Module CH923
Statistics for Data Analysis

Relationships Between Variables

Introduction

A range of techniques are available that can be used to model, or describe, the relationship between an observed response variable and one or more explanatory variables. There are a number of reasons why we might wish to do this, but the two principal ones are:

i) To investigate and test hypothetical mathematical models for biological systems,

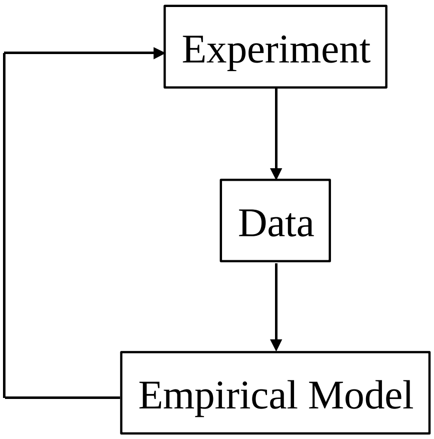
ii) To predict the values of one variable from another.

Often a particular investigation may stem from both of these motives. Examples of relationships that may need to be investigated are:

(a) The response of a crop to varying amounts of several fertilisers – the objectives may be to establish the form of the relationship, or to predict the optimum combination of fertilisers;
(b) The relationship between various meteorological measurements and crop yield – the most obvious objective would be to try and understand possible causative mechanisms of meteorological effects on crop growth;
(c) The effectiveness of different rates of insecticide in killing a particular insect – here we are probably interested in identifying the rate required to kill a certain proportion of the insects;
(d) The relationship between leaf area and leaf weight for several varieties of a plant and various ages of leaves for each plant – in this situation we could be interested in predicting leaf area, an important variable in photosynthesis, which is difficult to measure, using leaf weight, which is relatively simple to measure.

The simplest form of relationship between two variables is a straight line, and this is known as simple linear regression.

Types of model



Before we develop the basic ideas behind regression modelling, it is worth spending a few moments considering the different types of model that we might meet. Our primary emphasis will be on *empirical modelling*, although probability and statistics have a role to play in some of the other types.

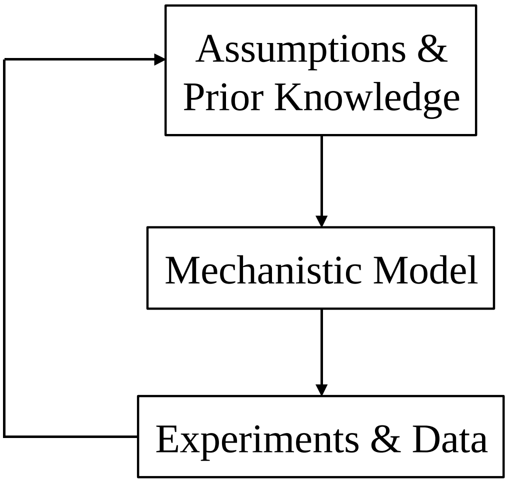
It should be stressed that the model types that we define below are not as clearly defined as they may appear. For example **mechanistic** models, which purport to describe reality using *mechanisms* will always involve *empiricism* at some stage of their construction. It is often convenient, however, to classify models using a combination of the descriptors below.

- **Empirical or Mechanistic**

Empirical models set out principally to describe; **mechanistic** models attempt to give a description with understanding (frequently based on differential equations).

More generally, **empirical** models are developed from experimental data – they describe the results of experiments, and the process of collecting experimental data and improving an empirical model may have to be repeated a number of times before an entirely satisfactory (or general) model can be produced.

By contrast, **mechanistic** models are developed from prior knowledge, but require validation and parameter estimation using experimental data.



