Subject: Re: regress
Posted by Brian Larsen on Wed, 18 Mar 2009 13:07:08 GMT
View Forum Message <> Reply to Message

> Step one should be to plot your two variables against each other. If
> you've got two many data points, you might need to look at a 2-D
> histogram, instead. Only use REGRESS if it seems plausible from such a
> plot that there is a linear relationship between them. If it looks like
> there's some other relationship between them, then you should fit to a
> curve that more closely resembles that relationship.

I point this out to everyone as much as to the original poster.
Plotting your data really is the key to initial understanding. To
make "looking" at plots more quantitative NIST has a nice statistics
guide that people should know about. See:
http://www.itl.nist.gov/div898/handbook/index.htm

I have written routines for many of the plots shown in the guide. Two
are 4-plot and 6-plot. See:
http://people.bu.edu/balarsen/IDLdoc/stats/fourplot.html
http://people.bu.edu/balarsen/IDLdoc/stats/sixplot.html
and the NIST guide for a description of what the plots show.

For example the 4-plot is testing these underlying assumptions about
regression:

1. Fixed Location:
   If the fixed location assumption holds, then the run sequence
plot will be flat and non-drifting.

2. Fixed Variation:
   If the fixed variation assumption holds, then the vertical
spread in the run sequence plot will be the approximately the same
over the entire horizontal axis.

3. Randomness:
   If the randomness assumption holds, then the lag plot will be
structureless and random.

4. Fixed Distribution:
   If the fixed distribution assumption holds, in particular if the
fixed normal distribution holds, then
    1. the histogram will be bell-shaped, and
    2. the normal probability plot will be linear.

The "scary" thing is that if any of these assumptions are violated in
a meaningful way then regression is invalid to use on a data set, but
it s done anyway....