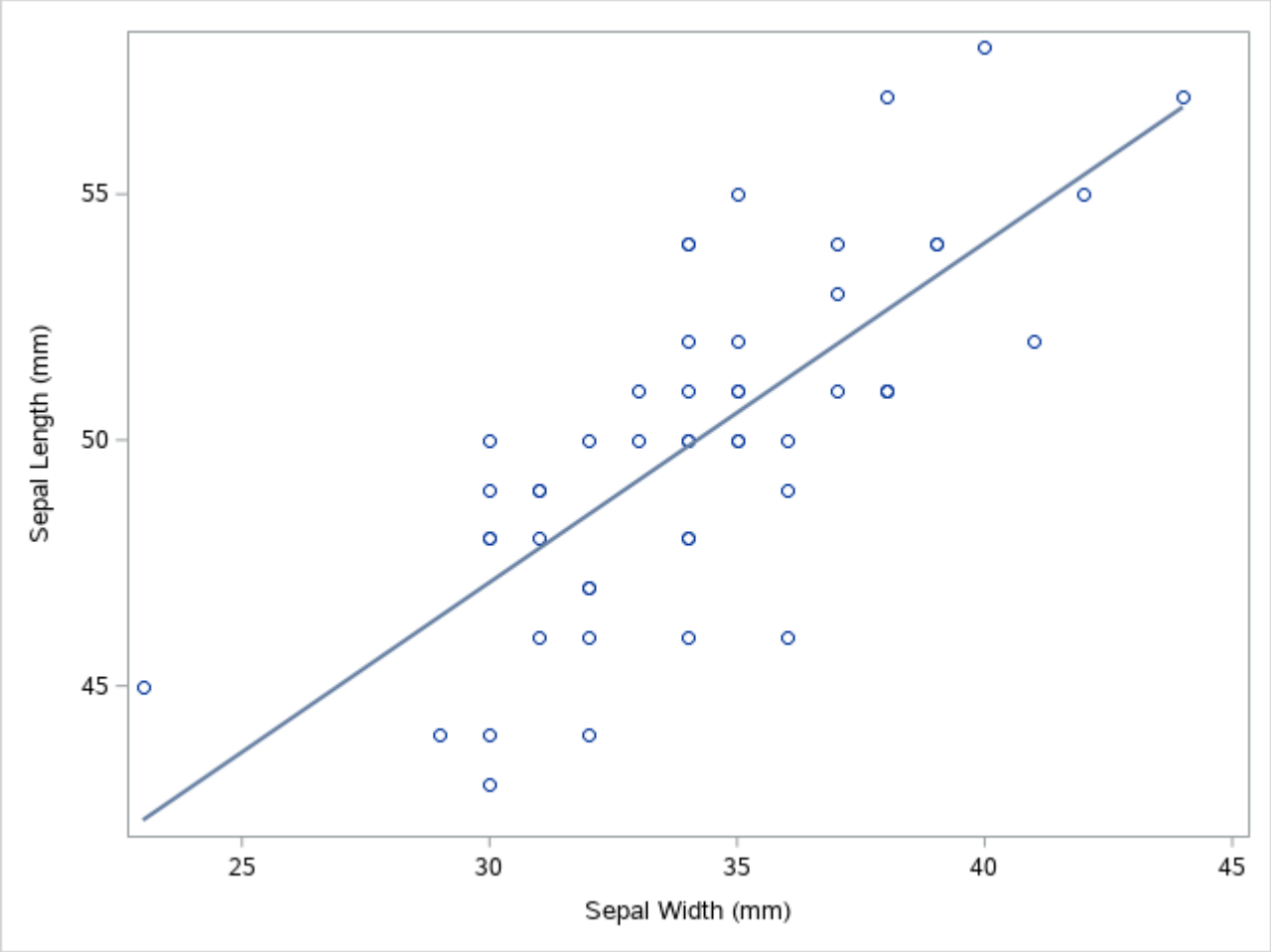

Figure 2: PROC SGSCATTER scatter plot of sepal width (x) against sepal length (y)

Checking our next two assumptions of linear regression we see the following:

- The regression line and data points suggests a linear relationship between sepal width and sepal length.
- The data points are homoscedastic meaning variance is consistent across all pairs of sepal width and sepal length. If, for example, the points display a cone shape at either end we would say they are heteroscedastic meaning variance is inconsistent across all pairs of sepal width and sepal length.

