

Re: Fitting: unbalanced data point density and weights

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2008-05/msg00260.html>

- *From:* Richard Ulrich <Rich.Ulrich@xxxxxxxxxxxx>
 - *Date:* Wed, 14 May 2008 17:37:15 -0400
-

On Wed, 14 May 2008 08:28:27 -0400, Paul Rubin <rubin@xxxxxxx> wrote:

Kristof Lebecki wrote:

Hello everybody,

I have a set of points $\{x,y\}$. When I plot them in a log-log scale they apparently agree with a power dependence function $F(x) = a \cdot x^p$. So, I am fitting these points with this function.

(Details: Mathematica package, function NonlinearRegress, its description for the help file: finds a least-squares fit to a list of data for a model that is a not linear combination of the given basis functions; the coefficients of the linear combination are the parameters of the fit).

The problem is that my data are evidently bad balanced: I have much more points on the "left" side (small x value), their errors are also much smaller (I use the errors to define weights, different for every point). On the other side, the x -density of the "right" points is much smaller. The points form a slight bow in the log-log scale – as the result, the fitted function matches "perfectly" left points while the other points are badly matched.

The question is: what to do to keep the data points appropriately balanced? A friend of mine saw once an article about this subject, but you know: "gone with the wind". ;)

Any advice? Should I present you a sketch of the problem?

(I hope this group fits to that subject).

If you regress $\log y$ on $\log x$ (linear regression), do you still see the

Re: Fitting: unbalanced data point density and weights

same phenomenon?

You could perhaps use weighted least squares, assigning greater weight to the points with larger x values. Another kludge would be to replicate the points with larger x (just put extra copies of them in the sample), until the sample was more balanced.

The underlying reason here is that the analysis is one that "minimizes the residuals" — in some sense. In Least Squares, the residuals are squared and summed, and the effect *can* be much the same in Maximum Likelihood. Or, the proper ML model may provide "weights" that are more appropriate than the squares of the actual residuals.

Paul gives a few approaches to re-emphasizing certain residuals — "weighting" is common; extra copies of cases is (as he says) a "kludge" to approximate a weighting function.

How important is it that you keep your original metric? In my biostatistical applications, I would be able to assume that I can take the logs of the variables — or some other transformation — and achieve my proper weighting. And analyze the new variables.

I don't think I would advocate either of those approaches, though, at least not at first. If the log-log linear regression again shows a good fit for small x and a poor fit for large x, I would take that to mean that the regression function is not really a power function, and I would try something else — perhaps a sum of two power functions with different exponents?

One of the important guides to a model is "what makes sense for the variables". That depends especially on the variables, and somewhat on the conventions of the field that uses the variables. If you mention what you have, someone here might be familiar with it.

—

Rich Ulrich

<http://www.pitt.edu/~wpilib/index.html>

.