

The GraphPad
Guide to
Nonlinear
Regression

Dr. Harvey Motulsky
President, Graphpad Software
April 1996

This booklet was published by GraphPad Software Inc., the creators of GraphPad Prism® – your complete solution for scientific graphics, curve fitting and statistics.

Copyright © 1995-96 by GraphPad Software, Inc. All rights reserved.

You may obtain additional copies of this booklet by downloading the document from the internet at <http://www.graphpad.com>. You may also order additional copies from GraphPad Software.

Companion booklets on radioligand binding and statistical comparisons may be available by the time you read this.

GraphPad Prism is a registered trademark of GraphPad Software, Inc.

To contact GraphPad Software:

Phone	800-388-4723 (US only) or 619-457-3909
Fax	(USA) 619-457-8141
Email	support@graphpad.com or sales@graphpad.com
World Wide Web	http://www.graphpad.com
Mail	GraphPad Software, Inc. 10855 Sorrento Valley Road #203 San Diego, CA 92121 USA

- INTRODUCTION TO NONLINEAR REGRESSION 3
- Why you should use nonlinear regression 3
 - Linear regression of transformed data is less accurate 3
 - Don't relegate scientific decisions to a computer program 4
 - The results of polynomial regression are often impossible to interpret scientifically 4
 - Cubic spline is not a data analysis method 4
- Terminology 4
- How nonlinear regression works 5
 - Comparison of linear and nonlinear regression 5
 - Iterations in nonlinear regression 5
- Decisions to make when using nonlinear regression 5
 - Choose a model 5
 - Prepare data for nonlinear regression 5
 - Estimate initial values 6
 - Constants 6
 - Weighting 6
 - Average replicates? 7
- INTERPRETING NONLINEAR REGRESSION RESULTS 8
- Assumptions of nonlinear regression 8
- Variables, standard errors, and confidence intervals 8
- Sum-of-squares, $s_{y,x}$, and R^2 8
- Residuals and the runs test 9
- How to tell if the nonlinear regression fit is any good 9
 - Did the fit converge on a solution? 9
 - Does the curve come close to the points? 10
 - Are the results scientifically plausible? 10
 - Do the data systematically deviate from the curve? 10
 - Are the confidence intervals wide? 10
 - Is the fit a local minimum? 11
- What to do when the fit is no good? 12
- Comparing two equations 13
 - How the F test works 13
 - Example 14
- Comparing fits to two data sets 14
 - 1. Compare the results of repeated experiments. 14
 - 2. Compare the results within one experiment. Simple approach. 15
 - 3. Compare the results within one experiment. More complicated approach. 15
 - Advantages and disadvantages of the three methods 16
- GRAPHPAD PRISM 17

Introduction to nonlinear regression

Nonlinear regression is a powerful tool for analyzing scientific data, especially in pharmacology and physiology. The goal of nonlinear regression is to fit a model to your data. The program finds the best-fit values of the variables in the model (perhaps rate constants, affinities, receptor number, etc.) which you can interpret scientifically. In most cases, the primary goal is to obtain those values and a secondary goal is to draw a graph of the fit curve.

In some situations, your only goal is to draw a curve. You don't care about models or equations, and don't want to obtain best-fit values. You just want a smooth curve through your points either for artistic reasons or to use as a standard curve. You may still use nonlinear regression in these situations, or you may use these alternatives:

- Polynomial regression.
- Cubic spline or LOWESS curve.
- A program that fits your data to thousands of equations and picks the best.

This chapter (and the next two) assumes that your goal is primarily to obtain the best-fit values of the variables – to fit a model to your data.

Why you should use nonlinear regression

Linear regression of transformed data is less accurate

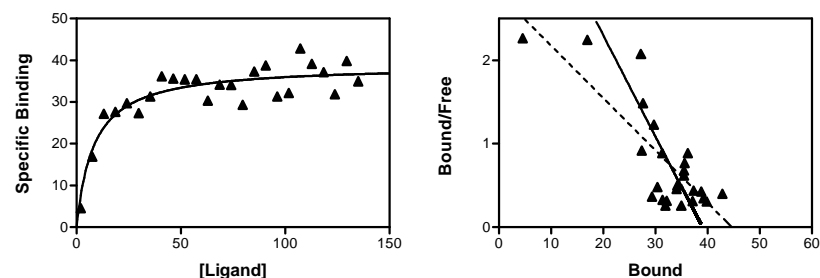
Before the age of microcomputers, nonlinear regression was not readily available to most scientists. Instead, scientists transformed their data to make a linear graph, and then analyzed the transformed data with linear regression. Examples include Lineweaver-Burke plots of enzyme kinetic data, Scatchard plots of binding data, and logarithmic plots of kinetic data.

These methods are outdated, and should not be used to analyze data. The problem is that the linear transformation distorts the experimental error. Linear regression assumes that the scatter of points around the line follows a Gaussian distribution, and that the standard deviation is the same at

every value of X. These assumptions are usually not true with the transformed data. A second problem is that some transformations alter the relationship between X and Y. For example, in a Scatchard plot the value of X (bound) is used to calculate Y (bound/free), and this violates the assumptions of linear regression.

Since the assumptions of linear regression are violated, the results of linear regression are incorrect. The values derived from the slope and intercept of the regression line are not the most accurate determinations of the variables in the model. Considering all the time and effort you put into collecting data, you want to use the best possible analysis technique. Nonlinear regression produces the most accurate results.