The other two assumptions of linear regression are no multicollinearity and independent errors also known as no auto correlation.

- Multicollinearity exists when two or more of the predictor variables have a strong linear relationship. The assumption of no multicollinearity is met because we have only one predictor variable which is sepal width.
- Auto correlation exists when the residuals also known as errors, which is the difference between the actual and predicted values, are not independent from one another. We shall assume the assumption of independent errors is met.

There are a number of statistical tests not covered in this paper that should be followed to perform more rigorous tests of the assumptions for linear regression. They are available as part of the SAS/STAT procedures. Anyone interested in learning more should check out the pre-requisite courses for the SAS Statistical Business Analyst certification at https://www.sas.com/en_us/certification/credentials/advanced-analytics/statistical-business-analyst.html.

Our data meets the assumptions for linear regression. We are ready to perform a simple linear regression to predict sepal length from sepal width for the iris Setosa species.

Notes on the code: Call the REG procedure to perform simple linear regression

- Simple linear regression means there is one predictor or input variable, which in this case is sepal width. Multiple linear regression uses more than one predictor or input variable.
- The predictor variable, sepal width appears to the right of the equal sign, whereas the output or predicted variable sepallength appears to the left of the equals sign.
- Same approach as previous code in using macro variable and where clause to select for species Setosa.

```
%let species=Setosa;

proc reg data=sashelp.iris(where=(species="&species.")) outest=est1;

  eq1: model  sepallength=sepalwidth;

run;
```

Please note the following in Figure 3: Output from PROC REG.

- Under Parameter Estimates the values of Pr > |t| are <.0001 for the input variable SepalWidth and the intercept. Under Analysis of Variance the value Pr > F is <.0001 for the model. The means the chance of the results being random is less than 0.01% at a 95% confidence limit. Because we use the default value of a 95% confidence limit in PROC REG, any value of Pr <.0500 (5%) is deemed statistically significant. Since the values of the input variable SepalWidth, the intercept and the model are all statistically significant, we can reject the null hypothesis that we cannot predict the value of sepal length based on the value of sepal width.
- Under Root MSE the value of Adjusted R Square (Adj R-Sq) is 0.5420. This means the variability in the values of sepal width explains 54.20% of the variability in the values of sepal length. This is a measure of predictive power. How high you want to get to for the value of Adjusted R Square is subjective depending on the type of model. Does the model predict who is going to live or die based on a course of treatment (go for as high

as possible without over fitting) or what proportion of people will select water versus soda at your next gathering (0.5 or better is good)? Generally speaking the higher the value of Adjusted R Square the better, but you need to be aware of the risk of over fitting your model against your sampled data to achieve a higher Adjusted R Square value. Always go with the Adjusted R Square rather than R Square, because Adjusted R Square contains a penalty for increasing the number of input variables in your model. This makes Adjusted R Square a more effective measure than R Square for comparing the performance of models with a different number of input variables.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	335.68848	335.68848	58.99	<.0001
Error	48	273.13152	5.69024		
Corrected Total	49	608.82000			

Root MSE	2.38542	R-Square	0.5514
Dependent Mean	50.06000	Adj R-Sq	0.5420
Coeff Var	4.76513		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	26.39001	3.10014	8.51	<.0001
SepalWidth	Sepal Width (mm)	1	0.69049	0.08990	7.68	<.0001

Figure 3: Output from PROC REG

There is more than one way of doing the same thing in SAS. The next couple of examples demonstrate creating a scatter plot using PROC SGPLOT instead of PROC SGSCATTER and creating multiple plots using PROC SGPANEL.

