SAS Tutorial: The Scatter Plot

Data Set: anscombe

Tutorial Instructions:

The scatterplot, and variations of the scatterplot, is arguably the most frequently used statistical graphic used in statistical modeling and Exploratory Data Analysis (EDA). In this tutorial we will learn how to produce scatter plots using the SAS procedures PROC SGSCATTER, PROC SGPLOT and PROC CORR. To illustrate these procedures we will use a famous example data set called *Anscombe's Quartet*.

You can read the original paper

> Anscombe (1973), "Graphs in Statistical Analysis", *The American Statistician*, Vol. 27, No. 1, pp. 17-21.

by downloading it from the NU Library. (The paper is also very easily found on the open internet.)

Anscombe's Quartet consists of four pairs of (X, Y) data that were constructed to have the same Pearson correlation coefficient (to three decimal places) despite the fact that they show four different types of statistical relationships between the X and Y pair. Keep in mind that this example was constructed at a time when computing (and graphics) was difficult, and simple linear regression (a regression model with one predictor variable) was more common place, again due to computing restrictions.

You can view the variables listed on the data set by using s PROC CONTENTS statement.

```
proc contents data=mydata.anscombe; run;
```

Part 1: PROC CORR

We will begin by using PROC CORR to produce the Pearson correlation coefficients for each of the X and Y pairs that are present in the anscombe data set. This is done to replicate the relationships present in the Anscombe example pp. 28-30 in *Regression Analysis By Example* (5th Edition).

```
TITLE "PROC CORR Example: Anscombe Quartet";
ods graphics on;
PROC CORR DATA=anscombe PLOT=(matrix);
      VAR X1;
      WITH Y1;
      TITLE2 "X1 and Y1 Correlation";
run;
ods graphics off;
```

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y1	11	7.50091	2.03157	82.51000	4.26000	10.84000
X1	11	9.00000	3.31662	99.00000	4.00000	14.00000

Pearson Correlation Coefficients, N = 11 Prob > |r| under H0: Rho=0	
	X1
Y1	0.81642 0.0022

Table 1: Correlation between X1 and Y1

Running the code provided will produce the output in Table 1. If you modify the code to look at the pairs (X2, Y2), (X3, Y3), and (X4, Y4), then you should get the same estimated value for the Pearson correlation coefficient to three decimal places. The code will also produce a scatterplot of the two variables.

