Molecular Life Sciences
DOI 10.1007/978-1-4614-6436-5_560-1

# Mathematics of Fitting Scientific Data

John Wesley Cain*
Department of Mathematics and Computer Science, University of Richmond, Richmond, VA, USA

## Synopsis

The ability to make predictions based upon scientific data is fundamentally important. Interpolation and extrapolation of data allow researchers to predict how a system will behave and sometimes elucidate the mechanisms responsible for observed behaviors. The usual way of fitting scientific data is via least squares regression, a systematic process for identifying curves that "best" fit a data set. This essay explains the process of least squares regression for fitting several types of curves (linear, power, exponential) to data sets. Also included are general guidelines for selecting which type of function to use, as well as a list of key issues to be aware of when fitting data.

## Introduction

Processing and interpreting experimental data is one of the most vitally important steps in the process of scientific research. Once data is collected, researchers are typically concerned with questions such as:

1. Does the data look "believable" in the sense that there are no obvious signs of faulty instrument calibration or poor experimental design/protocol? For example, if temperature recordings are measured from a sample that is moved from an oven into a freezer and the data indicates that the temperature increases, then clearly something is wrong.
2. How might one quantify relationships between variables involved in a data set?
3. Once a relationship between variables is quantified, how can that information be used to make predictions beyond what may be experimentally measurable (interpolation and extrapolation)?

In what follows, assume that the first question is answered in the affirmative, as there is no point in trying to draw conclusions from a "bad" data set. The statistical procedure known as *regression analysis* specifically addresses the other questions. This article covers the basics of regression analysis for data sets involving *one* independent variable (usually time, $t$) and *one* dependent variable. Although there are well-established techniques for handling multiple independent and dependent variables, the relevant mathematics is more technical. One goal of a regression analysis is to generate the *regression function*, the graph of which is a smooth curve that is designed to stay as close as possible to the data points in a "holistic" sense to be made precise below. Regression functions allow one to interpolate and extrapolate data by choosing some test value of the independent variable, substituting it into the formula for the regression function, and using the result to estimate the corresponding value of the dependent variable.

*Email: jcain2@richmond.edu

DOI 10.1007/978-1-4614-6436-5_560-1

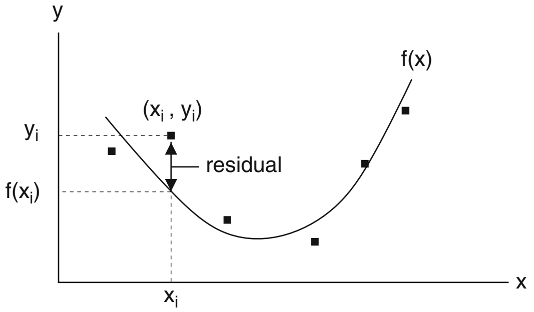

**Fig. 1** Schematic illustration of a regression function $f(x)$ fit to data points (*solid squares*). The residual error is shown for one of the data points $(x_i, y_i)$

## Least Squares Regression and Three Special Cases

Consider, for example, an experiment involving a simple decay process $A \rightarrow \star$ in which some chemical species $A$ decays to an inert product, and suppose that the concentration of $A$ is measured at various times. To establish notation, let $t_1, t_2, \ldots t_n$ denote the times at which the measurements were recorded, and let $c_i$ denote the concentration of $A$ recorded at time $t_i$. The data can be rendered graphically by plotting all of the ordered pairs $(t_i, c_i)$; here the ordering suggestively indicates that concentration is regarded as the dependent variable and time as the independent variable. The purpose of least squares regression can be stated as follows: given some class of mathematical function (e.g., linear, exponential, sinusoidal, etc.) which is believed to mimic the data set, identify the *specific* function within that class "best" fits the data points in a sense to be made precise in the next paragraph. For the decay process described above, it is natural to try a simple exponential function, one of the form $f(t) = \beta e^{mt}$ where $\beta$ and $m$ are *parameters*. In that case, least squares regression should produce the specific values of $\beta$ and $m$ for which the exponential function fits the data set "optimally."