Examples

i) Empirical model

Crop yield versus weed density – a commonly used relationship to describe the relationship between crop weight and changing weed density is the *rectangular hyperbola* ie.

$$\text{Crop Yield} = \frac{b}{1 + d * (\text{WeedDensity})}$$

The form of the relationship is not based on any knowledge of the mechanism, but has simply been observed to describe the shape of the relationship for similar data collected previously.

ii) Mechanistic Model

An exponential growth model for a bacterial culture. Denoting the number of bacteria at time t by N(t), a simple model can be expressed as

$$\frac{dN}{dt} = kN$$

When solved this leads to the *exponential growth model*, ie

$$N(t) = N(0) * e^{kt}$$

This model involves a simple mechanistic assumption that each individual reproduces at a constant rate. Experiments are needed to *verify* that this relationship holds, and to estimate *k.*

- **Deterministic or Stochastic**

A **deterministic** model makes definite predictions for quantities (such as crop yield, or rainfall); in contrast a **stochastic** model contains random elements so that it will predict, for example, the mean yield, but also predicts the potential variability of the yield.

Examples

i) Deterministic

The exponential growth model above is a deterministic model, as it makes definite prediction for the number of bacteria at time t. This is a deterministic model because it deals with very large populations, so that the random component is insignificant.

ii) Stochastic

A typical example is insect dispersal, where insects have an initial spatial distribution from which they disperse. In this case the population is small, so it is not possible to predict the *exact* spatial distribution of insects after a given interval of time. However we can model the *expected* spatial distribution of insects, and the variability of each point estimate.

- **Static or Dynamic**

A **static** model does not involve time, so assumes that the processes reach equilibrium fairly rapidly; a **dynamic** model describes the behaviour of a biological system through time.

Examples

i) Static

The forces acting on the structural members of a farm building.

ii) Dynamic

The exponential growth model above predicts the number of bacteria at any time t.

The role of statistics in *mechanistic* modelling

A mechanistic model is likely to contain *parameters* whose values are not explicitly known; these parameters need to be estimated using experimental data. For example, for the simple exponential model of growth, the growth rate, *k*, needs to be estimated from experimental results. Usually the models are considerably more complex than this simple model.

Within the development of a mechanistic model, the following activities will use statistical analysis:-

- Estimation of model parameters from experimental data.
- Estimating the precision of the parameter estimates (ie the standard error of the parameter estimates, or confidence regions for them).
- Making inferences using the parameters (ie is a particular parameter different from zero; is

a particular variable important in the model?)

- Estimating the precision of model predictions.
- Model validation – does the "best-fit" model describe the experimental results adequately? Is there evidence that the model "behaves" wrongly?

These activities require the use of **regression** techniques, both **linear**, and **non-linear**, as will be described in more detail later.

Approaches in *empirical* modelling

The aim of an **empirical model** is to describe experimental results. An empirical model will usually enable large sets of data to be reduced to a parsimonious description consisting of one or more *biologically sensible* equations with *biologically meaningful* parameters. Empirical models can also describe complex interactions.

In general the procedure followed can be described in three steps:

(a) **Model specification**

Decide on one or more equations that give a "sensible" description of the experimental results. These are chosen from past experience (both yours, and what appears in the literature) and/or by plotting the experimental data and comparing the shape of response to a number of standard curves.

(b) **Parameter estimation**

Fit the equations to estimate the unknown parameters of the equations (eg with the rectangular hyperbola the parameters B and d are estimated).

Statistical technique: Regression (linear & non-linear)

(c) **Model simplification**

There may be other factors affecting what is going on (eg in studies of crop yield versus weed density another experimental factor may be included such as different herbicide doses). Are some of the model parameters unaffected by treatment? (ie is the weed competitivity affected by dose?).

Statistical technique: Parallel curves/lines – discriminating between nested models.

Fitting models - parameter estimation

Model fitting is an essential requirement for any of the techniques outlined above – the parameter estimation and inference needed for mechanistic modelling, and the data summarising, parameter estimation and model simplification required for empirical modelling.

