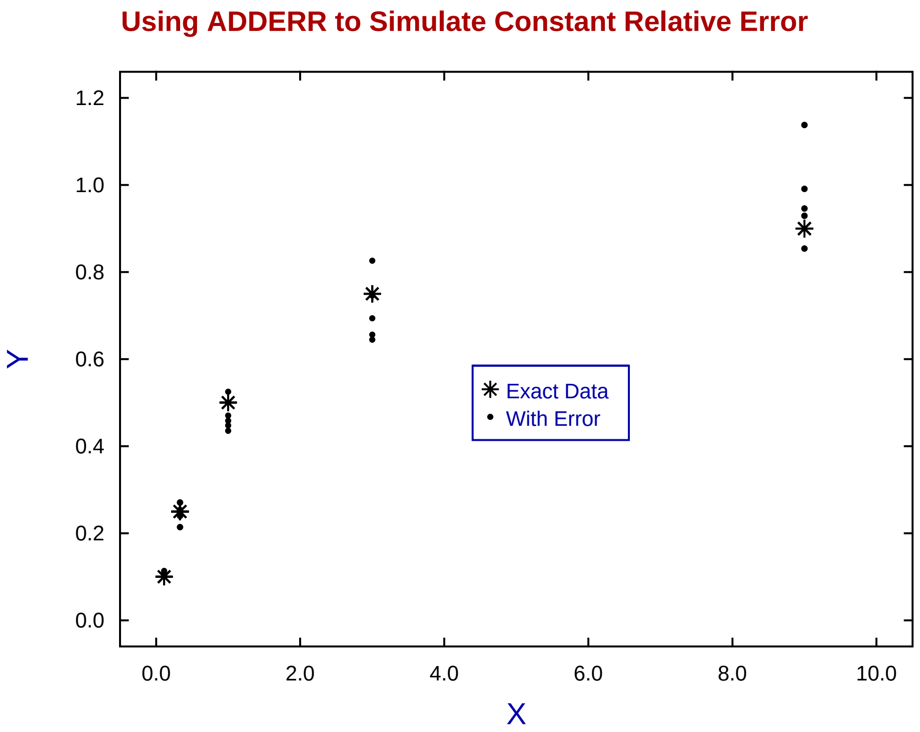
Tutorials and worked examples for simulation,
curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

To check the conclusions from model fitting it is advisable to simulate the model of interest, say $y = f(x)$, then study the robustness of parameter estimation and model discrimination with data simulated according to the assumed experimental design and variance model $V(y)$.

Example 1: Constant relative error, $V(y) = (\%y)^2$

From the main SimFIT menu choose [A/Z], open program **adderr**, then read in the test file `mmfit.tf1`. This has exact data for the Michaelis-Menten model generated by program **makdat**. After choosing the defaults to add five replicates with 7.5% constant relative error the following plot resulted.