This figure shows the problem of transforming data. The left panel shows data that follows a rectangular hyperbola (binding isotherm). The right panel is a Scatchard plot of the same data. The solid curve on the left was determined by nonlinear regression. The solid line on the right shows how that same curve would look after a Scatchard transformation. The dotted line shows the linear regression fit of the transformed data. The transformation amplified and distorted the scatter, and thus the linear regression fit does not yield the most accurate values for B_{max} and K_d .



Transformations can be very useful when used appropriately. When analyzing data, follow these rules:

- You should transform your data when the transformation makes the variability more consistent and more Gaussian.
- You should not transform data when the transformation makes the variability less consistent and less Gaussian.

- You should not perform transforms (such as the Scatchard transform) that destroy the relationship between X and Y.
- You should not transform the data merely to make it linear. Since nonlinear regression is easy, there is no reason to force your data into a linear form.

Although it is usually inappropriate to analyze transformed data, it is often helpful to display data after a linear transform. Many people find it easier to visually interpret transformed data. This makes sense because the human eye and brain evolved to detect edges (lines) — not to detect rectangular hyperbolas or exponential decay curves. Even if you analyze your data with nonlinear regression, it may make sense to display transformed data.

Don't relegate scientific decisions to a computer program

The goal of nonlinear regression is to fit a model to your data. The program finds the best-fit values of the variables in the model (perhaps rate constants, affinities, receptor number, etc.) which you can interpret scientifically. Choosing a model is a scientific decision. You should base your choice on your understanding of chemistry or physiology (or genetics, etc.). The choice should not be based solely on the shape of the graph.

Some programs automatically fit data to hundreds or thousands of equations and then present you with the equation(s) that fit the data best. Using such a program is appealing because it frees you from the need to choose an equation. The problem is that the program has no understanding of the scientific context of your experiment. The equations that fit the data best are unlikely to correspond to scientifically meaningful models. You will not be able to interpret the best-fit values of the variables, and the results are unlikely to be useful for data analysis.

This kind of approach is very useful in three situations:

- Your only goal is to plot an attractive curve.
- You wish to create a standard curve for interpolating unknown values.
- You need an equation to use within a computer simulation.

In all three situations, it doesn't matter whether the equation corresponds to a biological, chemical or physical model. What matters is that the equation accurately predict Y from X within the range of your data.

This approach can be useful in some situations. Don't use it when the goal of curve fitting is to fit the data to a model based on chemical, physical, or biological principles. Don't use a computer program to avoid making a scientific decision.

The results of polynomial regression are often impossible to interpret scientifically

Beware of the term "curve fitting". The term is often used to refer not to nonlinear regression, but rather to polynomial regression. This method fits data to a polynomial equation: $Y = A + BX + CX^2 + DX^3 \dots$. Programmers prefer polynomial regression, because it is so much easier to program. That's why it is built in to so many spreadsheet and graphics programs. But few biological or chemical models are described by polynomial equations, so polynomial regression is of limited usefulness to scientists.

Cubic spline is not a data analysis method

Cubic spline curves are smooth curves that go through every data point. In some cases, a cubic spline curve can look attractive on a graph and work well as a standard curve for interpolation. The curve does not correspond to any equation (or rather the equation differs for every pair of points) so cubic spline is not useful in data analysis.

Terminology

A model is a formal presentation of a chemical or physiological idea. To be useful for nonlinear regression, the model must be expressed as an equation that defines Y, the outcome you measure, as a function of X and one or more variables that you want to fit. We use the term variable to refer to the terms in the equation you want to fit. In the context of nonlinear regression, the term variable does not refer to X and Y. Some programs and books use the word parameters rather than variables.

How nonlinear regression works

Comparison of linear and nonlinear regression

A line is described by a simple equation that calculates Y from X, slope and intercept. The purpose of linear regression is to find values for the slope and intercept that define the line that comes closest to the data. More precisely, it finds the line that minimizes the sum of the square of the vertical distances of the points from the line.

The goal of minimizing the sum-of-squares in linear regression can be achieved quite simply. A bit of algebra (shown in many statistics books) derives equations that define the slope and intercept. Put the data in, and the answers come out. There is no chance for ambiguity.

Nonlinear regression is more general. It can fit data to any equation that defines Y as a function of X and one or more variables. It finds the values of those variables that generate the curve that comes closest to the data. More precisely, the goal is to minimize the sum of the squares of the vertical distances of the points from the curve.

Except for a few special cases, it is not possible to directly solve the equation to find the values of the variables that minimize the sum-of-squares. Instead nonlinear regression requires an iterative approach.

Iterations in nonlinear regression

Here are the steps that every nonlinear regression program follows:

1. Start with an initial estimated value for each variable in the equation.
2. Generate the curve defined by the initial values. Calculate the sum-of-squares (the sum of the squares of the vertical distances of the points from the curve).
3. Adjust the variables to make the curve come closer to the data points. There are several algorithms for adjusting the variables. The most commonly used method was derived by Levenberg and Marquardt (often called simply the Marquardt method).
4. Adjust the variables again so that the curve comes even closer to the points.

5. Keep adjusting the variables until the adjustments make virtually no difference in the sum-of-squares.
6. Report the best-fit results. The precise values you obtain will depend in part on the initial values chosen in step 1 and the stopping criteria of step 5. This means that repeat analyses of the same data will not always give exactly the same results.

Decisions to make when using nonlinear regression

When you use a program for nonlinear regression, you must make the following decisions.

Choose a model

To use nonlinear regression, you must first define a mathematical model based on theory. The first step is to choose a model. For example, many kinds of binding data are explained by the law of mass action. The next step is to express the model as an equation defines Y as a function of X and one or more variables. Some programs also let you define the model as a differential equation that defines dY/dX as a function of one or more variables.