The general process of least squares regression for data involving a single independent variable $x$ and a single dependent variable $y$ is as follows. Given data points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ and a function $f(x)$ defined for all possible choices of $x$, define the residual sum of squares (*RSS*) as the quantity

$$RSS = \sum_{i=1}^{n} [y_i - f(x_i)]^2.$$

The goal of least squares regression is to identify the specific choice of $f(x)$ (from within some class of mathematical functions) for which the RSS is minimized.

It is not terribly difficult to understand the rationale behind minimizing the above sum, particularly with some graphical intuition (Fig. 1). The quantity $y_i - f(x_i)$ is called a *residual error* and measures the vertical separation between the data point $(x_i, y_i)$ and the point $(x_i, f(x_i))$ on the graph of $f(x)$ for the same choice of the dependent variable (i.e., $x = x_\mathrm{i}$). Squaring a residual always returns a *nonnegative* quantity regardless of whether $f(x_i)$ is an underestimate or overestimate of $y_i$, and $[y_i - f(x_i)]^2$ increases with the amount of separation between $y_i$ and $f(x_i)$. The main purpose of squaring the residuals (as opposed to not doing so) is to prevent "cancellations" of errors when $f(x)$ underestimates some data points but overestimates others; i.e., the sum should impose a penalty whenever $f(x)$ deviates from the data, no matter how that deviation occurs. More concretely, consider a data set with just two points. If one draws a line which overestimates one point by 1,000 and underestimates the other point by 1,000, it would be inappropriate to declare the fit perfect on the basis that the errors cancel out. The RSS offers a holistic indicator of the goodness of fit between $f(x)$ and the actual data, with larger RSS signifying larger error. Minimizing the RSS assures that, at least in some aggregate sense, the graph of $f(x)$ ought to remain "close" to the data for all values of the independent variable $x$ at which measurements were provided.

There are three special classes of regression functions $f(x)$ that are worth singling out: linear, exponential, and power. Those functions model a wide variety of natural phenomena and have the rare advantage of precise formulas for minimizing the RSS error.

## Linear Regression

Lines of the form $f(x) = mx + b$ are the simplest class of functions commonly used to fit experimental data. By varying the parameters $m$ (the slope) and $b$ (the intercept), it is possible to represent every possible line of finite slope, and the purpose of linear least squares regression is to find the choices of $m$ and $b$ that minimize the RSS for a given data set $(x_i, y_i)$, $i = 1, 2, \ldots n$. It happens that there is an exact formula for these optimal choices of $m$ and $b$: let $\overline{x}$ and $\overline{y}$ denote the mean (average) values of the values $x_i$ and $y_i$, respectively, for a given data set. Then the regression line for the data set is achieved by choosing

$$m = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

and

$$b = \overline{y} - m\overline{x}.$$

These formulas are obtained by differentiating the RSS sum with respect to the parameters $m$ and $b$ (separately), setting the two results equal to 0, and solving the resulting system of two equations for $m$ and $b$. For readers familiar with statistical parlance, notice that the slope $m$ of the regression line is the ratio of the covariance of the data sets $x_i$ and $y_i$ to the variance of the data set $x_i$. An example illustrating the computation of a regression line for a sample data set is provided in the next section.

## Exponential Regression

An exponential function is one of the form $f(x) = \beta e^{mx}$, where $\beta$ and $m$ are parameters, and can simulate growth or decay according to whether $m$ is positive or negative. (Note: If the "natural" base $e = 2.71828\ldots$ feels unnatural to the reader, converting an exponential function to another base such as 2 or 10 is not difficult and requires only minor modification of the following material.) Henceforth, assume that the parameter $\beta$ and the measurements $y_i$ are positive, a reasonable assumption in biochemistry contexts in which the dependent variable is a (nonnegative) chemical concentration. This assumption is not needed for general exponential regression but allows avoidance of technical issues when taking logarithms in what follows. The problem of determining the best-fit exponential function to a data set $(x_i, y_i)$, i.e., finding the optimal $\beta$ and $m$, can be reduced to a problem of linear regression. By taking the natural logarithm of $f(x) = \beta e^{mx}$ and defining $b = \ln \beta$, one obtains

