STATISTICS SURVIVAL SERIES

5 Common Regression Analysis Mistakes and How to Avoid Them

GraphPad

Refresher: The Principles of Regression Analysis

What you're trying to learn from regression analysis is how your dependent variable(s) increase or decrease as your independent variables change. You do this by fitting a model to your data.

Before we jump into some common mistakes made during regression analysis, let's quickly refresh on the underlying principles.

Regression analysis helps you understand the relationship between dependent and independent variables in order to help you make predictions. What you're trying to learn from regression analysis is how your dependent variable(s) increase or decrease as your independent variables change. You do this by fitting a model to your data.

This model is not designed to describe the relationship perfectly. Rather, your goal is to find as simple a model as possible that comes close to describing your system so you understand the system, reach valid scientific conclusions, and design new experiments.

Here is a simple, real-world example that most of you have probably encountered to help understand the goal of regression analysis.

If you drive a gas-powered car, you've realized that there is a relationship between the amount of money that you spend to put gasoline in the vehicle with how far the vehicle can go before you need to add more. By understanding the relationship between these two factors (amount of money spent on fuel and distance you can travel), you're essentially performing a regression analysis. What's more, you can also use this analysis to make predictions about future events. If you need to make a long road trip and know about how far you'll be driving, you can use the relationship described by the regression analysis to estimate how much money you'll need to spend on gas. That's one simple form of regression analysis, but the applications are vast.

The Idea of Regression

All models in regression analysis define an outcome (Y) as a function of one or more parameters and an independent variable (X) [or several independent variables]. The goal is to adjust the values of the model's parameters to find the line or curve that comes closest to your data.

For example, with linear regression, the goal is to find the best-fit values of the slope and intercept that optimizes the distance of the line to the data. With nonlinear regression of a normalized dose-response curve, the goal is to adjust the values of the EC50 (the concentration that provokes a response halfway between the minimum and maximum responses) and the slope of the curve.

The Goals of Regression

Typically, scientists use regression with one of three distinct goals:

To fit a model to your data in order to obtain best-fit values of the parameters, or to compare the fits of alternative models. If this is your goal, you must pick a model (or two alternative models) carefully, and pay attention to all the results. The whole point is to obtain best-fit values for the parameters, so you need to understand what those parameters mean scientifically.

To fit a smooth curve in order to interpolate values from the curve, or perhaps to draw a graph with a smooth curve. If this is your goal, you can assess it purely by looking at the graph of data and curve. There is no need to learn much theory.

To make predictions based on your data. Unlike the case of simply fitting a smooth curve, in order to make predictions, you must understand how the data were generated as well as why you should select a specific model (and its parameters) for that data.

Complexity of Regression Analysis

Although many scientists perform regression analysis more than any other statistical technique, many state they don't understand the underlying principles. It is a flexible and powerful tool. It can also be complex.

The following resources in this guide will help you understand the common regression analysis mistakes, and provide advice so you can avoid them.

MISTAKE #1

Using Linear Regression Instead of Nonlinear Regression

Before analyzing your data with linear regression, stop and ask yourself whether it might make more sense to fit your data with nonlinear regression.

How Linear and Nonlinear Regression Work

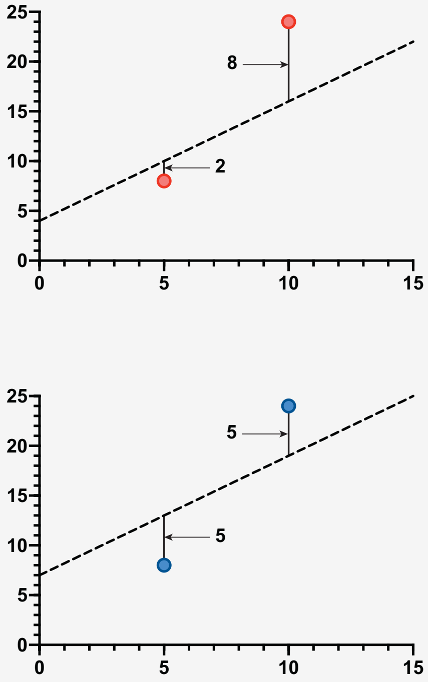
A line is described by a simple equation that calculates Y from X, slope and intercept (y = mx + b), slope and intercept. The purpose of linear regression is to find values for the slope (m) and intercept (b) that define the line that comes closest to the data. Just like linear regression, nonlinear regression also attempts to find the values of the parameters that make the line or curve come as close as possible to the data.

