



SAS/GRAPH and regression

by Katrien Declercq

Academic

All sorts of nice graphs can be made using SAS/GRAPH, but did you know that you can have a regression line adjusted to your data by choosing the right symbol definition? The following code illustrates some of the possibilities using the SASHELP.CLASS dataset where a regression is done of the weight of the children on their height.

```
ods html path='c:\temp' body="regression.html";
options reset=all gunit=pct ftext=swiss htext=3 target=winprtc rotate=landscape
        device=java;

axis1 label=('Height');
axis2 label=(angle=90 'Weight');

symbol1 v=dot i=r1 width=2 cv=black ci=red;
symbol2 v=dot i=r1 clm90 width=2 cv=black ci=red co=green;
symbol3 v=dot i=r1 cli90 width=2 cv=black ci=red co=green;
symbol4 v=dot i=r10 width=2 cv=black ci=red;
symbol5 v=dot i=rq width=2 cv=black ci=red;

title1 h=4 'Regression of Weight on Height (SASHELP.CLASS)';
proc plot data=sashelp.class;

    title2 'Simple Linear Regression';
    plot weight*height=1/haxis=axis1 vaxis=axis2 regeqn;
    run;

    title2 'Simple Linear Regression with 90% Confidence Limits for Mean Predicted
    Values';
    plot weight*height=2/haxis=axis1 vaxis=axis2 regeqn;
    run;

    title2 'Simple Linear Regression with 90% Confidence Limits for Individual
    Predicted Values';
    plot weight*height=3/haxis=axis1 vaxis=axis2 regeqn;
    run;

    title2 'Simple Linear regression through the origin';
    plot weight*height=4/haxis=axis1 vaxis=axis2 regeqn hzero vzero;
    run;

    title2 'Quadratic Regression';
    plot weight*height=5/haxis=axis1 vaxis=axis2 regeqn;
    run;

quit;
ods html close;
```

In the first plot, the first symbol definition is used to fit a simple linear regression line to the data and the option REGEQN serves to print the regression equation on the graph (by default it only appears in the log). In the second plot, the same regression is fitted as in the first plot, but the 90% confidence limits for mean predicted values are superimposed as are the 90% confidence limits for individual predicted values in the third plot. The latter are much wider since for the prediction of individual observations, not only the variability in the estimated parameters is taken into account, but also the variability of the individual observations about the regression line. In the fourth plot, the regression line is forced through the origin and the options HZERO and VZERO force the horizontal and vertical axes to start at zero. Clearly, this regression doesn't fit these data very well, but in some situations, you may need to force the regression line through the origin. In the fifth plot, a quadratic regression line is fitted to the data, meaning that a second order term of the independent variable (height) is added to the model. In order to know whether this quadratic regression fits the data significantly better than the simple linear regression line, a test can be done to test the hypothesis that the coefficient of the squared independent variable is equal to zero. This can be read immediately from the output from PROC REG. Also PROC REG should be used in all cases to find out more details on the regression model (goodness of fit, hypothesis tests for the individual parameter estimates) (cfr. Newsletter 15).

The INTERPOL option in the SYMBOL statement is to be used to fit a regression line on a graph of two variables with as value R<type><0><CLM | CLI<50...99>, where type can be L for linear regression, Q for quadratic regression and C for cubic regression. The 0 can be used to force the regression line through the origin, i.e. to fit a regression equation without intercept. CLM or CLI may be used to add the confidence limits for mean predicted values or the individual predictions respectively. Confidence levels from 50 to 99 can be specified, with a default of 95. The options CV, CI and CO are used to specify the colors for the symbols, the regression line and the confidence limits.