$$\ln f(x) = mx + b,$$

**Table 1** Data showing 11 measurements of chemical concentration (mmol/L) during a decay process over a 5-min span. The last row shows the natural logarithm of the concentrations in the middle row, for use in exponential regression

| Time (min) | 0.0 | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Conc (mmol/L) | 7.72 | 5.39 | 4.98 | 3.80 | 2.69 | 2.41 | 2.06 | 1.27 | 1.29 | 0.812 | 0.603 |
| ln(conc) | 2.04 | 1.68 | 1.61 | 1.34 | 0.990 | 0.879 | 0.721 | 0.239 | 0.252 | −0.208 | −0.505 |

a form that looks familiar from the above discussion of linear regression. Consequently, if one defines $Y_i = \ln y_i$ and forms a new data set $(x_i, Y_i)$, the problem of fitting an exponential function $\beta e^{mx}$ to the original data set is reduced to the problem of fitting a line $mx + b$ to the new data set. The optimal choices of $m$ and $b$ are given by the formulas from the linear regression case above, after which one may compute $\beta = e^b$. Incidentally, a plot of the points $(x_i, Y_i) = (x_i, \ln y_i)$ is called a semilog plot of the original data set.

Table 1 shows sample data of concentration (Row 2 in the table) over time (Row 1) for a simple chemical decay process. The natural logarithms of concentrations appear in Row 3 and are included for the purpose of fitting a regression line to the semilog plot. The left panel of Fig. 2, a semilog plot, includes a graph of the best-fit regression line ($m = -0.488$, $b = 2.04$) as computed by the formulas appearing in the linear regression section above. The right panel of the figure shows the corresponding exponential regression function $f(x) = \beta e^{mx}$, where $\beta = e^b = 7.69$.

Using the regression function $f(x) = \beta e^{mx} = 7.69e^{-0.488x}$ to interpolate and extrapolate the data in Table 1 is straightforward. To estimate the concentration at time $t = 1.25$ min (interpolation), simply calculate $f(1.25) \approx 4.18$ mmol/L. Similarly, one may extrapolate the concentration at time $t = 8.0$ min by computing $f(8.0) \approx 0.155$ mmol/L.

## Power Regression

A power function is one of the form $f(x) = \beta x^m$ where $\beta$ and $m$ are parameters. The important distinction between power and exponential functions is that, in the former, the base is variable and the exponent is constant, whereas in the latter, the base is constant and the exponent is variable. Much like the process of exponential regression, determining the best-fit power function to a data set can be distilled to a problem involving linear regression. To avoid similar technicalities to those mentioned previously, suppose that the data set $(x_i, y_i)$ is such that both coordinates are positive. Taking logarithms of both sides of the relationship $f(x) = \beta x^m$ and defining $b = \ln \beta$, one obtains