Both linear and nonlinear regression find the values of the parameters (slope and intercept for linear regression) that make the line or curve come as close as possible to the data. More precisely, this process attempts to minimize the sum of the squares of the vertical distances of the points from the curve.

Linear regression accomplishes this goal using math that can be completely explained with simple algebra (shown in many statistics books). Put the data in, and the answers come out. There is no chance for ambiguity. You could even do the calculations by hand, if you wanted to.

Nonlinear regression uses a computationally intensive, iterative approach that can only be explained using calculus and matrix algebra. The method requires initial estimated values for each parameter.

> The purpose of **linear regression** is to find values for the slope (m) and intercept (b) that define the line that comes closest to the data.



Why Minimize the Sum of Squared Distances?

When performing linear (or nonlinear) regression, you want your curve to come as close to as many data points as possible. Intuitively, you may think that minimizing the sum of the actual distances of the points to the fit line (or curve) would work. However, imagine a curve passing by two points of a larger data set: one at a distance of 2 units, and the other at a distance of 8 units. The sum of distances in this scenario would be 10 units. A second possible fit curve could pass by these same points at a distance of 5 units each, and again, the sum of distances would be 10 units. Linear and nonlinear regression assume that the error in measurements follows a Gaussian distribution. As such, it's far more likely to have two medium size deviations than to have one small deviation and one large deviation. Calculating the sum of squared distances in the previous example results in a value of 68 ($8^2 + 2^2$) for the first fit, but only 50 ($5^2 + 5^2$) for the second fit. Minimizing the sum of squared distances of *all* of the points in a data set provides the line (or curve) most likely to be correct.

Linear Regression is a Special Case of Nonlinear Regression

Nonlinear regression can fit any model, including a linear one. Therefore, linear regression is just a special case of nonlinear regression.

Even if your goal is to fit a straight line through your data, there are many situations where it makes sense to choose nonlinear regression rather than linear regression.

While using nonlinear regression to analyze data is only slightly more difficult than using linear regression, your choice of linear or nonlinear regression should be based on the model you are fitting.

Even if your goal is to fit a straight line through your data, there are many situations where it makes sense to **choose nonlinear regression rather than linear regression.**

Tip: Avoid Outdated Data Transformations

If you have transformed nonlinear data to create a linear relationship, you will almost certainly be better off fitting your original data using nonlinear regression.

Before nonlinear regression was readily available, the best way to analyze nonlinear data was to transform the data to create a linear graph, and then analyze the transformed data with linear regression. Examples include Lineweaver-Burk plots of enzyme kinetic data, Scatchard plots of binding data, and logarithmic plots of kinetic data.

> 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 rarely true after transforming data.**

These methods are **outdated**, and should not be used to analyze data.

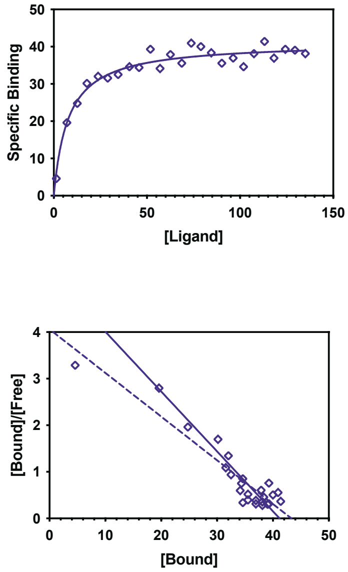
The problem with these methods is that the transformation distorts the experimental error. Linear regression assumes that the scatter of points around the line follows a Gaussian distribution and that the standard deviation is the same at every value of X. These assumptions are rarely true after transforming data.

Furthermore, some transformations alter the relationship between X and Y. For example, in a Scatchard plot the value of X represents the concentration of ligand bound by a receptor ([bound]), while Y represents the ratio of the concentration of bound ligand vs. the concentration of free ligand ([bound]/[free]). When linear regression is performed, X ([bound]) is used to calculate Y ([bound]/[free]), but this violates the assumption of linear regression that all uncertainty is in Y while X is known precisely. It doesn't make sense to minimize the sum of squares of the vertical distances of points from the line, if the same experimental error appears in both X and Y directions.