Choosing a model is a scientific decision, not a statistical one. The model needs to make sense in scientific terms.

You may also fit two different models to your data, and then use statistical methods (F test) to compare them. See "Comparing two equations" on page 13.

Prepare data for nonlinear regression

When preparing data for nonlinear regression, keep these points in mind:

- It matters which variable is X and which is Y. X should be the variable you control or manipulate. Y is the variable you measure. Nonlinear regression finds the curve that lets you best predict Y from X.
- Use reasonable units. In pure mathematics, it doesn't matter whether you express your results as 1 picomolar or 10^{-12} molar, as

1 nanovolt or 10^{-9} volts. When computers do the calculating, however, it can matter. Calculation problems such as round off errors are far more likely when the values are very high or very low. We recommend that you scale your data to avoid values less than 10^{-4} or greater than 10^4 .

- Don't smooth. You lose information when you smooth data, and this won't get you a better fit.
- If you are fitting data to a sigmoidal dose-response or competitive binding curve, enter the X values as the logarithm of concentration, rather than the concentration itself.

Estimate initial values

Nonlinear regression is an iterative procedure. The program must start with estimated values for each variable that are in the right "ball park" — say within a factor of five of the actual value. It then adjusts these initial values to improve the fit. It then adjusts the values again and again until the improvement is tiny.

Some programs automatically provide initial values automatically. With other programs, you must enter the values manually. If you have "clean" data that clearly define a curve, then it usually doesn't matter if the initial values are fairly far from the correct values. You'll get the same answer no matter what initial values you use, unless the initial values are very far from correct.

Initial values matter more when your data have a lot of scatter, don't span a large enough range of X values to define a full curve, or don't really fit the model. In these cases, you may get different answers depending on which initial values you use. See "Is the fit a local minimum?" on page 11.

You'll find it easy to estimate initial values if you have looked at a graph of the data, and understand the model and what all the variables mean. Remember, you just need an estimate. It doesn't have to be very accurate. If you are having problems estimating an initial value:

- Check that you have chosen a model that makes scientific sense.

- Make sure you understand what each variable in the equation means.
- Put away your data, and spend an hour or two generating curves using the model. Change the variables one at a time, and see how they influence the shape of the curve.

Constants

You don't have to fit every variable in the equation. In many situations it makes sense to fix some of the variables to constant values. For example, you might want to define the bottom plateau of a dose-response curve or an exponential decay curve to equal zero.

Weighting

In general, the goal of nonlinear regression is to find the values of the variables in the model that make the curve come as close as possible to the points. Usually this is done by minimizing the sum of the squares of the vertical distances of the data points from the curve. This is appropriate when you expect that the scatter of points around the curve is Gaussian and unrelated to the Y values of the points. (Note to those who have studied advanced statistics: If those assumptions are true, minimizing the sum-of-squares is equivalent to finding the maximum likelihood estimate of the variables).

With many experimental protocols, you don't expect the experimental scatter to be the same, on average, for all points. Instead, you expect the experimental scatter to be a constant percentage of the Y value. If this is the case, points with high Y values will have more scatter than points with low Y values. When the program minimizes the sum of squares, points with high Y values will have a larger influence while points with smaller Y values will be relatively ignored. You can get around this problem by minimizing the sum of the square of the relative distances. This procedure is termed weighting the values by $1/Y^2$. Because it prevents large points from being over-weighted, the term unweighting seems more intuitive.

It is also possible to weight the data in other ways. The goal, always, is to end up with a measure of goodness-of-fit that weights all the data points equally.

Average replicates?

If you collected replicate Y values at every value of X, there are two ways to analyze the data:

- Treats each replicate as a separate point.
- Average the replicate Y values, and treat the mean as a single point.

Deciding which approach to use can be difficult.

The advantage of the first approach is that you have more data points and thus more degrees of freedom. However, you should only use that approach when the experimental error of each replicate is no more closely related to the other replicates than to other data points. Here are two examples where you should analyze each replicate:

- You are doing a radioligand binding experiment. All the data were obtained from one tissue preparation and each replicate was determined from a separate incubation (separate test tube). The sources of experimental error are the same for each tube. If one value happens to be a bit high, there is no reason to expect the other replicates to be high as well.
- You are doing an electrophysiology study. You apply a voltage across a cell membrane and measure conductance. Each data point was obtained from a separate cell. The possible sources of experimental error are independent for each cell. If one cell happens to have a high conductance, there is no reason to expect the replicate cells (those that you apply the same voltage to) to also have high conductance.

You should not treat each replicate as a separate point when the experimental error of the replicates are related. You should average the replicates instead, and analyze the averages. Here are two examples where you should average the replicates:

- The experiment was only performed with a single replicate at each value of X, and you measure radioactivity as Y. Each tube is counted three times, and the three counts are treated as replicates. Any experimental error while conducting the experiment would appear in all the replicates. The replicates are not independent.

- The experiment is a dose-response curve. At each dose, you use a different animal but measure the response three times. The three measurements are not independent. If an animal happens to respond more than the others, that will affect all the replicates. The replicates are not independent.