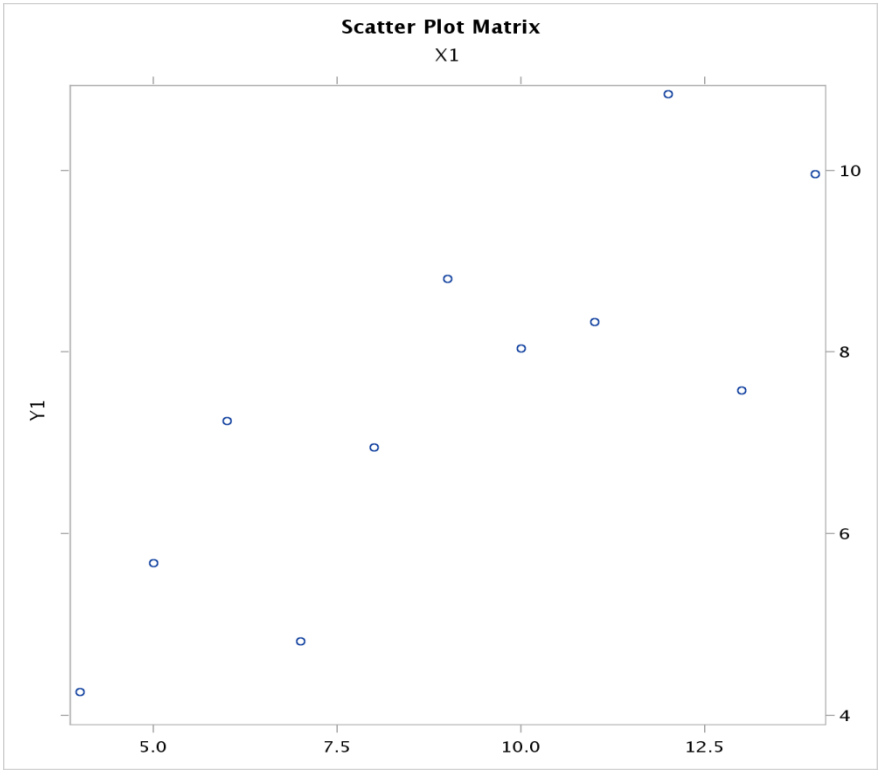


Figure 2: Scatterplot of X1 and Y1 using PROC CORR

The Pearson correlation coefficient is supposed to be a measure of the strength of the linear relationship between X and Y. A value of 0.816 would suggest a very strong linear relationship. Is this true? Do all of our data pairs exhibit a strong linear relationship? Do we see a strong linear relationship in Figure 1? Maybe we should examine some scatterplots in order to discern the relationship between the X and Y variables.

Part 2: PROC SGPLOT

When using PROC CORR, the scatterplot was produced as a plotting option. Now we will use PROC SGPLOT to directly produce a scatter plot with "smoothed curves" defined by a LOESS curve (a scatterplot smoother) and a fitted regression line between the two variables. A deviance between the LOESS curve and the fitted regression line indicates the model mis-fit that would occur by fitting the simple linear regression model.

```
ods graphics on;
PROC SGPLOT DATA=anscombe;
      scatter X=X1 Y=Y1;
      title "X1 and Y1 Scatter Plot - No Smoothers";
run;
ods graphics off;
```

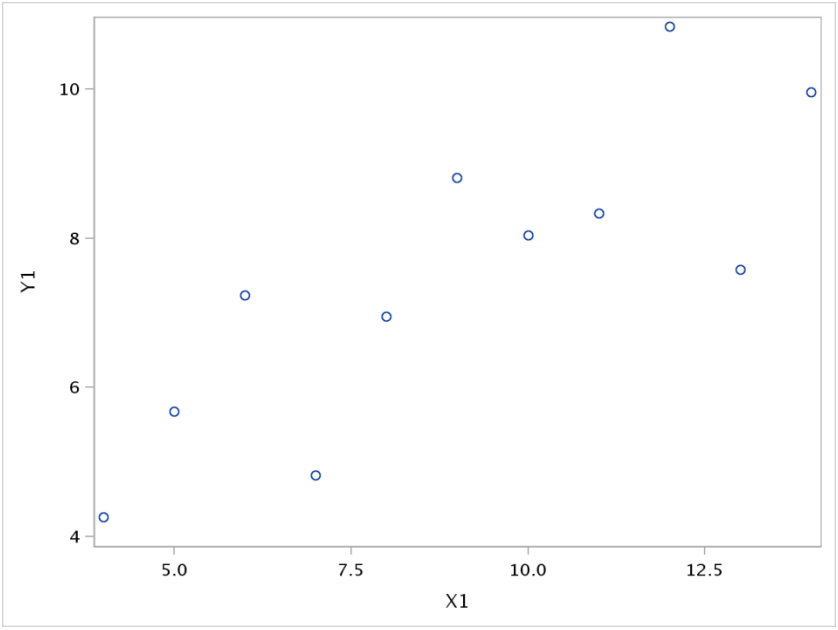


Figure 2: Scatterplot of X1 and Y1 using PROC SGPLOT

```
ods graphics on;
PROC SGPLOT DATA=anscombe;
      LOESS X=X1 Y=Y1 / NOMARKERS;
      REG X=X1 Y=Y1;
      title "X1 and Y1 Scatter Plot with LOESS and Regression";
run;
ods graphics off;
```