There are three steps to the model fitting process:

i) **Choosing the correct distribution of the predicted variable**

Empirical models are looking for relationships between a *predicted*, or *dependent*, variable (eg crop yield) and *explanatory*, or *independent*, variables (eg weed density). The *distribution* of the dependent variable needs to be taken account of in any fitting process so that the "best" parameter estimates are obtained, and that valid inferences are drawn when investigating model simplification. Some common distributions are:

- The Normal distribution - the most widely appropriate distribution.
- The Poisson distribution - used for data consisting of counts, ie insects caught in a field trap.
- The binomial distribution - used when data consists of proportion of individuals in a small group responding to a stimulus, eg insect mortality in probit analysis.
- The multinomial distribution - used when each individual in a small group can be classified into one of several groups.

ii) **Choosing an appropriate equation**

This has been covered in the previous section (*Model specification*).

iii) **Fitting the equation**

Model fitting generally uses a technique called *maximum likelihood*, which estimates the model parameters to be those which are most likely given the observed data. When the predicted variable being modelled is Normally distributed this reduces to a technique called *least squares*, which will be presented in more detail later. Combinations of certain probability distributions for the predicted variable with certain predictive equations allow model parameters to be estimated using a special case of maximum likelihood, and are called *generalised linear models*, more details of which will be presented later.

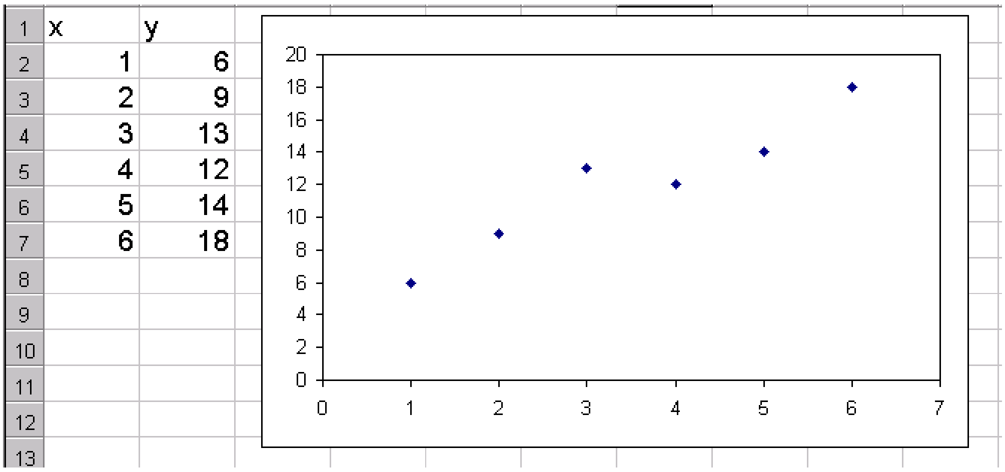
Simple linear regression

The method of *simple linear regression* is concerned with linear relationships between just two variables. The word **simple** here does not mean **easy**, but is used to contrast with *multiple linear regression* which involves three or more variables. The term **linear** in the name refers to the relationship between the response variable and the parameters, but in this simplest case we are also only interested in fitting a straight-line relationship between the two variables.

Finding the best line

In the simplest situation we will have a series of pairs of observations of y and x, where y, the **dependent variable**, is assumed to depend on x, the **independent variable**. As in designed experiments, we must assume that the units on which observations are made are variable. So, if we plot the values of y against the corresponding values of x, we do not get a series of points exactly on a straight line, but a scatter of points about an apparently straight line.

An artificial example of the kind of data we should expect to find is given in the Excel plot below, where the *x*-variable might be an applied level of nutrient, and the *y*-variable might be the increase in plant weight over a given period of time.