Interpreting results

Assumptions of nonlinear regression

The results of nonlinear regression are meaningful only if these assumptions are true (or nearly true):

- The model is correct. Nonlinear regression adjusts the variables in the equation you chose to minimize the sum-of-squares. It does not attempt to find a better equation.
- The variability of values around the curve follow a Gaussian distribution. Even though no biological variable follows a Gaussian distribution exactly, it is sufficient that the variation be approximately Gaussian.
- The SD of the variability is the same everywhere, regardless of the value of X. The assumption is termed homoscedasticity. If the SD is not constant but rather is proportional to the value of Y, you should weight the data to minimize the sum-of-squares of the relative distances.
- The model assumes that you know X exactly. This is rarely the case, but it is sufficient to assume that any imprecision in measuring X is very small compared to the variability in Y.
- The errors are independent. The deviation of each value from the curve should be random, and should not be correlated with the deviation of the previous or next point. If there is any carryover from one sample to the next, this assumption will be violated.

Variables, standard errors, and confidence intervals

Along with the best-fit value of each variable in the equation, nonlinear regression programs usually report its standard error and 95% confidence interval.

By itself, the standard error is difficult to interpret. It is used to calculate the 95% confidence interval, which is easier to interpret.

This is what the CI is supposed to mean: If all the assumptions of nonlinear regression are true, there is a 95% chance that the true value of the variable lies within the interval. More precisely, if you perform nonlinear regression many times (on different data sets) you expect the confidence interval to include the true value 95% of the time, but to exclude the true value the other 5% of the time.

Three factors can make the confidence interval too narrow:

- The CI is based only on the scatter of data points around the curve within this one experiment. If you repeat the experiment many times, the scatter between the results is likely to be greater than predicted from the CI based on one experiment.
- The CI can only be interpreted if you accept the assumptions of nonlinear regression. See "Assumptions of nonlinear regression" on page 8.
- The confidence intervals from linear regression are calculated using straightforward mathematical methods. If you accept the assumptions of linear regression, then you can interpret the 95% CI of slope and intercept quite rigorously. It is not straightforward to calculate the 95% CI of variables from nonlinear regression – mathematical shortcuts are needed. These shortcut intervals (reported by most programs) are sometimes referred to as asymptotic confidence intervals. In some cases these intervals can be too narrow (too optimistic).

Because of these problems, you shouldn't interpret the confidence intervals too rigorously. Rather than focusing on the CI reported from analysis of this one experiment, you should repeat the experiment several times.

Sum-of-squares, $s_{y,x}$, and R^2

The sum-of-squares (SS) is the sum of the square of the vertical distances of the points from the curve. Nonlinear regression works by varying the values of the variables to minimize the sum-of-squares. It is expressed in the square of the units used for the Y values.

The value $s_{y,x}$ is the standard deviation of the vertical distances of the points from the line. Since the distances of the points from the line are

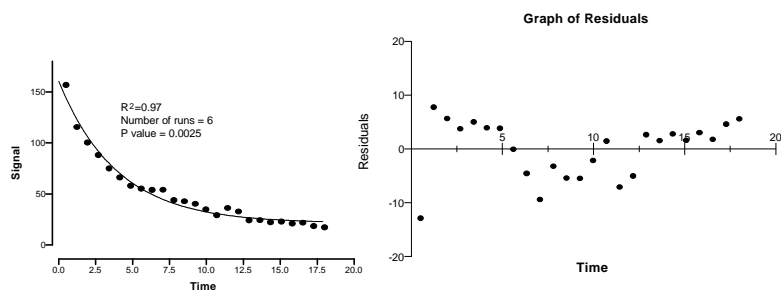
called residuals, $s_{y,x}$ is the standard deviation of the residuals. Its value is expressed in the same units as Y.

The value R^2 is a measure of goodness of fit. It is a fraction between 0.0 and 1.0, and has no units. When R^2 equals 0.0, the best-fit curve fits the data no better than a horizontal line going through the mean of all Y values. In this case, knowing X does not help you predict Y. When $R^2 = 1.0$, all points lie exactly on the curve with no scatter. If you know X you can calculate Y exactly.

You can think of R^2 as the fraction of the total variance of Y that is explained by the model (equation). Mathematically, it is defined by this equation: $R^2 = 1.0 - SS_{\text{reg}}/SS_{\text{tot}}$, where SS_{reg} is the sum-of-squares of the points from the regression curve and SS_{tot} is the sum-of-squares of the distances of the points from a horizontal line where Y equals the mean of all the data points.

Residuals and the runs test

A residual is the distance of a point from the curve. A residual is positive when the point is above the curve, and is negative when the point is below the curve. The residual table has the same X values as the original data, but each Y value is replaced by the vertical distance of the point from the curve.



An example is shown above. If you look carefully at the curve on the left, you'll see that the data points are not randomly distributed above and below the curve. There are clusters of points all above or all below. This is much easier to see on the graph of the residuals on the right. The points are not randomly scattered above and below the X axis.

The runs test determines whether your data differ significantly from the equation you selected. A run is a series of consecutive points that are either all above or all below the regression curve. Another way of saying this is that a run is a series of points whose residuals are either all positive or all negative.

If the data points are randomly distributed above and below the regression curve, it is possible to calculate the expected number of runs. If there are fewer runs than expected, it may mean that the regression model is wrong. The P value from the runs test answers this question: If the data really follow the linear or nonlinear equation used to create the line or curve, what is the chance of obtaining as few (or fewer) runs as observed in this experiment? If the P value is small, you'd be inclined to conclude that the data really don't follow the model.

The P values are always one-tail, asking about the probability of observing as few runs (or fewer) than observed. If you observe more runs than expected, the P value will be higher than 0.50.

If the runs test reports a low P value, you should suspect that the data don't really follow the equation you have selected.

In the example above, the equation does not adequately match the data. There are only six runs, and the P value for the runs test is tiny. This means that the data systematically deviate from the curve. Most likely, the data were fit to the wrong equation.