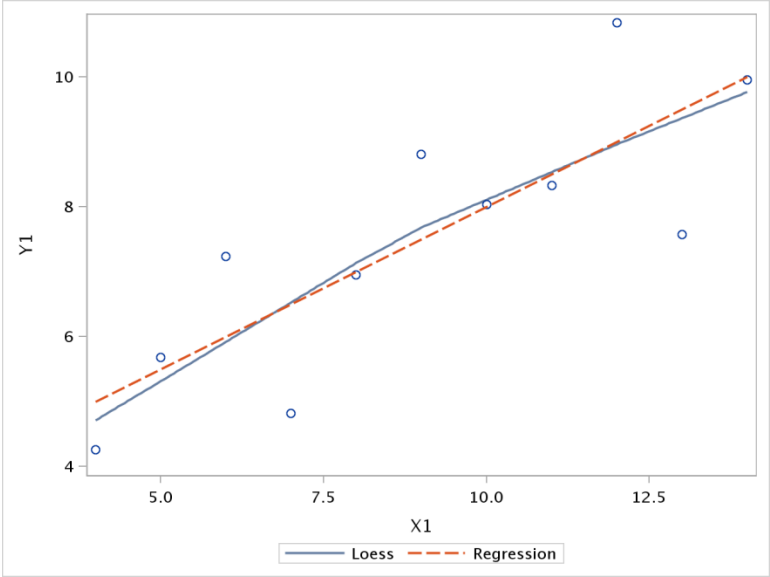


Figure 3: Scatterplot of X1 and Y1 with scatterplot smoothers using PROC SGPLOT

We can see that the overlays of the smoothed curves help us visualize the relationships present in the data that may be hard to visualize in the scatterplot alone. It is a very common practice to include the LOESS curve in a scatterplot. In fact, it is so common that almost all scatterplots should include the LOESS curve in order to enhance the visualization power of the scatterplot.

Part 3: PROC SGSCATTER

A third way to produce a scatterplot in SAS is to use the procedure PROC SGSCATTER. Like PROC SGPLOT, PROC SGSCATTER will allow us to produce a scatterplot with the LOESS and fitted regression line overlays for enhanced visualization.

```
ods graphics on;
PROC SGSCATTER data=anscombe;
      compare X=X1 Y=Y1 / loess reg;
      title "X1 and Y1 Scatter with Loess and Regression";
run;
ods graphics off;
```