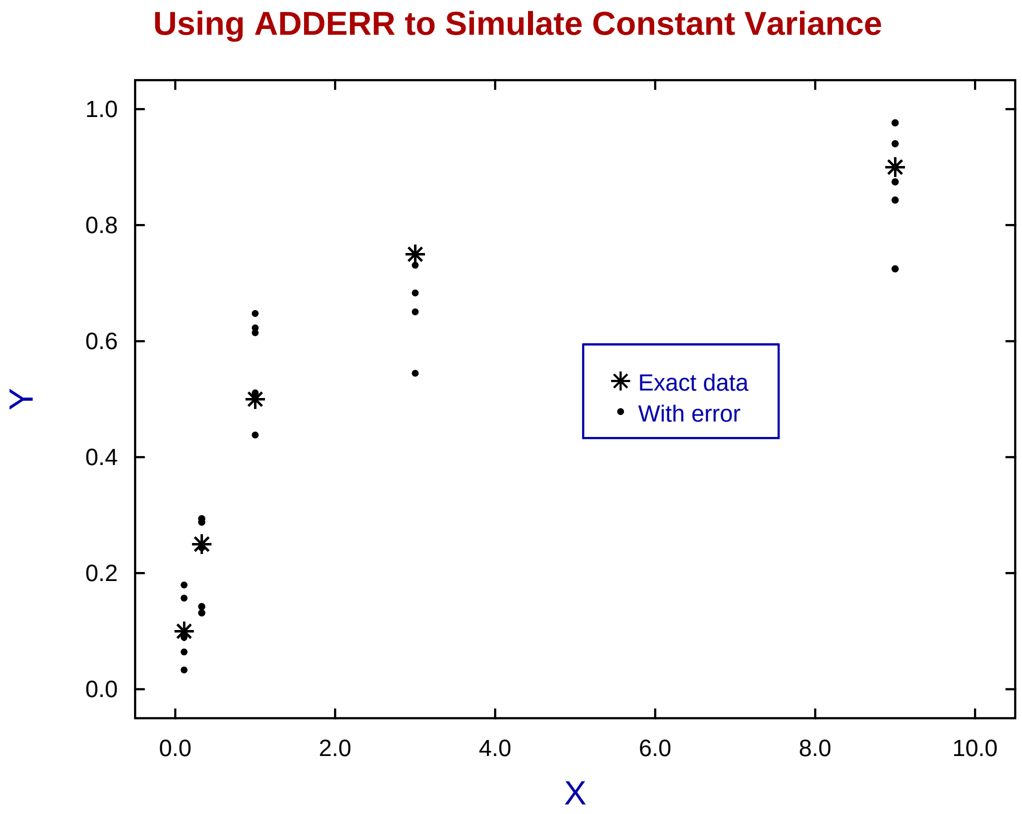
Constant relative error assumes that the observations O_i have experimental error which is a normal variate E_i with mean zero, and standard deviation equal to a percentage of the absolute true value $T_i = y(x_i)$, i.e.

$$\begin{aligned} O_i &= T_i + E_i \\ E_i &\approx N(\mu, \sigma_i^2) \\ \mu &= 0 \\ \sigma_i &= \%|T_i| \end{aligned}$$

Although this choice attempts to represent the reality that experimental error is proportional to response, it does tend to underestimate the error at low response so that the fitting may be dominated by small observations.

Example 2: Constant variance, $V(y) = k^2$

Note that program **adderr** can write a simulated data set to file containing the replicates along with weighting factors appropriate for the error type assumed. Actually most curve fitting is unweighted, that is with all weighting factors equal to one, which corresponds to assuming constant variance. The next plot illustrates adding error of the constant variance type to the exact data.



Here the observations O_i are assumed to result from adding a normal variate E_i with mean equal to zero and constant variance to the true value $T_i = y(x_i)$, i.e.

$$\begin{aligned} O_i &= T_i + E_i \\ E_i &\approx N(\mu, \sigma^2) \\ \mu &= 0 \\ \sigma &= k, \text{ a fixed positive constant} \end{aligned}$$

It will be seen from this plot that assuming constant variance presumes that all data points are of equal importance, which tends to bias towards a fit dominated by the largest observations.

In order to fit a model to data it is necessary to understand the type of weighting required, and this can only be known by performing replicate observations in order to discover the best statistical model for the error variance. Program **adderr** provides many options for simulating experimental error as will now be summarized.

Theory

Weighting curve fitting data

Curve fitting is undertaken when n observations y_i have been made in the belief that the true model is known to be a function of independent variable(s) x and some unknown parameters Θ, but contaminated by experimental or observational error ϵ_i. The intention is to obtain meaningful estimates for the model parameters in order to interpret the observations in the light of some basic scientific principles.

