



## OPINION

# Linearization Plots

## Time for Progress in Regression

by Martin L. Lobemeier

*Posted March 3, 2000 · Issue 73*

---

### *Abstract*

Linear regression used to be the norm for statistical analyses. In the age of computers and statistical software packages, researchers should consider using nonlinear regression, an often more appropriate analysis. In either case, testing the chosen model is a necessary step in good statistical practice.

---

In the life sciences, linearization often is used to simplify the analysis of quantitative data that could be represented more accurately by nonlinear mathematical models. Scientists have traditionally preferred linear regression methods to nonlinear regression methods because of their inherent simplicity. For instance, hermodynamic and kinetic studies of enzyme activity, receptor-ligand interactions, and ion channels historically have depended on linear transformations such as the Lineweaver-Burk plot [1] and the Scatchard plot [2]. However, linearization methods can generate systematic error; therefore, researchers must choose a regression technique appropriate to the data set, whether linear or nonlinear. In this Opinion, I will discuss choosing a regression technique and the drawbacks of linear and nonlinear methods.

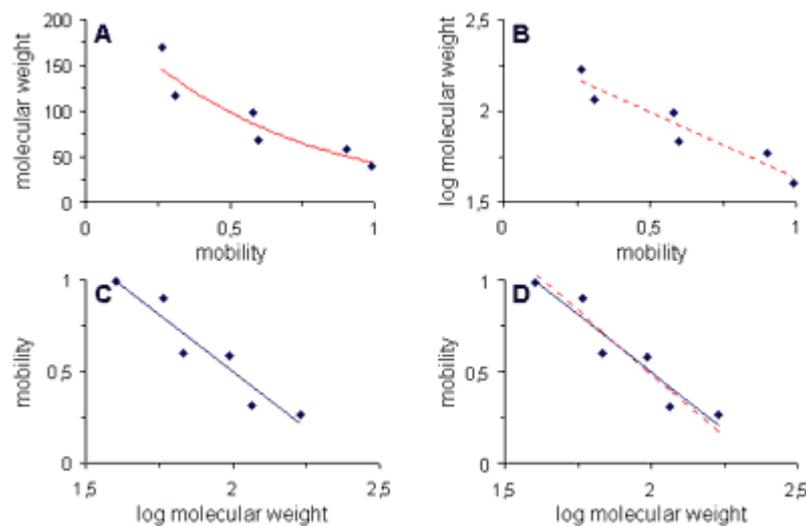
Linear regression techniques require the dependent variable to be plotted against an independent or experimental variable. A common source of error is in declaring variables. Several assumptions are made in the process:

1. The independent variable is without error or contains negligible error only.
2. For each value of the independent variable there is a subpopulation of dependent variable values that is distributed normally.
3. The variances of all dependent variable subpopulations are equally distributed.
4. The means of the subpopulations of the dependent variable all lie on the same straight line (the assumption of linearity).

Despite the assumption of negligible error, it is inherent in the nature of experimental systems that error will arise in the measurement of variables. Hence, it is important that the measured variables be treated as the dependent ones. In linear regression analysis, the distance of the data points to the regression line is minimized in the direction of the y-axis; therefore it is essential that the experimental error be distributed in the same direction.

A pertinent example is the estimation of the molecular weight of proteins in analytical sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), as carried out by Weber and Osborn [3] (figure 1). From the graphs shown in figures 1B and 1C, it appears to be a straightforward calculation for the molecular weight of unknown peptides as compared to standards using linear regression. However, if you compare the resulting regression lines from both plots without changing the data set, as shown in figure 1D, you see that the lines differ, so will the calculated molecular weights of the investigated peptides. More often [4,5] than not [6] the regression line is incorrectly plotted, as shown in figure figure 1B.

Figure 1



Molecular weight calibration of SDS-PAGE gels.