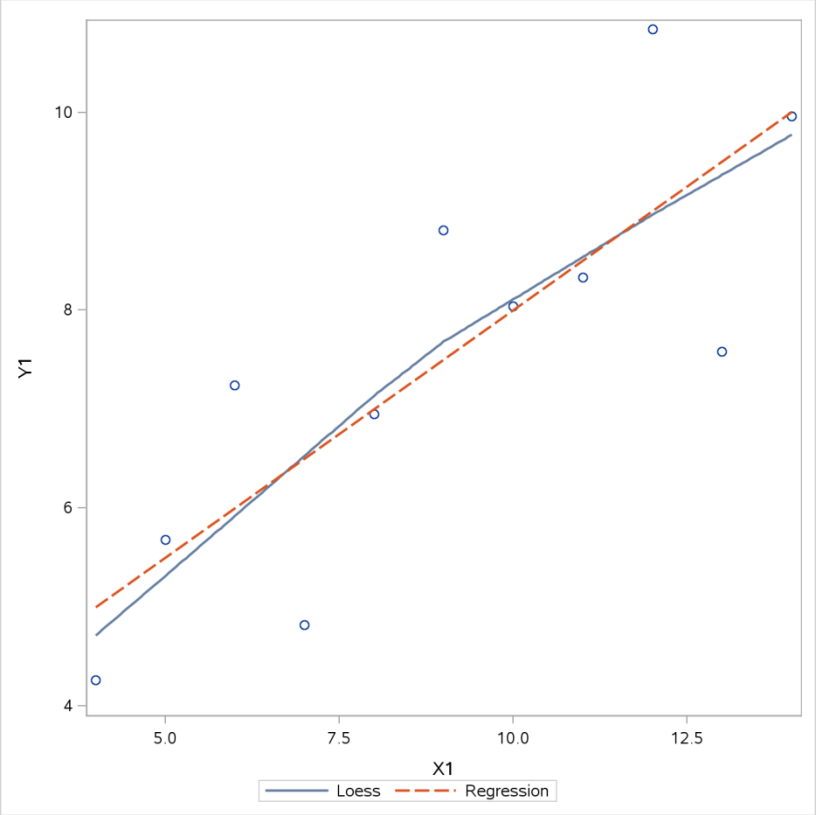


Figure 4: X1 and Y1 Scatter Plot with scatterplot smoothers using PROC SGSCATTER

Notice that in all three of these examples, we are producing scatterplots using a different SAS procedure, and each SAS procedure has a slightly different syntax. In addition each SAS procedure has a different set of options, some of which include producing a different set of plots other than the scatterplot. Sometimes a SAS procedure can have multiple syntaxes to produce the same output. Try this code as an example.

```
ods graphics on;
PROC SGSCATTER data=anscombe;
      plot Y1*X1 / loess reg;
      title "X1 and Y1 Scatter with Loess and Regression";
run;

PROC SGSCATTER data=anscombe;
      compare Y=Y1 X=X1 / loess reg;
      title "X1 and Y1 Scatter with Loess and Regression";
run;
ods graphics off;
```

You should make the scatterplots for all four pairs and make sure that you understand how the Pearson correlation coefficient can be misleading and why statistical graphics are important.

Part 4: PROC SGPANEL

When you need to produce a "family" of plots, it is convenient to display the plots together as a single unit called a *panel*. SAS will allow you to create a panel of plots by using PROC SGPANEL. However, in order to use the procedure PROC SGPANEL we will need to convert our dataset from "wide" format to "long" format.

Here is how we can convert our dataset from wide format to long format.

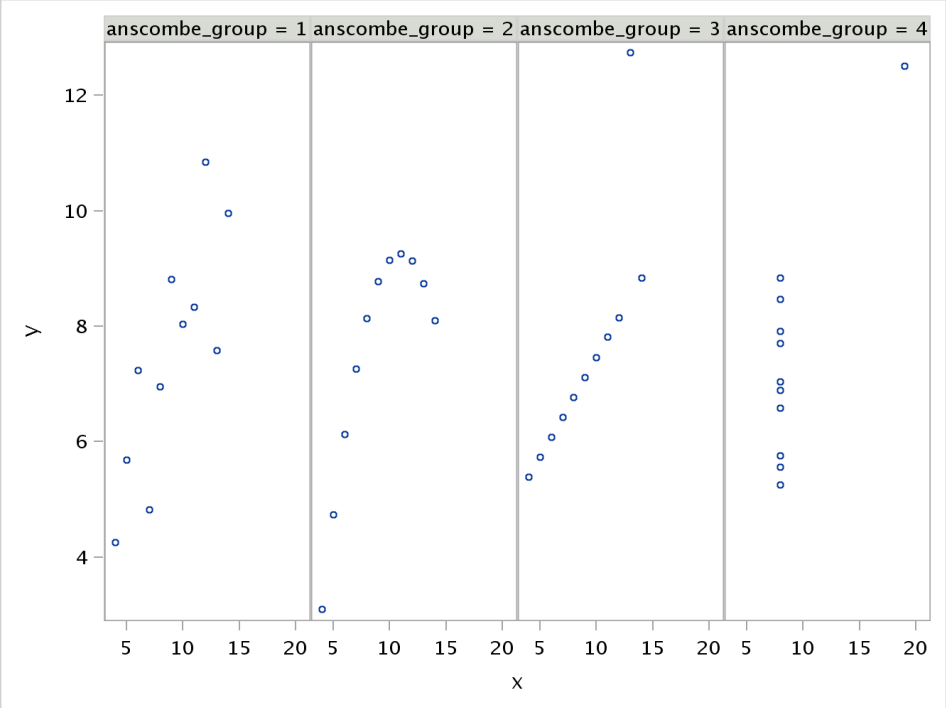
```
PROC PRINT data=mydata.anscombe; RUN; QUIT;

DATA long;
      set mydata.anscombe (keep=x1 y1 rename=(x1=x y1=y) in=a)
      mydata.anscombe (keep=x2 y2 rename=(x2=x y2=y) in=b)
      mydata.anscombe (keep=x3 y3 rename=(x3=x y3=y) in=c)
      mydata.anscombe (keep=x4 y4 rename=(x4=x y4=y) in=d)
      ;

      if (a=1) then anscombe_group=1;
      else if (b=1) then anscombe_group=2;
      else if (c=1) then anscombe_group=3;
      else if (d=1) then anscombe_group=4;
      else anscombe_group=0;
RUN;

PROC PRINT data=long; RUN; QUIT;
```