How to tell if the nonlinear regression fit is any good

Before accepting the results that any curve fitting program gives you, ask yourself the following questions:

Did the fit converge on a solution?

Nonlinear regression stops its iterations when it can't improve the fit by adjusting to the values of any of the variables. At that point, the program is said to have converged on the best-fit. In some cases, the program gets stuck. It doesn't know whether the fit would improve by increasing or decreasing the value of a variable. When this happens, the program stops and says that it was unable to converge on a solution. No results are reported.

Does the curve come close to the points?

In rare cases, the fit may be far from the data points. This may happen, for example, if you picked the wrong equation. Look at the graph to make sure this didn't happen.

Also look at the R^2 value. It is the fraction of the overall variance in Y that is "explained" by the model. See "R" on page 8. If R^2 is low, the curve does not come close to the points. If R^2 is high, you can conclude that the curve comes closer to the points than would a horizontal line through the mean Y value. But don't over interpret a high R^2 . It does not mean that you have chosen the equation that best describes the data. It also does not mean that the fit is unique — other values of the variables may generate a curve that fits just as well.

Are the results scientifically plausible?

A computer program can only follow a procedure to fit a curve. It is up to you to figure out what they mean. Before accepting the results, ask yourself if the results make any sense.

The mathematics of curve fitting sometimes yields results that make no scientific sense. For example with noisy or incomplete data, nonlinear regression can calculate negative rate constants, fractions greater than 1.0, and negative K_D values. It's up to you to realize that these are nonsense.

If the results make no scientific sense, you should conclude that the fit is no good, regardless of R^2 and regardless of how close the curve comes to the points. Try a simpler equation, or try fixing some variables to constant values.

Also check that the best-fit values of the variables make sense in light of the range of the data. Don't trust the results if the top plateau of a sigmoid curve is far larger than the highest data point. Don't trust the results if an EC_{50} value is not within the range of your X values.

Do the data systematically deviate from the curve?

If the data really follow the model described by your equation, the data points should randomly bounce above and below the curve. The distance of the points from the curve should also be random, and not be related to the value of X.

The best way to look for systematic deviations of the points from the curve is to inspect a graph of the residuals and to look at the runs test. See "Residuals and the runs test" on page 9. With a good fit, the residuals should be randomly distributed between positive and negative values and the P value from the runs test will be high.

If the runs test reports a low P value, you should suspect that the data don't really follow the equation you have selected.

Are the confidence intervals wide?

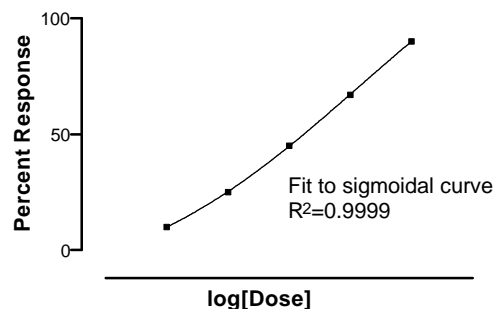
Most nonlinear regression programs report the standard error of each variable, and its 95% confidence interval. You can be approximately 95% sure that the true value of the variable lies within the confidence interval.

The confidence interval will be very wide (i.e. the standard error will be very large) when the fit is not unique. This means that curves generated from other values of the variables would fit nearly as well.

Confidence intervals are wide in these circumstances:

- You have not collected data over a wide enough range of X values. See the first example below.
- You have not collected data in an important part of the curve. See the second example below.
- The data are very scattered.
- The equation contains redundant variables. For example, the confidence intervals would be very wide if you fit this equation: $Y = A + B + C \cdot X$. This equation describes a line, but the intercept is defined by the sum of A plus B. There is no way for the program to know how to apportion the value between A and B, so both will have very wide confidence intervals.

Example 1. Data not collected over a wide range of X.



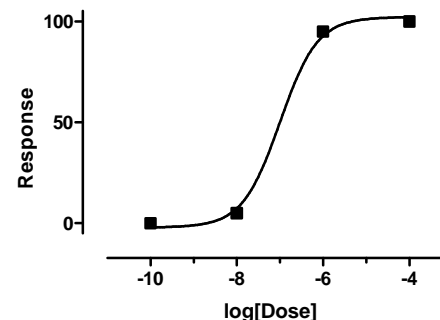
This best-fit dose response curve has wide confidence intervals. The 95% CI for the EC_{50} extends over six orders of magnitude!

The explanation is simple. The data were fit to a sigmoidal equation with four variables: the top plateau, the bottom plateau, the slope, and the EC_{50} (the $\log[Dose]$ when response = 50%). But the data do not form plateaus at either the top or the bottom, so best-fit values for the plateaus are very uncertain. The information is simply not in the data. Since the data do not clearly define zero and one hundred, the value for the EC_{50} is very imprecise. The program indicates this by reporting large standard errors and wide confidence intervals. The fit is not unique. You could find other values of the variables that fit the data equally well.

In this example, it might make scientific sense to set the bottom plateau to 0% and the top plateau to 100% (if the plateaus were defined by other controls not shown on the graph). If you did this, the equation would fit fine and the confidence interval would be narrow.

Note that the problem with the fit is not obvious by inspecting a graph, because the curve goes very close to the points. The value of R^2 (0.9999) is also not helpful. That value just tells you that the curve comes close to the points, but does not tell you whether the fit is unique.