Notes on the code: Create scatterplot using sgplot instead of sgscatter

- We shall take the same approach as the previous code in using macro variable and where clause to select for species Setosa. Use of scatter key word to create a scatter plot
- Use of reg keyword with / clm and cli options to create a regression line with confidence limits for the mean CLM and individual predicted values CLI

```
%let species=Setosa;
proc sgplot data=sashelp.iris(where=(species="&species."));
  scatter x=sepalwidth y=sepallength;
  title "Sgplot scatter plot of Sepal Width (x) against Sepal Length (y)";
  reg x=sepalwidth y=sepallength /clm cli;
run;
```

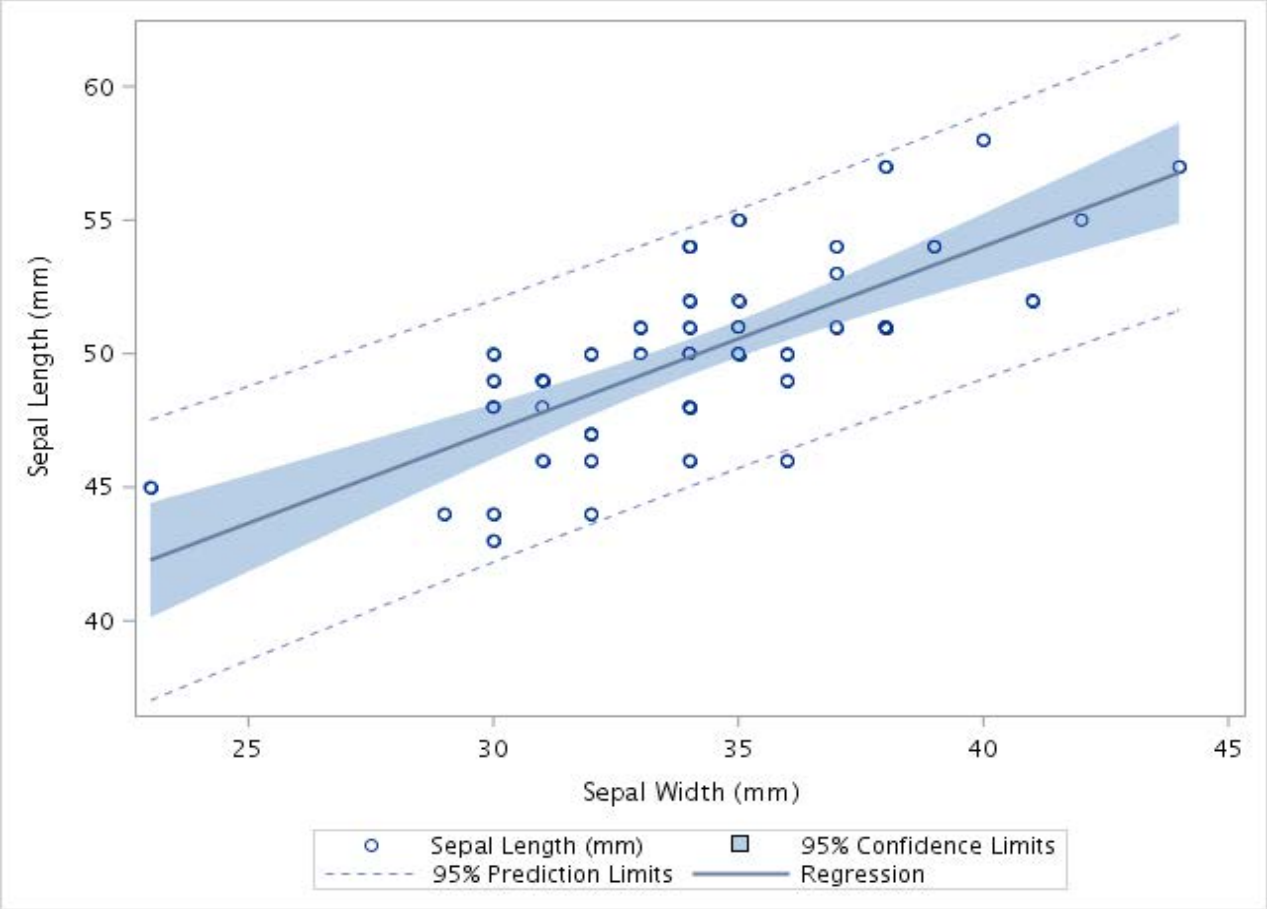

Figure 4: PROC SGPLOT scatter plot of sepal width (x) against sepal length (y)