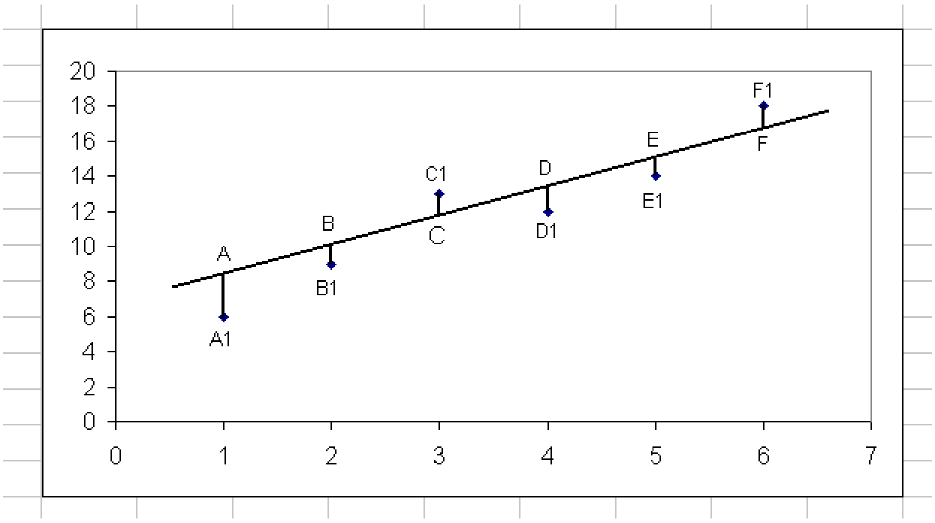
1	x	y
2	1	6
3	2	9
4	3	13
5	4	12
6	5	14
7	6	18
8		
9		
10		
11		
12		
13		

In order to choose a best line we must define a criterion for judging the adequacy of the fit of any particular line. The standard criterion used is that of minimising the sum of squared vertical deviations of the observed points about the line. This criterion is known as the principle of least squares (the same idea behind the analysis of variance of designed experiments). The use of squared deviations should not be entirely surprising, since this is the basis for variances – what we are doing is minimising the variance of the observations about the line.

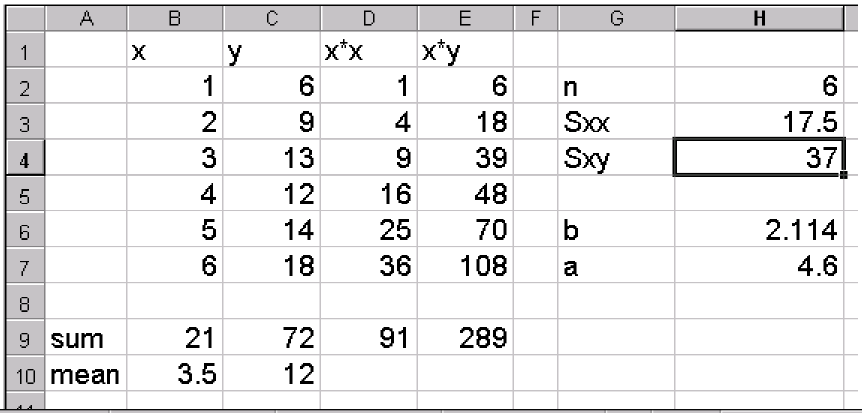
The criterion is illustrated in the plot below for an arbitrary line fitted by eye to the data. The measure of adequacy of the line A to F in fitting the points A1, B1, …, F1, is given by the sum

$$S_r = (\mathrm{A1} - \mathrm{A})^2 + (\mathrm{B1} - \mathrm{B})^2 + (\mathrm{C1} - \mathrm{C})^2 + (\mathrm{D1} - \mathrm{D})^2 + (\mathrm{E1} - \mathrm{E})^2 + (\mathrm{F1} - \mathrm{F})^2$$

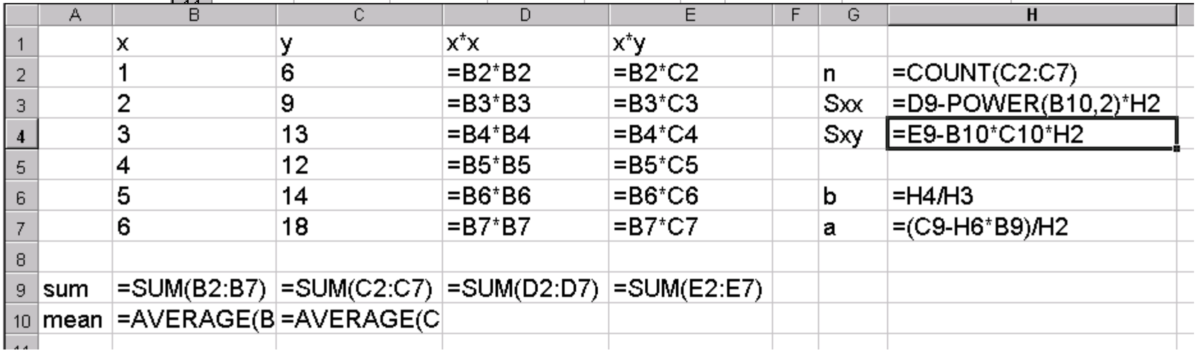
where S_r is referred to as the residual sum of squares, since it measures the residual variation of the *y* observations about the line, i.e. that part of the original variation between the *y* observations which is not attributable to the estimated linear dependence of *y* on *x*. The line which gives the smallest residual sum of squares, S_r, is defined to be the best fitting line.



