

Plotting Confidence Intervals & Prediction Bands with SAS/GRAPH

Huixing Tang, David Swift, and Kerril Bauerly
The Psychological Corporation

This poster presents a method for generating a regression plot using SAS/GRAPH shading (SAS/GRAPH Software, Volumes 1 and 2, Version 6 (SAS Institute, 1990). The information displayed on the graph includes the following:

- Predicted value
- Observed mean
- Observed individual value
- Confidence interval
- Prediction band

A step by step description of the procedure will be provided using a simulated data set of science and mathematics test scores. For this illustration, scores on the mathematics test, MATH, are used as the dependent variable, and the scores on the science test, SCIENCE, are used as the independent or predictor variable in a simple linear regression analysis. Both variables are stored in the data set INPUTDAT. The resulting regression plot will show how well science performance predicts math performance at both the group mean and the individual score level.

Step 1. Running regression analysis with PROC REG

PROC REG is run on the data to generate: (1) a predicted value, (2) an observed value, (3) the upper and lower bound for the confidence interval, and (4) the upper and the lower bound for the prediction band. These values are saved in an output data set with an OUTPUT statement. The code for this step is as follows:

```
proc reg noprint data=inputdat;  
    model math=science;  
    output out=regout
```

```

l95=low95
l95m=low95m
u95=up95
u95m=up95m
p=predict;

```

Step 2. Computing the observed mean at each level of the predictor variable

The observed individual MATH scores in the REGOUT data set are used to compute the observed mean at each predictor value. The observed means are then merged with the variables in the REGOUT data set. This is done in the following code:

```

proc sort data=regout;
    by science;

proc means noprint;
    var math;
    by science;
    output out=meanmath mean=meanmath;

data plotdata;
    merge regout meanmath;
    by science;

```

Step 3. Generating the plot with PROC GPLOT

PROC GPLOT is used to generate the regression plot using the variables in PLOTDATA.. The code is given as follows:

```

options source2;

goptions device=gddm382l gddmnickname=gddmgray
    gddmtoken=img240x ftext=centb ctext=black cback=white
    gunit=percent colors=(black blue red pink green cyan yellow
    dab orange purple dag libg gold gray) htitle=3 htext=4;

```

```

axis1 label=(h=3.5 angle=-90 rotate=90 'Math')
      order=( 5 to 80 by 5) value=( h=3);

axis2 label=(h=3.5 'Science') order=(0 to 60 by 5)
      value=(h=3) offset=(3);

symbol1 c=white i=sm50 line=1 width=1;
symbol2 c=gray i=sm50 line=1 width=20;
symbol3 c=gold i=sm50 line=1 width=20;
symbol4 c=white i=sm50 line=1 width=1;
symbol5 c=black i=sm50 line=1 width=1;
symbol6 c=black i=none value=dot height=1.8;
symbol7 c=black i=none value=square height=.7;

pattern1 c=white value=solid;
pattern2 c=gray value=solid;
pattern3 c=gold value=solid;
pattern4 c=gray value=solid;

legend1 cborder=white origin=(25, ) label=none mode=protect
      shape=symbol(3,2) value=(h=2.25 '
      'Prediction band'
      'Confidence interval'
      'Predicted' 'Observed mean' 'Individual')
      across=1 frame position=(top left inside) offset=(-14,+3);

proc gplot data=plotdata;
  plot low95*scrs low95m*scrs up95m*scrs up95*scrs
  predict*scrs meanma*scrs mars*scrs
  /areas=4 overlay vaxis=axis1 haxis=axis2 legend=legend1;
  title1 "Predicting Math Performance From Science
  Performance";
run;
quit;

```

In the above code, the four PATTERN statements correspond to the four background areas specified in the AREAS option of the PLOT statement. The four background areas are shaded based on the colors specified in the PATTERN statements. The boundaries of the shading are

determined by the lines drawn from the first four variable combinations in the PLOT statement. The colors selected in the SYMBOL statements were selected to match the shading of their respective areas. The general color scheme used in the PATTERN and SYMBOL statements were selected to fit the gray scale color scheme of the host system. Notice the WIDTH option on the SYMBOL2 and SYMBOL3 statements were increased to 20. This was done to enlarge the width of the legend symbols for the prediction and confidence intervals and does not affect the overall graph.

References

SAS Institute, Inc. (1990). SAS/GRAPH Software, Version 6, Volumes 1 and 2. Cary, NC: Author.

Predicting Math Performance from Science Performance