Since the assumptions of linear regression are violated, 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 technique for analyzing your data. Nonlinear regression produces the most accurate results.

The Problem with Transforming Data

The figures here show the problem of transforming data. The top panel shows data that follow a rectangular hyperbola (binding isotherm). The bottom panel is a Scatchard plot of the same data. The solid curve in the top panel was determined by nonlinear regression. The solid line in the bottom panel shows how that same curve would look after performing a Scatchard transformation. In contrast, the dotted line in the bottom panel shows a line generated by a linear regression fit to the data after it had already been transformed. Scatchard plots can be used to determine the receptor number (B_{max}, determined as the X-intercept of the linear regression line) and the equilibrium dissociation constant (K_d, determined as the negative reciprocal of the slope). Since the Scatchard transformation amplified and distorted the scatter of the points, the linear regression fit performed after transformation will not yield the most accurate values for B_{max} and K_d.

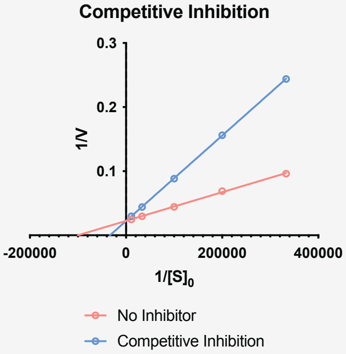


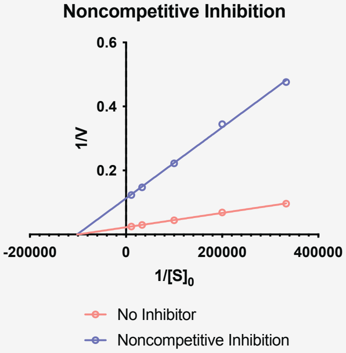
Linear Transformations Are Still Useful, But Not for Analysis

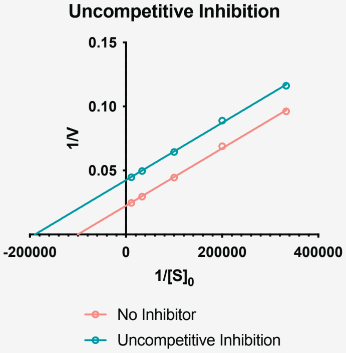
Although it is usually inappropriate to analyze transformed data, it is often helpful to display data after a linear transformation. Many people find it easier to visually interpret transformed data. This makes sense because the human eye and brain evolved to detect edges (lines) — not to detect rectangular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

Take, for example, these Lineweaver-Burk plots showing various types of inhibition of an enzyme-substrate reaction

Using these linear transformations, you can quickly see that for competitive inhibition the two lines intersect at the y-axis, while for noncompetitive inhibition the lines intersect at the x-axis, and for uncompetitive inhibition the two lines are parallel. Using linear transformations to present data and make quick, general inferences this way is perfectly acceptable, but nonlinear regression should still be used to calculate the desired parameters from your data.







The Bottom Line: Don't use linear regression just to avoid using nonlinear regression. Fitting curves with nonlinear regression is not difficult, and provides more accurate estimates for the parameters of your data.

CHAPTER 3

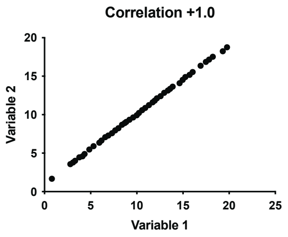
MISTAKE #2

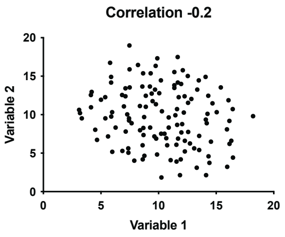
Confusing Linear Regression With Correlation

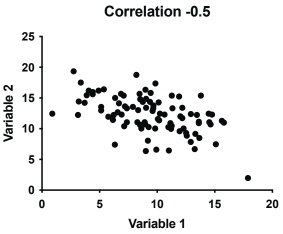
First off, correlation and linear regression are not the same. Understanding each of these terms and their differences will help you avoid the mistake of confusing these two seemingly similar concepts.