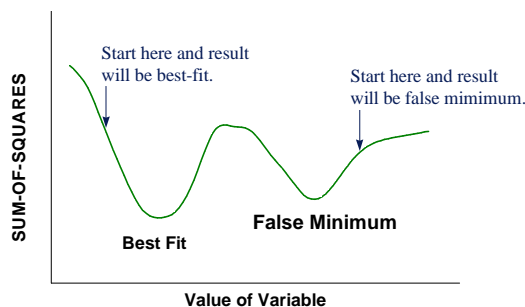
Example 2. No data in an important part of the curve.



This dose-response curve also has wide confidence intervals. Even if you constrain the bottom to be zero and the top to be 100 and the slope to equal 1.0, the 95% CI for the EC_{50} extends over almost an order of magnitude. The problem is simple. The EC_{50} is the concentration at which the response is half-maximal, and this example has no data near that point.

Is the fit a local minimum?

The nonlinear regression procedure adjusts the variables in small steps in order to improve the goodness-of-fit. If nonlinear regression converges on an answer, you can be sure that altering any of the variables a little bit will make the fit worse. But it is theoretically possible that large changes in the variables might lead to much better goodness-of-fit. Thus the curve that the program decides is the "best" may really not be the best.



Think of latitude and longitude as representing two variables you are trying to fit. Now think of altitude as the sum-of-squares. Nonlinear works iteratively to reduce the sum-of-squares. This is like walking downhill to find the bottom of the valley. When nonlinear regression has converged, changing any variable increases the sum-of-squares. When you are at the bottom of the valley, every direction leads uphill. But there may be a much deeper valley over the ridge that you are unaware of. In nonlinear regression, large changes in variables might decrease the sum-of-squares.

This problem (called finding a local minimum) is intrinsic to nonlinear regression, no matter what program you use. You will rarely encounter a local minimum if your data have little scatter, you collected data over an appropriate range of X values, and you have chosen an appropriate equation.

To continue the analogy, the confidence intervals for the variables are very wide when the bottom of the valley is very flat. You can walk a great distance without changing elevation. You can change the values of the variables a great deal without changing the goodness-of-fit.

To test for the presence of a false minimum:

1. Note the values of the variables and the sum-of-squares from the first fit.
2. Make a large change to the initial values of one or more variables and run the fit again.
3. Repeat step 2 several times.
4. Ideally, curve fitting programs will report nearly the same sum-of-squares and same variables regardless of the initial values. If the values are different, accept the ones with the lowest sum-of-squares.

What to do when the fit is no good?

The last section explained that a fit is bad when the fit did not converge, the results make no sense, or the confidence intervals are wide. If you encounter any of these situations, here are some things to try.

Potential problem	Solution
The equation simply does not describe the data.	Try a different equation.
The initial values are too far from their correct values.	Enter different initial values. If you are using a user-defined equation, check the rules for initial values.
The range of X values is too narrow to define the curve completely.	If possible, collect more data. Other wise, hold one of the variables to a constant value.
You have not collected enough data in a critical range of X values.	Collect more data in the important regions.
Your data are very scattered and don't really define a curve.	Try to collect less scattered data. If you are combining several experiments, normalize the data for each experiment to an internal control.
The equation includes more than one component, but your data don't follow a multicomponent model.	Use a simpler equation.
Your numbers are too large.	If your Y values are huge, change the units. Don't use values greater than about 10^4 .
Your numbers are too small.	If your Y values are tiny, change the units. Don't use values less than about 10^{-4} .

Comparing two equations

Sometimes you don't know which of two equations is more appropriate for your data. You want to fit both equations, and let the program compare the results. For example, you might want to fit a competitive binding curve to models with both one and two binding sites. Or you might want to fit a dissociation kinetics curve to exponential decay equations with both one and two phases.

Goodness of fit is quantified by the sum-of-squares. Therefore you might imagine that you can simply define the "best" equation as the one that gives the smaller sum-of-squares. That rule makes sense when both equations have the same number of variables.

Most often, however, you wish to compare equations with different numbers of variables. If the more complicated equation fits worse than the simpler equation, then you should clearly stick with the simpler equation. However, the curve generated by the more complicated equation (the one with more variables) will nearly always come closer to the points because it has more inflection points (it wiggles more). The question is whether this decrease in sum-of-squares is worth the "cost" of the additional variables (loss of degrees of freedom). The F test addresses this question by calculating a P value that answers this question: If the simpler model is really correct, what is the chance that you'd randomly obtain data that fits the more complicated model so much better? If the P value is low, you conclude that the more complicated model is significantly better than the simpler model.

The results of the F test are only strictly valid when the simpler equation is a special case of the more complicated equation. For example, you can compare a one-site vs. two-site binding curve.

How the F test works

First fit the more complicated model (Model 2) and calculate its goodness-of-fit as the sum-of-squares. Now fit the simpler model (Model 1). Even if this simpler model is correct, you expect it to fit worse (have the higher sum-of-squares) because it has fewer inflection points (more degrees of freedom). In fact, statisticians can prove that the relative increase in the

sum of squares is expected to equal the relative increase in degrees of freedom. In other words, if the simpler model is correct you expect that:

$$(SS1 - SS2) / SS2 \approx (DF1 - DF2) / DF2$$

If the more complicated model is correct, then you expect the relative increase in sum-of-squares (going from complicated to simple model) to be greater than the relative increase in degrees of freedom:

$$(SS1 - SS2) / SS2 > (DF1 - DF2) / DF2$$

The F ratio quantifies the relationship between the relative increase in sum-of-squares and the relative increase in degrees of freedom.

$$F = \frac{(SS1 - SS2) / SS2}{(DF1 - DF2) / DF2}$$

If the simpler model is correct you expect to get an F ratio near 1.0. If the ratio is much greater than 1.0, there are two possibilities:

- The more complicated model is correct.
- The simpler model is correct, but random scatter led the more complicated model to fit better. The P value tells you how rare this coincidence would be.

