

Anscombe's Quartet

[SAS 9.2] I discovered these data sets during a presentation by Michael Blastland at the SAS Professionals Convention at SAS UK in Marlow in July 2011. Anscombe's quartet comprises four data sets that have identical simple statistical properties. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician F J Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties. The data sets can be read into SAS data sets using the following SAS code:

```
DATA anscombe;
  INFILE DATALINES DLM=',' DSD TRUNCOVER;
  LENGTH dset $3;
  INPUT x y @; dset='I'; OUTPUT;
  INPUT x y @; dset='II'; OUTPUT;
  INPUT x y @; dset='III'; OUTPUT;
  INPUT x y; dset='IV'; OUTPUT;
  DATALINES;
10.0,8.04,10.0,9.14,10.0,7.46,8.0,6.58
8.0,6.95,8.0,8.14,8.0,6.77,8.0,5.76
13.0,7.58,13.0,8.74,13.0,12.74,8.0,7.71
9.0,8.81,9.0,8.77,9.0,7.11,8.0,8.84
11.0,8.33,11.0,9.26,11.0,7.81,8.0,8.47
14.0,9.96,14.0,8.10,14.0,8.84,8.0,7.04
6.0,7.24,6.0,6.13,6.0,6.08,8.0,5.25
4.0,4.26,4.0,3.10,4.0,5.39,19.0,12.50
12.0,10.84,12.0,9.13,12.0,8.15,8.0,5.56
7.0,4.82,7.0,7.26,7.0,6.42,8.0,7.91
5.0,5.68,5.0,4.74,5.0,5.73,8.0,6.89
RUN;
PROC SORT DATA=anscombe; BY dset x; RUN;
```

The data sets have essentially the same statistical analysis:

```
dset=I
Number of Observations Read      11
      Sum of      Mean
Source    DF  Squares  Square  F Value Pr > F
Model      1   27.51000 27.51000 17.99   0.0022
Error      9   13.76269  1.52919
Corrected
  Total    10   41.27269
Root MSE=1.23660  R-Square=0.6665
Dependent Mean=7.50091  Adj R-Sq=0.6295
Coeff Var=16.48605

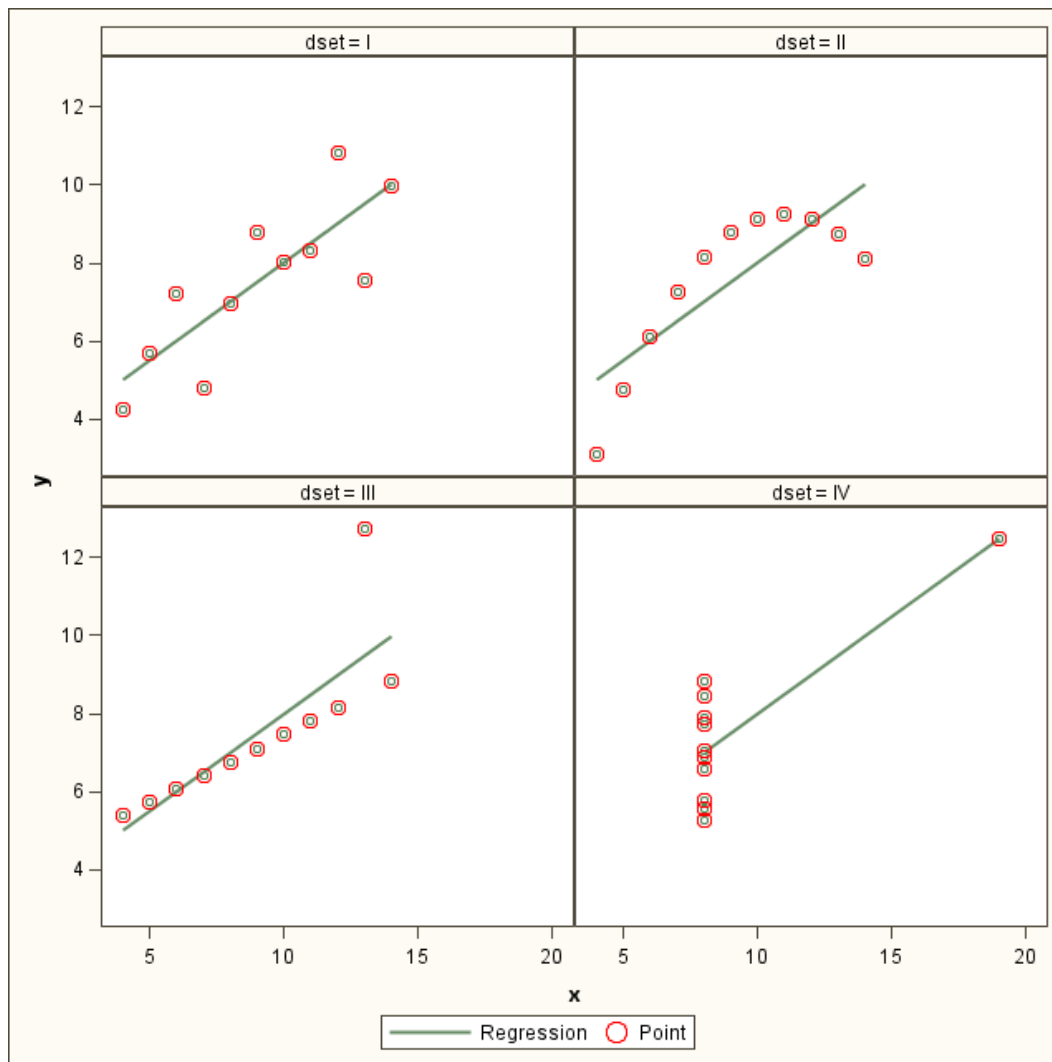
      Parameter Standard t
Variable DF  Estimate  Error  Value Pr > |t|
Intercept 1    3.00009    1.12475  2.67  0.0257
x          1    0.50009    0.11791  4.24  0.0022
```

The statistical analysis above was generated using the following SAS code:

```
PROC REG DATA=anscombe RSQUARE; BY dset; MODEL y=x; RUN;
```

The following SGPanel code generated the graphs, which plainly show that, although the statistical analysis of the 4 data sets is almost identical, the graphs of the data sets are very different!

```
PROC SGPanel DATA=anscombe;
  PANELBY dset;
  REG X=x Y=y / LEGENDLABEL="Regression";
  SCATTER X=x Y=y / LEGENDLABEL="Point" MARKERATTRS=(COLOR=RED SIZE=10PT);
RUN;
```



Anscombe's quartet comprises four datasets that have identical simple statistical properties, yet appear very different when graphed. **Never accept numerical statistics alone, but request a graph too!**

Philip R Holland, Holland Numerics Ltd

*Data taken from the Wikipedia article about Anscombe's Quartet at en.wikipedia.org/wiki/Anscombe's_quartet
This article was first published in Issue 54 of VIEWS News at www.sascommunity.org/wiki/VIEWS_News_backissues*