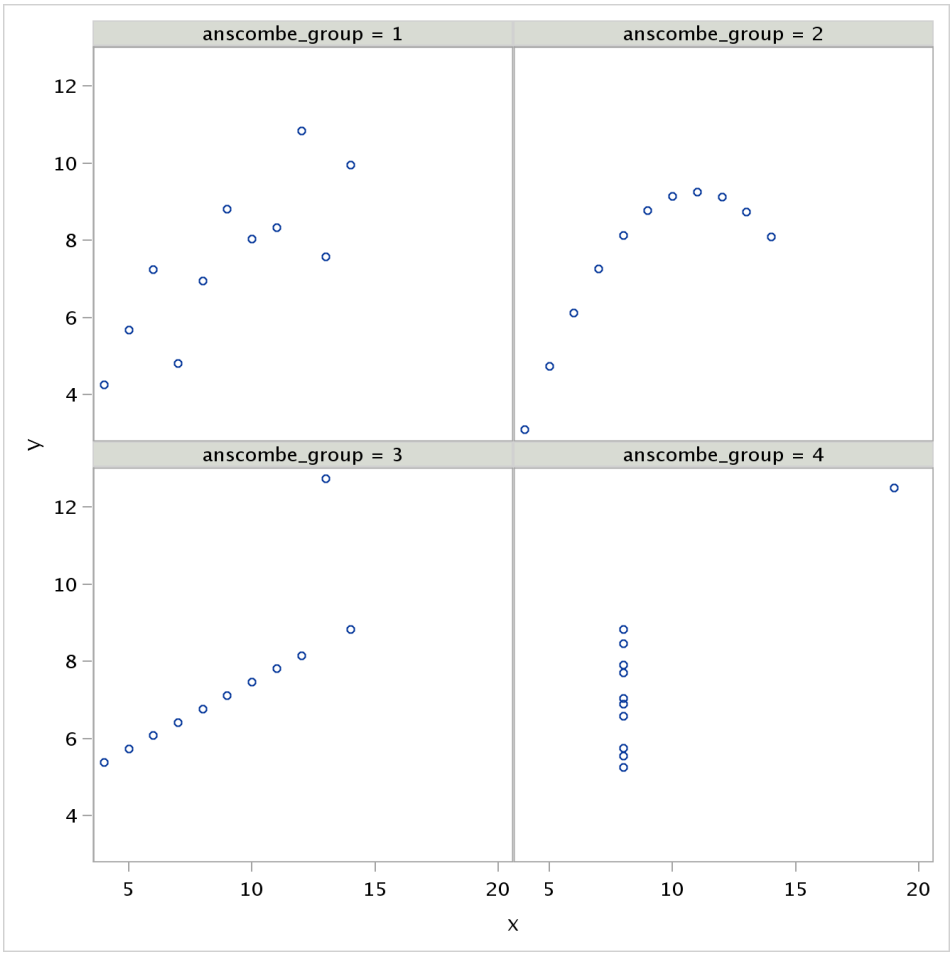
The PROC PRINT statements can be used to see the data sets as they are stored. The original dataset is in wide format because it has many columns with the "type" or "group" denoted by the column name. The long dataset is in long format because it has two columns for the values and a third column to denote the "type" or "group". Some SAS procedures will require wide format, some will require long format, and some can use either.

Here is an example of how we would panel the Anscombe scatterplots.

```
ods graphics on;
PROC SGPANEL data=long;
      PANELBY anscombe_group / COLUMNS=4 ROWS=1;
      SCATTER X=x Y=y;
      TITLE 'Panel of Scatterplots';
RUN; QUIT;
ods graphics off;
```

```
ods graphics on;
PROC SGPANEL data=long;
      PANELBY anscombe_group / COLUMNS=2 ROWS=2;
      SCATTER X=x Y=y;
      TITLE 'Panel of Scatterplots';
RUN; QUIT;
ods graphics off;
```




anscombe_group = 1
anscombe_group = 2
anscombe_group = 3
anscombe_group = 4
y
x
12
10
8
6
4
5
10
15
20