Notes on the code: Modify PROC SGPLOT code to create a panel plot for all species

- Change sgplot to sgpanel
- Add panelby statement for species

```
proc sgpanel data=sashelp.iris;
  panelby species;
  scatter x=sepalwidth y=sepallength;
  title "Scatter plot of Sepal Width (x) against Sepal Length (y)";
  reg x=sepalwidth y=sepallength /clm cli;
```

```
run;
```

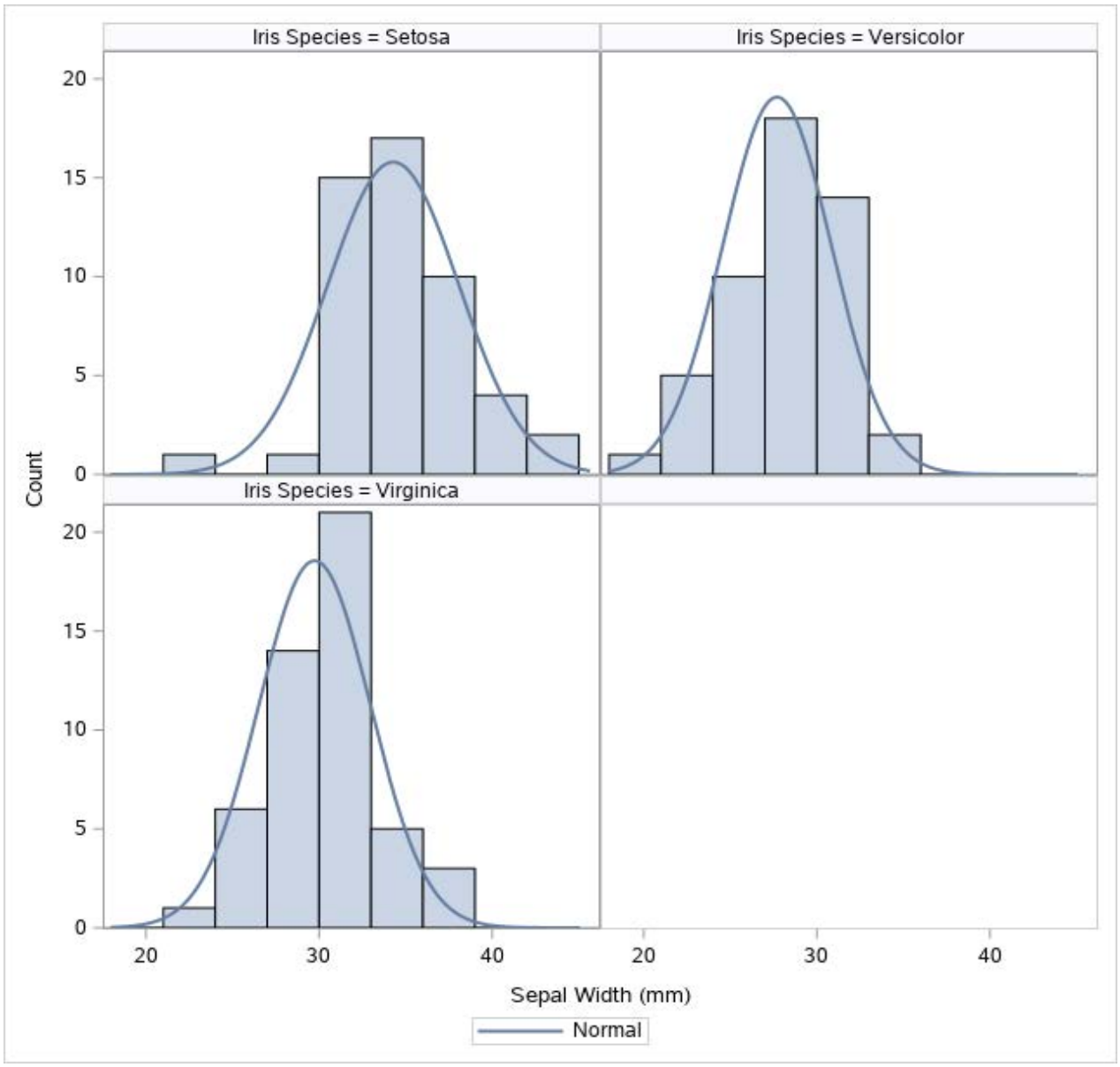

Figure 5: Sepal width frequency distribution using PROC SGPANEL

DEMO 2: COMPARE SALES DATA BY YEAR USING PROC SGPLOT

We will use PROC SGPLOT to create a line graph with 2 y axes to allow us to compare clothing sales and profits by year across quarters from the data in SASHELP.ORSALES.