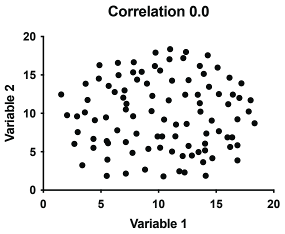
What is Correlation?

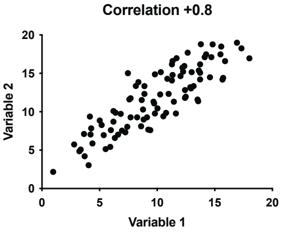
In the previous chapter, you saw that linear regression is a technique used to find a best line that can predict values of Y from values of X. In comparison, correlation is a technique used to quantify the degree to which two variables are related. However, correlation does not fit a line through the data points. This technique is simply used to compute a correlation coefficient (r) that tells you how much one variable tends to change as the other variable changes. The value of r provides this information: when r is 0.0, there is no relationship; when r is positive, the trend in the data is that one variable increases as the other variable increases; when r is negative, the trend is that one variable increases as the other decreases. The value of r can range from +1 to -1, and provides some information on how “strongly” the two variables are correlated. That’s it!











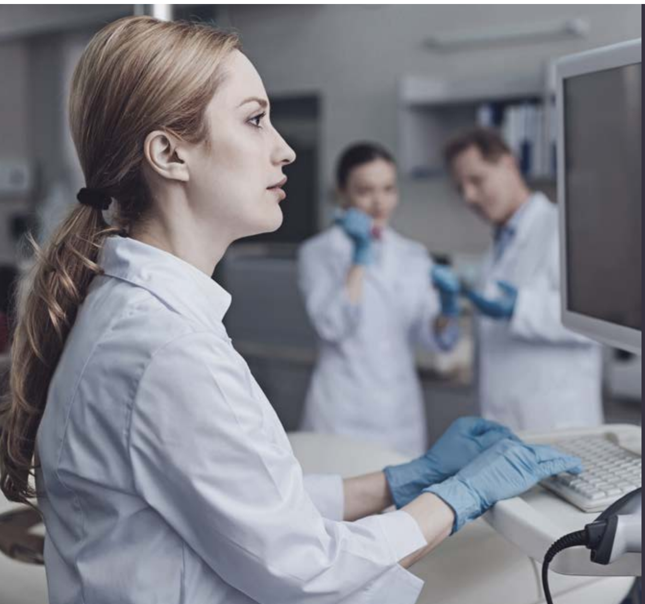
So What Are the Differences?

Aside from linear regression generating a line and correlation simply providing a measurement of how the variables tend to change with respect to each other, there are some important differences between these two techniques. First, linear regression is usually used when X is a variable that you manipulate (time, concentration, etc.) to obtain a measurement for Y. In this relationship, X is referred to as the "independent" variable, while Y is the "dependent" variable. As an example, imagine you wanted to measure the growth of a plant exposed to different amounts of sunlight. You would select various exposure times of sunlight (your X or independent variable) and you would measure the resulting growth (your Y or dependent variable) after some time.

In contrast, correlation is almost always used when you measure both variables. It is rarely appropriate when one variable is something you experimentally manipulate. Say, for example, you wanted to see the relationship between the height and weight of various professional athletes. You could collect both measurements for a number of different players, but you would not be experimentally determining either.

Another difference between these two techniques is that the decision of which variable you call "X" and which you call "Y" matters in regression. The line that best predicts Y from X is not the same as the line that predicts X from Y (and scientifically, this often makes no sense anyway). With correlation, you don't have to think about cause and effect. It doesn't matter which of the two variables you call "X" and which you call "Y".

> Say, for example, you wanted to see the relationship between the height and weight of various professional athletes. **You could collect both measurements for a number of different players,** but you would not be experimentally determining either.



The Bottom Line

By now, you've seen that linear regression and correlation are definitely not the same thing. However, these two techniques do share some similarities.