- A.** Plot of molecular weight against mobility as described by Weber and Osborn.
- B.** The same plot, but molecular weight substituted with the logarithm of the molecular weight to give a straight line.
- C.** Modified Weber and Osborn plot with mobility plotted against log molecular weight.
- D.** Both regression lines drawn into one graph with the dashed red line as  $x = f(y)$  and the solid blue line as  $y = f(x)$ .

When using the plot of Weber and Osborn for the molecular weight calibration of gels - i.e., plotting the molecular weight onto the y-axis and the mobility onto the x-axis - the molecular weight is handled as if it is dependent. This does not make much sense for a calibration. The mobility is the variable that is being measured; therefore it contains the experimental error. Thus the graph in figure 1C is the only appropriate analysis method to use for calibrating SDS-PAGE gels.

In many cases, the data may look better than in this example. Even in good runs, the difference between the regression lines will be at least 2%, and perhaps as high as 10% (in bad runs for low and high molecular weight peptides within the same data set). Peptides in the middle range are not strongly affected by the analysis method (compare figure 1D). One could argue, so what? The experimental error lies within the same range as the model-dependent difference. But the decision of which model to use defines the theory behind the experiment.

The model of Weber and Osborn implies that the proteins change their molecular weight during electrophoretic separation, while the value of their electrophoretic mobility is static, in which case a molecular weight calibration would not be possible. The modified model, which treats the mobility as the dependent variable, accounts for almost all sources of error (such as the accuracy of measurement, whether scanning digitally or visually with a ruler).

Presenting the data correctly is important. Probably no one would decide to plot substrate concentration against enzyme activity. The correct way is the reverse - enzyme activity against substrate concentration (figure 2A). The data points for this graph were calculated for a rectangular hyperbola that describes a Michaelis-Menten kinetic reaction or a simple receptor-ligand model. A constant error was added (blue line) and subtracted (red line) from each data point and all values were transformed into the linearization plot of Lineweaver-Burk (figure 2B) and the Scatchard plot (figure 2C). The blue and the red lines, which were parallel to the calculated hyperbola in the direct plot, are now distorted along the x-axis. Another characteristic is that the error is no longer symmetrically distributed with respect to the calculated hyperbola.

The Scatchard plot shows the most severe distortion, because both axes contain experimental error. As stated before, one of the most fundamental rules for linear regression is that a dependent variable is plotted against an independent variable. The Scatchard plot sets the dependent variable on both axes, i.e., both axes are interdependent. This invalidates the fundamental rule of regression, and the linear regression technique should not be applied [7].

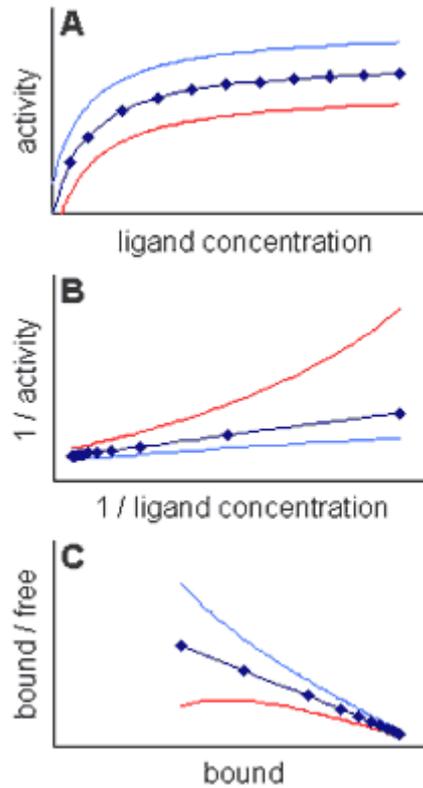
Nonetheless, at least when teaching students, both plots are often used as tools to visualize data, because it is easy to understand the data when presented as a straight line. Especially for the Scatchard plot, the model fits only in rare cases [8] and the disadvantages mentioned earlier far exceed the accuracy of parameters.