First we need to aggregate and format our data for how we wish to plot it.

```
proc sql;
   create table orsales_qtr as
   select year
         ,substr(quarter,5,2) as qtr format $2.
         ,sum(profit) as profit format dollar13.
         ,sum(quantity) as quantity format comma15.
   from sashelp.orsales
   group by year,calculated qtr
   order by year,calculated qtr;
quit;
```

Then we run PROC SGPLOT to create a line graph with two y axes to compare sales and profits by year across quarters.

Notes on the code.

- Use one color for each of four years, datacontrastcolors
- Use solid line for quantity, lineattrs=(pattern=solid)
- Use dashed line for profit, lineattrs=(pattern=longdash)
- Group lines by year, group=year
- Use yaxis and y2axis to reference left and right hand vertical axes
- Inset description at bottom of plot, inset 'text' / position=bottom

```
proc sgplot data=orsales_qtr;
  title;
  title1 color=black "Orion Sales 1999 - 2002";
    styleattrs datacontrastcolors=(purple green orange blue);
    xaxis type=discrete label='Quarter';
    yaxis label='Units Sold - Solid Line' grid minor;
    y2axis label='Profit $ - Dashed Line' minor;
    series x=qtr y=quantity / group=year lineattrs=(pattern=solid);
    series x=qtr y=profit / group=year lineattrs=(pattern=longdash) y2axis;
    INSET 'Units Sold and Profit by Quarter'/ POSITION = BOTTOM BORDER
        TEXTATTRS=(Size=11 Weight=Bold);
run;
```