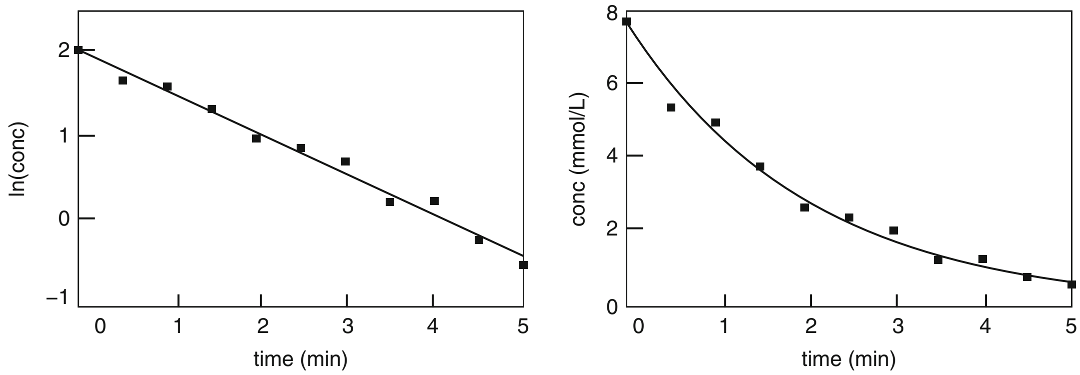

**Fig. 2** *Left*: Semilog plot (Row 3 vs Row 1 of Table 1) along with the least squares regression line. *Right*: Exponential regression fit to concentration-versus-time data (Row 2 vs. Row 1 from Table 1)

DOI 10.1007/978-1-4614-6436-5_560-1

$$\ln f(x) = m \ln x + b,$$

implying that $\ln f(x)$ depends linearly on $\ln x$. Hence, if one transforms the original data set $(x_i, y_i)$ by defining new variables $X_i = \ln x_i$ and $Y_i = \ln y_i$, then the problem of finding the best-fit power function $\beta x^m$ to the original data is reduced to the linear regression problem for the points $(X_i, Y_i)$. The optimal choices of $m$ and $b$ are then provided by the above formulas in the linear regression section, after which one may compute $\beta = e^b$. A plot of the points $(X_i, Y_i) = (\ln x_i, \ln y_i)$ is called a log-log plot of the original data set. As an example of when power regression might be appropriate, consider a simple, nonreversible kinetic process $2A \rightarrow B$ in which two molecules of species A combine to form species B. Then the law of mass action would suggest that the rate of accumulation of B (say, in moles per unit time) depends quadratically on the number of moles of A. Hence, given data points $(x_i, y_i')$ where $\mathrm{x}_i$ and $\mathrm{y}_i'$ denote the concentration of A and rate of accumulation of B (respectively) at the ith measurement, one may fit a power function $\beta x^m$ by fitting a regression line to the log-log plot of the data set. If the law of mass action applies, one would expect the regression process to produce an optimal exponent of $m \approx 2$ in this example, while the optimal choice of $\beta$ represents an approximation of the kinetic constant for the process.

## Choosing Functions and Parameters

The above are the most common types of functions fit to experimental data sets, but there are many other possibilities. The general guiding principle when selecting which class of functions to use is this: make reasonable choices that can be justified based upon scientific principles and/or mathematical models of the presumed dynamics of the system (Mathematical Models in the Sciences). For a simple decay process, an exponential function $f(x) = \beta e^{mx}$ is probably appropriate. For a saturation process, a sigmoidal function (S-curve; see example below) may suffice, or alternatively, a function of the form $f(x) = \alpha - \beta e^{mx}$, where $\alpha$, $\beta$, and $m$ are positive constants. Fitting data involving biochemical oscillators might involve sinusoidal functions; however, such fits are far less common in the literature, largely because the underlying dynamics of oscillatory reactions are too complex to be described using simple combinations of (scaled) sine and cosine functions.

Here are a few other remarks to bear in mind:

1. **Optimization can be computationally difficult**: The luxury of precise mathematical formulas for optimal parameter choices (as in the linear, power, and exponential cases above) is a rarity. No such formulas exist for general classes of regression functions, in which case performing least squares fits may require sophisticated computer algorithms.
2. **Computer software**: Of course, statistical and mathematical exploration of a data set is facilitated through use of a computer. There are several major statistical software packages available for performing least squares regression (among other things), some of which are available cost-free. Many of them allow the user to select from a variety of function classes, including the ones described above.
3. **Number of parameters**: Increasing the number of parameters (degrees of freedom) in the class of functions being used to fit data vastly complicates the computation of a regression curve (even for a computer). For example, suppose that one wishes to fit a sigmoidal function (S-curve) of the form