Nowadays, an appropriate way to evaluate enzyme kinetics or receptor-ligand interactions is to use a software package that offers nonlinear regression for mathematical analysis. There are some expensive programs available, but they often lack a proper graphical output. There is also a huge supply of free software [9,10] that, in combination with other standard documentary applications, offers a good alternative for a low budget [11].

The advantage of nonlinear regression is that a mathematical model is directly fitted to the untransformed data. The disadvantage is that the investigator has to carefully decide which model to apply and has to ascertain that the chosen model best fits or describes the experiment. The next example shows how this can be achieved.

Since its invention in 1951, the Lowry method for estimating protein determination [12] has led to different approaches in the analysis of the standard curve. In the original paper the authors state that the curvature of the data is the most difficult problem with this method; the detailed reaction mechanism remains unknown [13].

Figure 2



Error distribution of linearization plots.

**A.** Direct plot.

**B.** Lineweaver-Burk plot.

**C.** Scatchard plot.

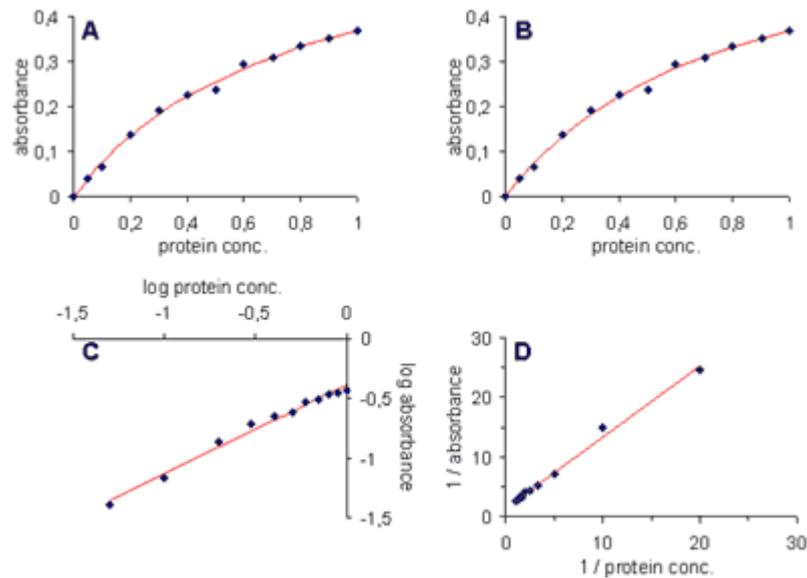
However, some authors have made approaches to linearize the standard curve in order to calculate the protein content of sample preparations rather than relying on a subjective eyeball fit. A log-log plot [14] and a double reciprocal plot [15] have been proposed to transform the curve into a straight line. Recent studies have shown that the data can be fitted to a polynomial [16] or a rectangular hyperbola [17], in a direct plot of absorbance vs. protein concentration.

Which of these methods is best? There are disagreements. The problem seems related more to the experimental procedure than to the analysis method. Researchers use different modified Lowry assays because a standardized method is lacking. The only way to find out which mathematical model is suitable for the analysis of a particular protein determination technique is to apply the different approaches to one dataset and compare the fits with each other.

Figure 3 shows data from a modified Lowry assay [18] with a protein standard solution of bovine serum albumin (BSA), ranging from 0 to 1 mg/ml protein. The task was to find a regression line that fit the data over this wide range. The 4th order polynomial fit of figure 3A was chosen because it projected a curve better than a 2nd order polynomial proposed by Harrington [16]. In general polynomials are not suitable for the calculation of a standard curve

because they possess local minima and maxima and might "nestle" to outliers and lead to the calculation of data far from the "true" values.

Figure 3



Analysis of a standard curve for the determination of BSA protein concentration.