The P value answers this question: If model 1 is really correct, what is the chance that you'd randomly obtain data that fits model 2 so much better? If the P value is low, you conclude that model 2 is significantly better than model 1.

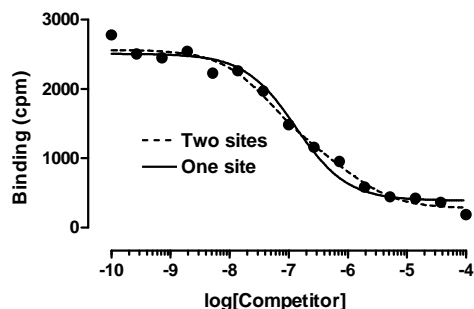
The equation is usually presented in this more conventional form.

$$F = \frac{(SS1 - SS2) / (DF1 - DF2)}{SS2 / DF2} \quad DF_n = (Df1 - DF2), \quad DF_d = DF2$$

If you are extremely familiar with analysis of variance, you'll appreciate that the F ratio is determined from this analysis of variance tables,

Source of variation	Sum-of-squares	df	MS
Difference	SS1 - SS2	DF1 - DF2	$\frac{SS1 - SS2}{DF1 - DF2}$
Model 2 (complicated)	SS2	DF2	SS2/DF2
Model 1 (simple)	SS1	DF1	

Example



This graph compares a one-site and two-site competitive binding curve. The results are shown here:

	Two-site	One-site	% Increase
Degrees of freedom	10	12	20.00%
Sum-of-squares	129800	248100	91.14%

In going from the two-site to the one-site model, we gained two degrees of freedom because the one-site model has two fewer variables. Since the two-site model has 10 degrees of freedom (15 data points minus 5 variables), the degrees of freedom increased 20%. If the one-site model were correct, you'd expect the sum-of-squares to also increase about 20% just by chance. In fact the sum-of-squares increased 91%. The percent increase was 4.56 times higher than expected ($91.1/20.0 = 4.56$). This is the F ratio ($F = 4.56$), and it corresponds to a P value of 0.039. If the one-site model is correct, there is only a 3.9% chance that you'd randomly obtain data that fits the two-site model so much better. Since this is below the

traditional threshold of 5%, you'd probably conclude that the two-site model fits significantly better than the one-site model.

Comparing fits to two data sets

The previous section discussed how to compare the fits of two different equations to one set of data. Here we discuss how to compare the fit of one equation to two different sets of data, for example comparing fits to data from control and treated preparations. Although this is a common situation, there is no clear consensus for how to do it. Three approaches are discussed below.

1. Compare the results of repeated experiments.

If you repeat the experiment several times, you can compare the best-fit value of a variable in control and treated preparations using a paired t test (or the analogous Wilcoxon nonparametric test).

For example, here are the results of a competitive binding curve performed in two groups of cells. The table shows the $\log(K_i)$ values.

Experiment	Control	Treated
1	-6.13	-6.53
2	-6.39	-6.86
3	-5.92	-6.31

Compare the results using a paired t test. The t ratio is 16.7, and the P value is 0.0036 (two-tail). If the treatment did not alter the $\log(K_i)$, there is only a 0.36% chance that you observe such a large difference (or larger) between $\log(K_i)$ by chance.

Notes that we compare $\log(K_i)$ values, not K_i values. When doing a paired t test, a key assumption is that the distribution of differences (treated - control) follow a Gaussian distribution. Since a competitive binding curve (similar to a dose response curve) is conducted with X values (concentration) equally spaced on a log scale, the uncertainty of X is reasonable symmetrical (and perhaps Gaussian) when expressed on a log scale. It is equally likely that the observed $\log K_i$ is 0.1 log units too high or 0.1 log units too low. In contrast, the uncertainty in K_i is not symmetrical.

2. Compare the results within one experiment. Simple approach.

When a nonlinear regression program reports the best-fit value of each variable for each data set, it also reports the standard error of the estimates. You can compare the best-fit values between two data sets using a t test.

For example, here are the results of fitting control and treated data to a competitive binding curve:

	Control	Treated
log(EC ₅₀)	-6.08	-6.29
SE	0.0677	0.0617
Sum-of-squares	19560	29320
df	12	12

Nonlinear regression fit three variables, Top, Bottom, and Log(EC₅₀). We only care about comparing the Log(EC₅₀) values. Compare these with a t test, basing the calculation on the best fit value and its SE. The only trick is deciding what value to enter for N. Follow this logic:

- For nonlinear regression, the number of degrees of freedom (df) equals the number of data points minus the number of variables fit. In this example, there were 15 data points, and three variables were fit. So there are 12 df.
- For an ordinary t test, the number of df for each sample equals one less than the number of data points.
- The t test calculations are based on the numbers of degrees of freedom. There is no way to enter degrees of freedom into most statistics programs – instead you enter N. Programs that perform t tests are programmed to compute the df as N-1. When comparing the results of nonlinear regression, enter N as the number of degrees of freedom plus 1. The program will subtract 1, and make the df correct. In this example, enter $N = 12 + 1 = 13$.

For this example, here are the values you enter into a statistics program:

	Control	Treated
Mean	-6.08	-6.29
SEM	0.0677	0.0617
N	13	13

The t ratio is 2.29 with 24 degrees of freedom. The two-tailed P value is 0.0309. If the treatment really didn't alter the log(EC₅₀), there is only a 3.09% chance that you'd observe this large of a difference (or more) by coincidence. Since the P value is so low, you conclude that the difference between the two EC₅₀ values is statistically significant.