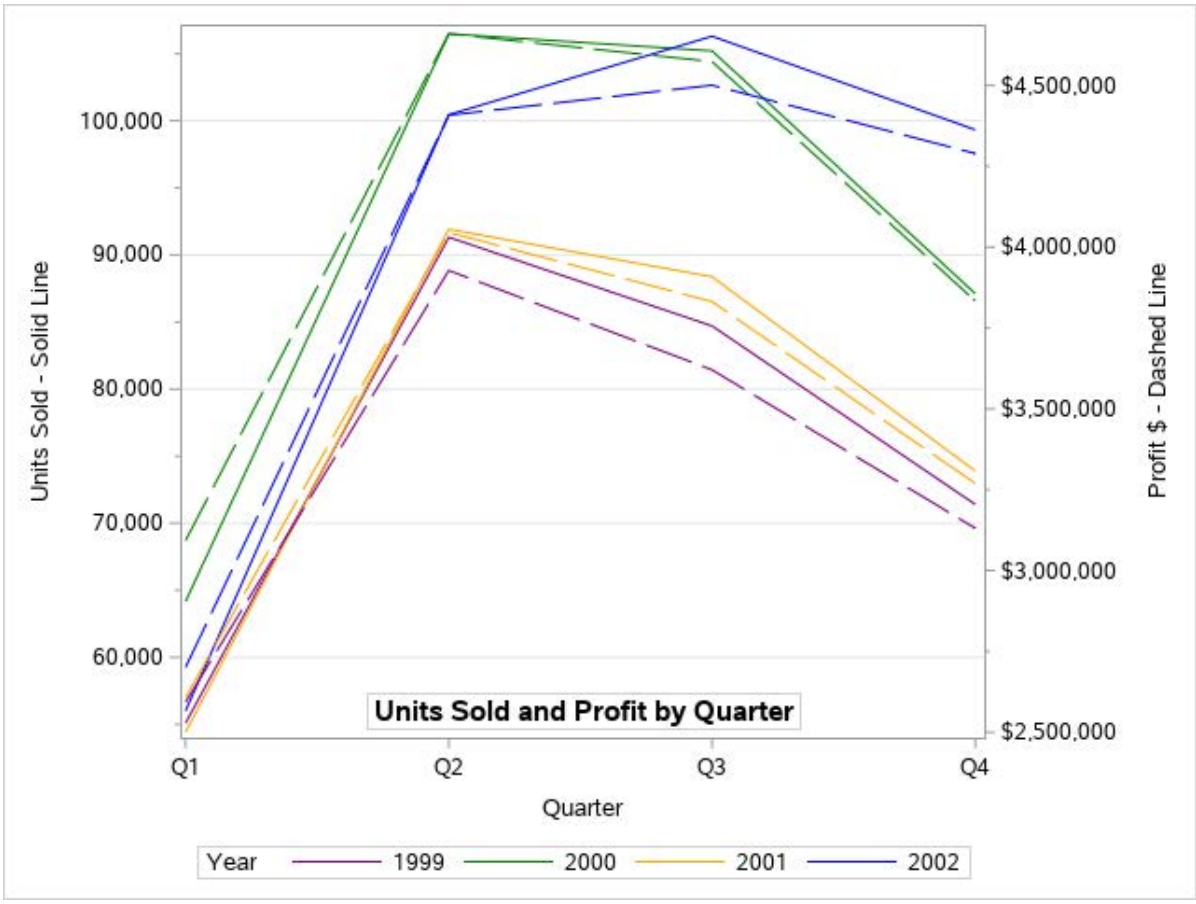

Figure 6. **Orion Sales 1999 – 2002**

As we look at the graph, the data starts to tell us a story.

- There is seasonality each year with lowest sales and profits in Q1 and Q4, and highest sales and profits in Q2 and Q3
- Units sold and profit follow similar trends within each year
- 2002 shows a different pattern to the other years
    - In Q3 2002 we see a separation in sales and profit
    - Q3 2002 has more sales and profits than Q2 2002. For the other years Q2 has more sales and profits than Q3

We should analyze the data further to understand why

- Sales and profits peak in the middle of the year. Don't assume seasonality is a fixture
- Q2 and Q3 2002 trend differently to the other years
- Sales and profits separate for Q3 2002

DEMO 3: PLOTTING IN 3 DIMENSIONS WITH SGRENDER

Example from SAS® 9.4 ODS Graphics: Procedures Guide, Sixth Edition

Create a stat graph template called surface that reads height, weight and density.

```
proc template;
  define statgraph surface;
  begingraph;
     layout overlay3d;
       surfaceplotparm x=height y=weight z=density;
     endlayout;
   endgraph;
  end;
run;
```

SASHELP.GRIDDED contains data on height, weight and density that the surface template converts into a 3-D rendering. We simply call PROC SGRENDER and apply the surface stat graph template to SASHELP.GRIDDED to render the 3-D image.

```
title;
title1 'Height, weight and density plot based on a custom statgraph template';
proc sgrender data=sashelp.gridded template=surface;
run;
title;
title1;
```