Linear regression quantifies the goodness of fit of the determined line with the term "r^2", sometimes shown as "R^2". If you put the same data into correlation (which as we showed is rarely appropriate), the square of r from correlation will equal r^2 from regression

MISTAKE #3

Fitting a Model to Smoothed Data

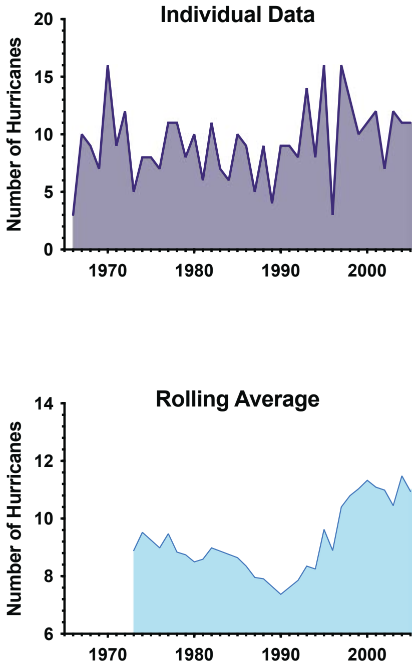
You should not fit a model to the rolling average data with linear or nonlinear regression, or compute a correlation coefficient.

The problem is that regression assumes that each value is independent of the others, but the rolling average are not at all independent of each other. Rather each value is included as part of the neighboring values.

Illustrative Example: Hurricanes Over Time

The figure here shows the number of hurricanes over time. The top panel shows the number of hurricanes in each year, which jumps around a lot. To make it easier to spot trends, the bottom panel shows a rolling average. The value plotted for each year is the average of the number of hurricanes for that year plus the prior eight years. This smoothing lets you see a clear trend.

But there is a problem. These are not real data. Instead, the values plotted in the left panel were chosen randomly (from a Poisson distribution, with a mean of 10). There is no pattern. Each value was randomly generated without regard to the previous (or later) values.



The Bottom Line: Creating the running average creates the impression of trends by ensuring that any large random swing to a high or low value is amplified, while variability is muted. This is misleading at best, and often times makes the research invalid.

MISTAKE #4

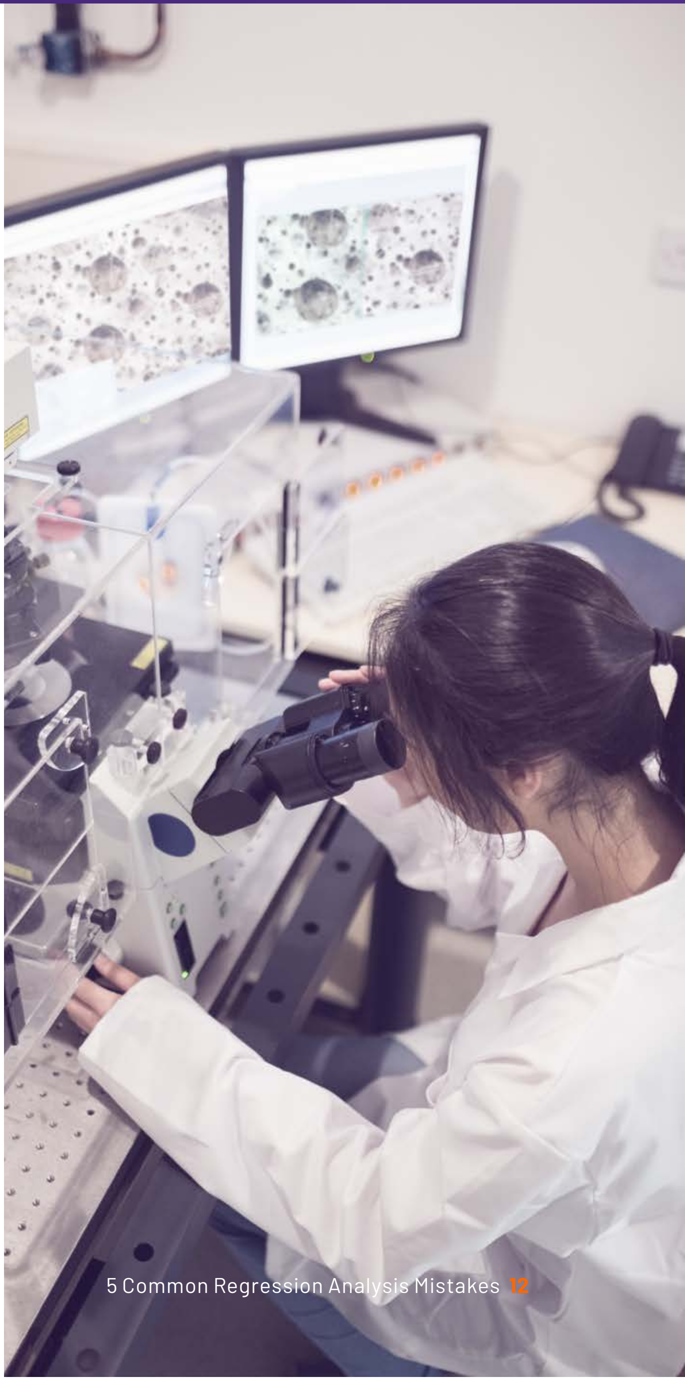
To Remove or Not to Remove: Data Points and Outliers

