**Producing and Interpreting Residuals Plots in SAS[[1]](#footnote-1)©**

 In a linear regression analysis it is assumed that the distribution of residuals is, in the population, normal at every level of predicted Y and constant in variance across levels of predicted Y. I shall illustrate how to check that assumption. Although I shall use a bivariate regression, the same technique would work for a multiple regression.

 Download **Residual-Plots.sas** and **Residual-Plots-Output.pdf** from my [SAS Programs page](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Programs.htm). Residual-Plots-Output.doc has the output with unessential parts trimmed out and with the most important parts highlighted. I recommending printing the “Producing and Interpreting Residuals Plots in SAS” document and bringing the Residual-Plots-Output.doc up in Word. That way you can see the annotated output on the screen while reading this document.

 Notice that variables X and Y are not skewed – I generated them with a normal random number generator. Notice that X2 and Y2 are skewed and that taking the square root of Y2 reduces its skewness greatly.

**proc** **reg** data=skew lineprinter; model Y=X;

output out=XY r=Y\_Resid; plot r.\*p.;

title 'Predicting Y from X.'; **run**;

**proc** **univariate** normal plot data=XY; var Y\_Resid;

 Here we predict Y from X and create a data set, XY, with the residuals, . The plot statement ask for a plot of the residuals versus the predicted scores. Look at the plot. On the residuals plots I have highlighted the regression line, . SAS has plotted numbers to indicate how many scores there are at each point in the plot. Notice that at each value of predicted Y the residuals are most dense near the regression line and become less dense as you move up or down away from the regression line, as would be expected if they were normally distributed at each level of predicted Y. Also note that the spread of the residuals does not change much with level of predicted Y, that is, there is homoscedasticity.

 The output from proc univariate shows that the marginal distribution of the residuals is normal.

 The next invocation of proc reg predicts Y from the skewed X2. Notice that the residuals plot of the proc univariate output reveal no problems. The skewness of X2 may be troublesome for the correlation model, but not for the regression model.

 The next invocation of proc reg predicts skewed Y2 from X. Notice that the residuals plots shows the residuals not to be normally distributed – they are pulled out (skewed) towards the top of the plot. Proc univariate also show that the marginal distribution of the residuals is skewed and the tests of normality reject the null hypothesis that these residuals came from a normally distributed population. I don’t generally pay much attention to these tests of normality, because with a large sample they will find a significant difference even when the deviation from normality is minor.

 The next invocation of proc reg predicts the square root of Y from X. Notice that the transformation did wonders, reducing the skewness of the residuals to a comfortable level.

 The next portion of the program reads a new data set, which is contained within the program. We predict new Y from new X. Look at the residuals plot. Oh my. Notice that the residuals are not symmetrically distributed about zero. They are mostly positive with low and high values of predicted Y and mostly negative with medium values of Y. If you were to find the means of the residuals at each level of Y and connect those means with the line you would get a curve with one bend. This strongly suggests that the relationship between X and Y is not linear and you should try a nonlinear model. Notice that the problem is not apparent when we look at the marginal distribution of the residuals.

**data** quadratic; set linear; X\_SQ=X\*X;

**proc** **reg** lineprinter; model Y=X X\_SQ; plot r.\*p. ;

plot Y\*X='.' p.\*X='X' / overlay;

 Here we fit the data with a polynomial regression, quadratic. The residuals plot looks much better. In the second plot we have the original data plotted with dots and the regression line plotted with X’s – SAS plots a question mark when one dot is overlaid on another, so treat the questions marks as part of the regression line too. I connected the predicted scores with lines within Word.

 In the final section of the program we read in another new data set and predict Y from X. The residuals plot shows clear evidence of heteroscedasticity. In this case, the error in predicted Y increases as the value of predicted Y increases. I have been told that transforming one the variables sometimes reduces heteroscedasticity, but in my experience it often does not help.

[Producing and Interpreting Residuals Plots With SPSS](http://core.ecu.edu/psyc/wuenschk/SPSS/Residual-Plots-SPSS.doc)

[Return to Wuensch's SAS Lessons Page](http://core.ecu.edu/psyc/wuenschk/SAS/SAS-Lessons.htm)

Copyright 2011, Karl L. Wuensch - All rights reserved.

1. © Copyright 2011, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)