Notes:

- This method only uses data from one experiment. The SE value is a measure of how precisely you have determined the log(EC₅₀) in this one experiment. It is not a measure of how reproducible the experiment is. Despite the impressive P value, I wouldn't trust these results until the experiment is repeated.
- Compare Log(EC₅₀), not EC₅₀. You want to express the variables in a form that makes the uncertainty as symmetrical and Gaussian as possible. Since a competitive binding curve (similar to a dose response curve) is conducted with X values (concentration) equally spaced on a log scale, the uncertainty of Log(EC₅₀) is reasonable symmetrical (and perhaps Gaussian) when expressed on a log scale. It is equally likely that the observed Log(EC₅₀) is 0.1 log units too high or 0.1 log units too low. In contrast, the uncertainty in EC₅₀ is not symmetrical.

3. Compare the results within one experiment. More complicated approach.

The method of the previous section only compared the value of the log(EC₅₀). This section describes a more general method to compare entire curves to ask whether the data sets differ at all.

The idea is to first fit the two curves separately, and then combine the values and fit one curve to all the data. Finally compare the sum of the

two individual sum-of-squares values with the sum-of-squares from the combined data.

Follow these steps:

1. Fit the two data sets separately. We did this in the previous section.
2. Total the sum-of-squares and df from the two fits. For this example the total sum of squares equals. $19560 + 29320 = 48880$, and the total df equals $12 + 12 = 24$. Since these are the results of fitting the two data sets separately, label these values SS_{separate} and DF_{separate}
3. Combine the two data sets into one. For this example, the combined data set has XY pairs, with each X value appearing in twice.
4. Fit the combined data set to the same equation. Note the SS and df. For this example, $SS = 165200$, and $df = 27$ (30 data points minus three variables). Call these values SS_{combined} and DF_{combined} .
5. You expect SS_{separate} to be smaller than SS_{combined} even if the curves are really identical simply because the separate fits have more degrees of freedom. The question is whether the SS values are more different than you'd expect to see by chance. To find out, calculate the F ratio using this equation:

$$F = \frac{(SS_{\text{combined}} - SS_{\text{separate}}) / (DF_{\text{combined}} - DF_{\text{separate}})}{SS_{\text{combined}} / DF_{\text{separate}}}$$

For this example, $F = 19.03$.

6. Determine the P value from F. There are $DF_{\text{combined}} - DF_{\text{separate}}$ degrees of freedom in the numerator, and DF_{separate} degrees of freedom in the denominator. If you don't have a program that does this, consult tables in the back of most statistics books.
7. For this example, the P value is less than 0.0001. If the treatment were really ineffective, there is less than a 0.01% chance that the two curves would differ as much (or more) than observed in this experiment. Since the P value is low, you'll conclude that the curves are really different.

Notes:

- This method only uses data from one experiment. I wouldn't trust these results until the experiment is repeated.
- This method compares the curves overall. It doesn't tell you which variable(s) are different. Differences might be due to something trivial like a different baseline, rather than something important like a different rate constant.

Advantages and disadvantages of the three methods

If you have repeated the experiment several times, I recommend that you use the first method. There are two advantages:

- Compared to the other methods discussed below, this method is far easier to understand and communicate to others.
- The entire test is based on the consistency of the results between repeat experiments. Since there are usually more causes for variability between experiments than within experiments, it makes sense to base the comparison on differences between experiments.

The disadvantage of the first method is that you are throwing away information. The calculations are based only on the best-fit value from each experiment, and ignores the SE of those values presented by the curve fitting program.

If you have performed the experiment only once, then you probably ought to repeat the experiment. Regardless of what statistical results you obtain, you shouldn't trust results from a single experiment. If you want to compare results in a single experiment, you can use method 2 or method 3.

The advantage of method 2 is that it focuses your thinking on a single variable. Generally, you care mostly about one variable (i.e. a rate constant or EC_{50}), and care less about the others. Method 2 compares the variable of interest.

Method 3 is the most general method. Since the method compares the entire curve, it does not force you to decide which variable(s) you wish to compare. This is both its advantage and disadvantage.

GraphPad Prism

All the graphs in this booklet were created using GraphPad Prism, a general-purpose curve fitting and scientific graphics program for Windows.

Although GraphPad Prism was designed to be a general purpose program, it is particularly well-suited for nonlinear regression:

- Prism provides a menu of commonly-used equations (including equations used for analysis of radioligand binding experiments). To fit a curve, all you have to do is pick the right equation. Prism does all the rest automatically, from picking initial values to graphing the curve.
- Prism can automatically compare two models with the F test.
- When analyzing competitive radioligand binding curves, Prism automatically calculates K_i from IC_{50} .
- You can use the best-fit curve as a standard curve. Enter Y values and Prism will determine X. Or enter X and Prism will determine Y.
- Prism can automatically graph a residual plot and perform the runs test.
- Prism's manual and help screens explain the principles of curve fitting with nonlinear regression and help you interpret the results. You don't have to be a statistics expert to use Prism.

Please visit our web site at <http://www.graphpad.com>. You can read about the Prism and download a free demo. Or contact GraphPad Software to request a brochure and demo disk by phone (619-457-3909), fax (619-457-8141) or email (sales@graphpad.com). The demo is not a slide show – it is a functional version of Prism with no limitations in data analysis. Try it out with your own data, and see for yourself why Prism is the best solution for analyzing and graphing scientific data.