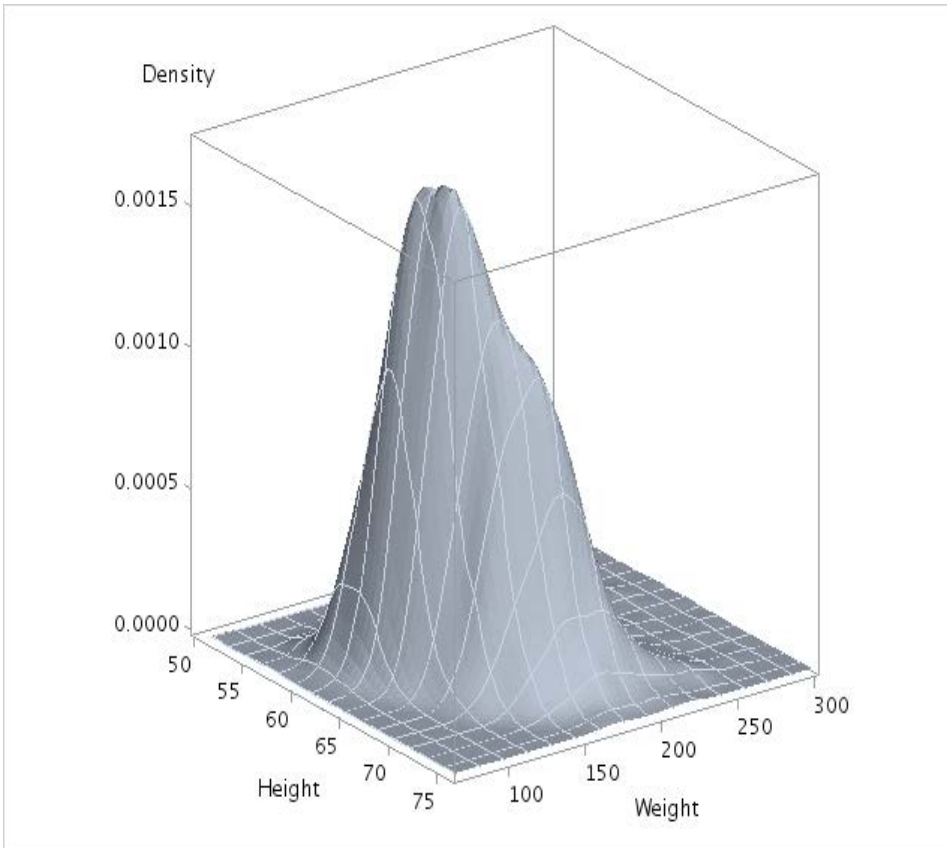

Figure 7. **Height, weight and density plot based on a custom statgraph template**