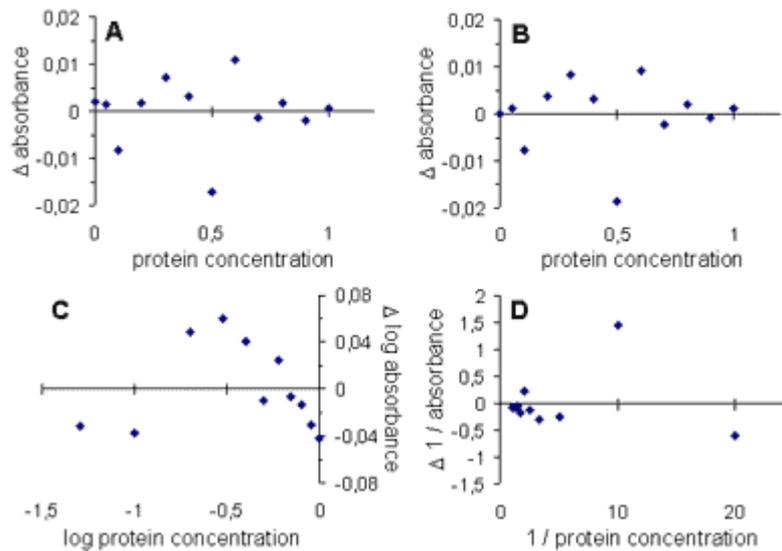
- A.** 4th order polynomial fit.
- B.** Data fitted to a rectangular hyperbola.
- C.** Stauffer's log-log plot.
- D.** Double reciprocal plot.

Polynomials are often used when a data set seems to follow a mathematical model but the formula is missing. Polynomials design smooth curves that betray the spectator's eye, as if a correlation is being described; therefore, they should not be used. I present them only in order to show all known analysis methods for the Lowry protein determination.

At first glance, the graphs do not show much difference between the fits. Indeed, all fits have an R<sup>2</sup> value greater than 0.99. But one should take a closer look and analyze the distribution of the residuals, i.e., the distance of the measured data points to the calculated best fit (the deviation or residual plot). A mathematical model that sufficiently describes a data set would show equally and normally distributed residuals along the x-axis.

From the graphs in figure 4 it is easy to conclude that the log-log plot (figure 3C) and the double reciprocal plot (figure 3D) represent the Lowry method only poorly, because the residuals are clearly not equally distributed and tend to build curves or clusters. The hyperbolic and the polynomial fit, on the other hand, show normally distributed residuals. However, only the rectangular hyperbola proves to be a good model for the evaluation of a standard curve of the Lowry protein determination.

Figure 4



Deviation plot of the analysis methods for the Lowry method. The graphs show the residuals of the data points to the best fit lines of figure 3.

A. Deviation plot of 4th order polynomial fit.

B. Deviation plot of rectangular hyperbola fit.

C. Deviation plot of Stauffer's log-log plot.

D. Deviation plot of double reciprocal plot.

Whichever regression technique is applied, it is essential to test the model. Doing so provides a reasonable explanation for the choice of method that best describes a reaction or experiment.

Since nonlinear regression is difficult to perform without computers, it was essential in the past to use linearization plots to evaluate experimental results. But with nonlinear regression programs available now, linearization plots are no longer an acceptable technique, even in order to simplify concepts for scientists in training.

*Martin L. Lobemeier is finishing his Ph.D. thesis about the LDL receptor in carp (Cyprinus carpio).*

*Andrzej Krauze is an illustrator, poster maker, cartoonist, and painter who illustrates regularly for HMS Beagle, The Guardian, The Sunday Telegraph, Bookseller, and New Statesman.*

## References

1. Lineweaver, H. and Burk, D. 1934. The determination of enzyme dissociation constants. *J. Am. Chem. Soc.* 56:658-666.
2. Scatchard, G. 1949. The attractions of proteins for small molecules and ions. *Ann. N.Y. Acad. Sci.* 51:660-672.