Nonlinear regression, like linear regression, assumes that the scatter of data around the ideal curve follows a Gaussian or normal distribution. This assumption leads to the familiar goal of regression: to minimize the sum of the squares of the vertical or Y-value distances between the points and the curve.

However, experimental mistakes can lead to erroneous values—outliers. Even a single outlier can dominate the sum-of-the-squares calculation, and lead to misleading results.

Is it 'Cheating' to Remove Outliers?

Some people feel that removing outliers is 'cheating'. It can be viewed that way when outliers are removed in an ad hoc manner, especially when you remove only outliers that get in the way of obtaining results you like. But leaving outliers in the data you analyze can also be perceived as 'cheating', as it can lead to invalid results.



An Approach to Removing Outliers

Here is a Bayesian way to think about systematic approaches to removing outliers. When a value is flagged as an outlier, there are two possibilities.

- A coincidence occurred, the kind of coincidence that happens in few percent of experiments even if the entire scatter is Gaussian (depending on how aggressively you define an outlier).
- A 'bad' point got included in your data.

Which possibility is more likely? It depends on your experimental system.

If your experimental system generates a 'bad' point in a few percent of experiments, then it makes sense to eliminate the point as an outlier. It is more likely to be a 'bad' point than a 'good' point that just happened to be far from the curve.

If your system is very pure and controlled, 'bad' points occur very rarely and it is more likely that the point is far from the curve due to chance (and not mistake) and you should leave it in. Alternatively, in that case you can change the threshold for defining outliers in order to only detect outliers that are much further away.

> Removing 'bad' points isn't necessarily a bad thing, but **outliers might also be telling you something important** about your research.

Remember that Outliers Aren't Always 'Bad' Points

In some situations, data points that at first appear to be outliers may not have been caused by experimental mistakes, but rather be the result of biological variation, or differences in some other variable that is not included in your model. Here, the presence of an outlier may be the most interesting finding of your study. It would be a big mistake to automatically exclude such outliers in this situation without further thought (or experimentation).

The Bottom Line

There are certainly circumstances that would require you to remove outliers. Consider each case before you do so to improve accuracy. Removing 'bad' points isn't necessarily a bad thing, but outliers might also be telling you something important about your research.

MISTAKE #5

Allowing a Program to Select a Model For You

The goal of nonlinear regression is to fit a model to your data. The program finds the best-fit values of the parameters 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 data on your graph.

However, some programs automatically fit data to 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.

> 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 data on your graph.**

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 parameters, so the results are unlikely to be useful.

Letting a program choose a model for you can be useful if your goal is to simply create a smooth curve for simulations or interpolations. In these situations, you don't care about the value of the parameters or the meaning of the model. You only care that the curve fit the data well and does not wiggle too much. Avoid this approach when the goal of curve fitting is to fit the data to a model based on chemical, physical, or biological principles.

The Bottom Line: Don't use a computer program as a way to avoid understanding your experimental system, or to avoid making scientific decisions.

Regression Analysis with GraphPad Prism

GraphPad Prism is the world's leading data analysis and graphing solution purpose-built for scientific research.

750,000 of the world's leading scientists use Prism to save time performing statistical analyses, make more accurate analysis choices, and elegantly graph and present their scientific research.

Download a free trial today—no credit cards, no commitments— and be on your way to sharing your research with the world!

www.graphpad.com

FOR MAC AND WINDOWS