PROC SGDEMO.SAS

```
/* Determine if iris sepal width is a predictor of iris sepal length */
/* Do this for species Setosa */
data iris;
  set sashelp.iris;
  if species = 'Setosa';
run;

/* Frequency histogram for sepal width */
/* See if assumptions for linear regression are met.  These include */
/*   Normal distribution */
proc sgplot data=iris;
  title "Sepal Width Frequency Distribution";
```

```
  histogram sepalwidth / scale=count;
  density sepalwidth / scale=count;
run;
title;

/* Using macro variable to filter for species */
%let species=Setosa;
proc sgplot data=sashelp.iris(where=(species="&species."));
  title "Sepal Width Frequency Distribution for &species.";
  histogram sepalwidth / scale=count;
  density sepalwidth / scale=count;
run;
title;

/* Use proc sgpanel to display scatter plots by species */
/* Same code as for sgplot but with panelby statement added */
proc sgpanel data=sashelp.iris;
  title "Sepal Width Frequency Distribution";
  panelby species;   /* Create a separate plot for each species */
  histogram sepalwidth / scale=count;
  density sepalwidth / scale=count;
run;
title;

/* Scatter plot for sepal width against Sepal Length with regression line */
/* See if assumptions for linear regression are met.  These include */
/*   Predictor variable (x) has a linear relationship with the response (y) */
/*   Homoscedasticity i.e. equal variance throughout sample */
proc sgscatter data=iris;
   plot sepallength*sepalwidth / reg;
   title "Sgscatter scatter plot of Sepal Width (x) against Sepal Length (y)";
run;

proc sgscatter data=sashelp.iris(where=(species="&species."));
    plot sepallength*sepalwidth / reg;
    title "Sgscatter scatter plot of Sepal Width (x) against Sepal Length (y) for &species.";
```

```
run;

/* Create scatter plot with regression line using sgplot.  */
/* Add confidence limits for the mean CLM and individual   */
/* predicted values CLI         */
%let species=Setosa;
proc sgplot data=sashelp.iris(where=(species="&species."));
 scatter x=sepalwidth y=sepallength;
 title "Sgplot scatter plot of Sepal Width (x) against Sepal Length (y)";
 reg x=sepalwidth y=sepallength /clm cli;
run;

/* Create scatter plots by species using sgpanel */
proc sgpanel data=sashelp.iris;
 panelby species;
 scatter x=sepalwidth y=sepallength;
 title "Scatter plot of Sepal Width (x) against Sepal Length (y)";
 reg x=sepalwidth y=sepallength /clm cli;
run;

/* If assumptions are met, we can perform linear regression */
/* Simple linear regression predicting sepal length from sepal width */
proc reg data=iris outest=est1;
   eq1: model  sepallength=sepalwidth;
run;

proc reg data=sashelp.iris(where=(species="&species.")) outest=est1;
   eq1: model  sepallength=sepalwidth;
run;
/* Display Orion Sales for quantity and profit by quarter and compare each quarter */
/* to the same quarter for 1999 - 2002 on a line chart */
proc sql;
  drop table orsales_qtr;
quit;

/* Prepare data for display on line chart with 2 y axes */
```

```
proc sql;
  create table orsales_qtr as
  select year
        ,substr(quarter,5,2) as qtr format $2.     /* Quarter */
             ,sum(profit) as profit format dollar13.    /* Format profit */
             ,sum(quantity) as quantity format comma15. /* Format quantity */
  from sashelp.orsales
  group by year
         ,calculated qtr
  order by year
         ,calculated qtr;
quit;

/* Create line chart with two y axes          */
/* Use one color for each of four quarters, datacontrastcolors    */
/* Use solid line for quantity, lineattrs=(pattern=solid)         */
/* Use dashed line for profit, lineattrs=(pattern=longdash)       */
/* Group lines by year, group=year                                */
/* Use yaxis and y2axis to reference left and right hand vertical axes  */
/* Inset description at bottom of plot, inset 'text' / position=bottom  */
proc sgplot data=orsales_qtr;
 title;
 title1 color=black "Orion Sales 1999 - 2002";
    styleattrs datacontrastcolors=(purple green orange blue);
    xaxis type=discrete label='Quarter';
    yaxis label='Units Sold - Solid Line' grid minor;
    y2axis label='Profit $ - Dashed Line' minor;
    series x=qtr y=quantity / group=year lineattrs=(pattern=solid);
    series x=qtr y=profit / group=year lineattrs=(pattern=longdash) y2axis;
    INSET 'Units Sold and Profit by Quarter'/ POSITION = BOTTOM BORDER
TEXTATTRS=(Size=11 Weight=Bold);
run;
title;
title1;

/* PROC SGRENDER */
/* Example from Example from SAS® 9.4 ODS Graphics: Procedures Guide, Sixth Edition */
```

```
/* Create stat graph template */
proc template;
  define statgraph surface;
  begingraph;
     layout overlay3d;
      surfaceplotparm x=height y=weight z=density;
     endlayout;
   endgraph;
  end;
run;
/* Generate graphics output from the template */
/* Input dataset contains information on height, weight and density */
title;
title1 'Height, weight and density plot based on a custom statgraph template';
proc sgrender data=sashelp.gridded template=surface;
run;
title;
title1;
```

CONCLUSION

As the examples demonstrate the SG procedures are very useful analysis and data visualization tools.

REFERENCES

Assumptions of Linear Regression – Statistics Solutions https://www.statisticssolutions.com/assumptions-of-linear-regression/

Cano, Gabe. *Convert Your Old Plots and Charts to New SG Plots and Charts: Here's How,* SGF 2012 http://support.sas.com/resources/papers/proceedings12/083-2012.pdf

Iris flower data set https://en.wikipedia.org/wiki/Iris_flower_data_set

SAS® 9.4 ODS Graphics: Procedures Guide, Sixth Edition

https://go.documentation.sas.com/?docsetId=grstatproc&docsetTarget=n0y3i6hxxcrnkmn1mq6zc61bsxrn.htm&docsetVersion=9.4&locale=en

Slaughter, Susan L and Delwiche, Laura D. *Using PROC SGPLOT for Quick High-Quality Graphs,* SGF 2010 http://support.sas.com/resources/papers/proceedings10/154-2010.pdf

Slaughter, Susan L and Delwiche, Laura D. *Graphing Made Easy with SGPLOT and SGPANEL Procedures,* SGF 2015 https://support.sas.com/resources/papers/proceedings15/2441-2015.pdf

CONTACT INFORMATION

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

Alex Chaplin
Bank of America
alex.chaplin@bofa.com