3. Weber, K. and Osborn, M. 1969. The reliability of molecular weight determinations by dodecyl sulfate-polyacrylamide gel electrophoresis. *J. Biol. Chem.* 244:4406-4412.
4. Stryer, L. 1988. *Biochemistry*. New York, W.H. Freeman.
5. Westermeier, R. 1990. *Elektrophorese-Praktikum*. Weinheim, VCH Verlagsgesellschaft.
6. Scopes, R.K. 1987. *Protein Purification: Principles and Practice*. 2nd edition. New York, Springer.
7. Kermode, J.C. 1989. The curvilinear Scatchard plot: Experimental artifact or receptor heterogeneity? *Biochem. Pharmacol.* 38:2053-2060.
8. Klotz, I.M. 1982. Numbers of receptor sites from Scatchard graphs: Facts and fantasies. *Science*. 217:1247-1249.
9. Duggleby, R.G. 1981. A nonlinear regression program for small computers. *Anal Biochem.* 110:9-18.
10. Ireland, C.R. and Long, S.P. 1984. *Microcomputers in Biology: A Practical Approach*. Oxford, IRL Press.
11. Lobemeier, M.L. 1999. A non-linear regression program for the evaluation of ligand-receptor interactions in the Microsoft Excel environment. *In: Elsner N. and Eysel U. (eds.) Göttingen Neurobiology Report 1999. Vol. II. 1999. New York, Georg Thieme Verlag.*
12. Lowry, O.J. et al. 1951. Protein measurement with the Folin phenol reagent. *J. Biol. Chem.* 193:265-275.
13. Legler, G., et al. 1985. On the chemical basis of the Lowry protein determination. *Anal. Biochem.* 150:278-287.
14. Stauffer, C.E. 1975. A linear standard curve for the Folin Lowry determination of protein. *Anal. Biochem.* 69:646-648.
15. Campbell, H.D. 1983. The Lowry Protein assay: Linearization of standard curve by double reciprocal plot. *Anal. Letters.* 16:1495-1507.
16. Harrington, C.R. 1990. Lowry protein assay containing sodium dodecyl sulfate in microtiter plates for protein determinations on fractions from brain tissue. *Anal. Biochem.* 186:285-287.
17. Grant, K.I. 1990. LOWRY: A simple BASIC program for the analysis of protein assays over a wide range of concentrations. *Comput. Meth. Progr. Biomed.* 31:237-242.
18. Wang, C.S. and Smith, R.L. 1975. Lowry determination in the presence of triton X-100. *Anal. Biochem.* 63:414-417.

### ***Endlinks***

GraphPad Guide to Nonlinear Regression - an introductory text from the makers of GraphPad software, written for scientists, not statisticians. Also includes links to related sites. (<http://www.graphpad.com/www/nonling1.html>)

Multiple Regression and Nonlinear Estimation - two chapters from StatSoft's Electronic Textbook offer a more technical and equation-oriented introduction. StatSoft is the developer of STATISTICA. (<http://www.statsoft.com/textbook/stmulreg.html>), (<http://www.statsoft.com/textbook/stnonlin.html>)

Statistics and Statistical Graphics Resources - an extensive collection of annotated links including many to software-related sites. (<http://www.math.yorku.ca/SCS/StatResource.html>)

Interactive Statistical Calculation Pages - a collection of more than 550 links to Web pages that perform statistical calculations, including regression analyses. (<http://members.aol.com/johnp71/javastat.html>)

Related HMS *Beagle* articles:

- Statistically Significant Sites - a review of statistics-related Web resources for biologists. (<http://news.bmn.com/hmsbeagle/68/reviews/insitu>)
- The Crack Baby Epidemic That Wasn't - discusses what statistics mean and don't mean. ([http://news.bmn.com/hmsbeagle/50/people/op\\_ed.html](http://news.bmn.com/hmsbeagle/50/people/op_ed.html))
- GraphPad Prism - a recent review of this statistical software package. (<http://news.bmn.com/hmsbeagle/54/reviews/sreview>)