A common approach is to appeal to the principle of maximum likelihood and assume that the observations are the sum of a deterministic model plus random error which is normally distributed with mean zero and standard deviation σ_i as in

$$\begin{aligned} y_i &= f(x_i, \Theta) + \epsilon_i \\ \epsilon_i &\approx N(0, \sigma_i^2). \end{aligned}$$

In this case the maximum likelihood estimate for the parameters Θ is obtained by minimizing the weighted sum of squares $WSSQ$ given by

$$WSSQ = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i, \Theta)}{\sigma_i} \right)^2 .$$

This poses a serious problem since the true standard deviations σ_i can never be known and must be replaced by some estimates. The aim of simulating experimental error is to examine the way that the accuracy with which the parameters Θ can be estimated depends on the choice of approximations s_i to the σ_i.

There are essentially four ways to choose s_i values.

1. Set all $s_i = 1$
   This is reasonable if the data are very noisy, or the range of values of the observations is relatively small.

2. Estimate s_i using replicates
   This reasonable if the number of replicates is sufficiently large (say at least 5) to make a sensible estimate for σ_i

3. Estimate $s_i = g(y_i)$
   This results in weights not being constant at fixed x.

4. Estimate $s_i = g(\hat{f}(x_i, \hat{\Theta}))$
   This results in weights not being constant at fixed x but has the added complication that the weights change at each iteration.

Method 1 assumes constant variance and uses the value of $WSSQ/NDOF$ at the solution point to estimate the variance, where $NDOF$ is the number of degrees of freedom, i.e. number of observations minus the number of parameters estimated. The other methods assume that the s_i supplied are proportional to the true σ_i, then use the value of $WSSQ/NDOF$ at the solution point to estimate the square of the proportionality factor. The choice of function $g(.)$ is often a variant of

$$g(t)^2 = A + Bt^{\lambda}$$

where A, B, and λ are either fixed or estimated from the data, and t can be taken as equal to the replicates or the best-fit function value.

Simulating experimental error

The output files from program **makdat** contain exact data for $y = f(x)$, which are then used to add random error to simulate experimental error. To do this, the output file then becomes an input file for program **adderr**. After adding random error, the input file is left unchanged and a new output file is produced as in this scheme.

$$\text{Model} \longrightarrow \textbf{makdat} \longrightarrow \boxed{\begin{array}{c}\text{Exact}\\ \text{data}\end{array}} \longrightarrow \textbf{adderr} \longrightarrow \boxed{\begin{array}{c}\text{Simulated}\\ \text{data}\end{array}}$$

There are numerous ways to use program **adderr**, including generating replicates. If in doubt, pick 7.5% constant relative error with 5 replicates, as this mimics many situations. Note: constant relative error cannot be used where $y = 0$ (which invokes a default value).

The options available with program **adderr** are as follows.

1. Single measurements: constant relative error
2. Single measurements: fixed constant variance
3. Single measurements: mixed power law error
4. Generate replicates: constant relative error
5. Generate replicates: fixed constant variance
6. Generate replicates: mixed power law error
7. Choose from selection of error distributions
8. Just add outliers to the data set supplied

In options 3 and 6 the model for the variance of observations is the mixed power law

$$V(y) = \sigma_0^2 + \sigma_1^2 y^2,$$

so that the error resembles constant variance white noise at low response levels with a transition to constant relative error at high response levels. Constant variance ($\sigma_1 = 0$) fails to account for the way variance always increases as the signal increases, while constant relative error ($\sigma_0 = 0$) exaggerates the importance of small response values. Note that using program **adderr** you can also simulate the effect of outliers or use a variety of error generating probability density functions, such as the Cauchy distribution which is a often a better model for experimental error.

When calculating variance $V(y)$ as a function of y it is possible to use the true y as the argument (option 1), or to use the mean of the simulated observations (option 2) in program **adderr**, but there are further points to be made.

- There are some experiments where the observations can never be negative, e.g. the size of a population or weight of a crop yield. Program **adderr** draws attention to such events and gives options that always result in positive simulated observations.
- Outliers are often encountered, that is observations that are not typical observations, and program **adderr** also allows this to be simulated. However many would argue that the best way to deal with outliers is for the experimentalists to view the scatter of observations then de-select extreme values as required.