DOI 10.1007/978-1-4614-6436-5_560-1

$$f(x) = \alpha + \frac{\beta}{1 + \exp((x - \gamma)/\delta)}$$

to a data set describing a saturation process. The four parameters in this class of functions offer considerable flexibility: $\alpha$ shifts the graph of $f(x)$ vertically, $\beta$ stretches the graph vertically, $\gamma$ shifts the graph horizontally, and $\delta$ stretches the graph horizontally. If a computer were to (naively) test a paltry 10 different values of each of those four parameters and calculate the RSS for each parameter set, there would be $10^4 = 10{,}000$ parameter sets to check. By contrast, the linear, exponential, and power regression functions described above each include only two parameters, so that a similar "exhaustive" search of parameter sets requires testing only $10^2 = 100$ parameter pairs. For complicated models involving dozens of parameters, even the fastest computers could never complete an exhaustive check of every possible parameter set.

4. **Reasonable parameter choices**: Software that performs general least squares regression is often automated with mathematical algorithms for seeking parameters that will minimize the RSS. Whenever possible, it is helpful to specify acceptable ranges that each parameter is allowed to take, thereby preventing the computer from searching highly unrealistic parameter choices. Visual inspection of Rows 1 and 3 in Table 1 suggests that the slope of the regression line is negative but certainly not less than $-1$, implying that a restriction $-1 \leq m \leq 0$ could be appropriate. Of course, this is merely an illustration – for linear least squares regression, there would be no need to restrict the allowable ranges of $m$ and $b$, because there are precise formulas for the optimal choices of those parameters. Generally, the more parameters there are, the more important it is to supply reasonable ranges for each parameter, and the narrower the range the better. Some software actually prompts the user to input a reasonable initial guess for the best-fit parameter values.
5. **Size of the data set**: Let $N$ denote the total number of *distinct* values of the independent variable within a data set. Then, without getting into technical details, the total number of parameters in the class of functions being used to fit the data should not exceed $N$ . For example, fitting a line $f(x) = mx + b$ (*two* parameters, $m$ and $b$) to a single data point would not make sense.
6. **Visually inspect the goodness of fit**: When a computer determines the best-fit parameter set for a regression curve, be sure to graph the curve on the same set of axes as the data itself. On some level, data fitting is a subjective pursuit in which the scientist must check that the function truly captures the desired trends within the data. Here is an extreme example of what might go wrong if data fitting is performed carelessly: Given a set of data points $(x_i, y_i)$ where $i = 1, 2, \ldots n$ for which the independent variable values $x_i$ are all distinct, it is straightforward to construct a polynomial of degree at most $(n + 1)$ which crosses all of the data points. However, the fact that such polynomials are called interpolating polynomials is something of a misnomer – they are often useless for the purpose of interpolation, oscillating wildly between consecutive data pairs and failing to capture overall trends within the data.

Graphing data and regression curves can also aid in spotting outliers, isolated data points that are likely due to significant measurement error or, more rarely, scientific anomaly. Removing outliers from a data set can improve the goodness of fit but should always be accompanied with a disclosure that the fit was achieved by neglecting some of the data points.

Molecular Life Sciences
DOI 10.1007/978-1-4614-6436-5_560-1

## Further Reading

This overview of scientific data fitting is far from comprehensive and makes no attempt to describe the various technicalities and assumptions underlying regression analysis. For a deeper introduction to the relevant mathematics and statistics, refer to the texts of Chatterjee and Hadi (2012), Kutner et al. (2004), or Montgomery et al. (2006).

## Cross-References

▶ Mathematical Models in the Sciences

## References

Chatterjee S, Hadi AS (2012) Regression analysis by example, 5th edn. Wiley, Hoboken

Kutner M, Nachtsheim C, Neter J (2004) Applied linear regression models, 4th edn. McGraw-Hill/Irwin, Chicago

Montgomery DC, Peck EA, Vining GG (2006) Introduction to linear regression analysis, 4th edn. Wiley-Interscience, Hoboken