We can calculate this criterion for any given line by calculating for each value of *x* for which there is an observation, the value of *y* on our given line, obtaining the difference between the observed and fitted line values of *y*, and summing the squares of these differences.



	A	B	C	D	E	F	G	H
1		x	y	x*x	x*y			
2		1	6	1	6		n	6
3		2	9	4	18		Sxx	17.5
4		3	13	9	39		Sxy	37
5		4	12	16	48			
6		5	14	25	70		b	2.114
7		6	18	36	108		a	4.6
8								
9	sum	21	72	91	289			
10	mean	3.5	12					



	A	B	C	D	E	F	G	H
1		x	y	x*x	x*y			
2		1	6	=B2*B2	=B2*C2		n	=COUNT(C2:C7)
3		2	9	=B3*B3	=B3*C3		Sxx	=D9-POWER(B10,2)*H2
4		3	13	=B4*B4	=B4*C4		Sxy	=E9-B10*C10*H2
5		4	12	=B5*B5	=B5*C5			
6		5	14	=B6*B6	=B6*C6		b	=H4/H3
7		6	18	=B7*B7	=B7*C7		a	=(C9-H6*B9)/H2
8								
9	sum	=SUM(B2:B7)	=SUM(C2:C7)	=SUM(D2:D7)	=SUM(E2:E7)			
10	mean	=AVERAGE(B	=AVERAGE(C					

Using the least squares principle of choosing the line to minimise the sum of squared deviations, we can express this sum algebraically as follows

$$S_r = \sum_i \left(y_i - \left(a + b\,x\right)\right)^2$$

where *a* and *b* are the intercept (value when *x* = 0) and slope of the fitted line respectively. It can be shown mathematically that the best fitting line has parameters *a* and *b* given by

$$b = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} \qquad a = \frac{\sum_i y_i - b \sum_i x_i}{n} = \bar{y} - b\,\bar{x}$$

where *n* is the number of observations of *y*. This line is the **regression** line of *y* on *x*. The numerator in the calculation of *b* is referred to as the corrected sum of products of *x* and *y* (written S_{xy}), and the denominator in this calculation is referred to as the corrected sum of squares of *x* (written S_{xx}). As in the calculation of the sample variance, these sums can be written in alternative forms for ease of calculation:

$$S_{xy} = \sum_i \left(x_i y_i\right) - n\,\bar{x}\,\bar{y} \qquad S_{xx} = \sum_i \left(x_i^2\right) - n\,\bar{x}^2$$

So, using these formulae, we can calculate the regression line for the artificial data on increase in plant weight against applied nutrient level plotted above (the Excel formulae are shown below).

Thus the fitted equation is $y = a + b\,x = 4.6 + 2.114x$

and, by setting x equal to 1, 2, 3, 4, 5 and 6 in this equation, we can obtain the fitted values of y .

Assessing the regression line

The first question we would expect to ask about the regression line we have just fitted is whether it represents a real relationship, i.e. is the slope of the assumed linear relationship genuinely different from zero? One way of answering this question is by considering the analysis of variance for fitting a regression line. We divide the total variation in the sample of y values into the residual variation about the fitted line (the sum we were minimising above to find the best line) and the variation between the fitted y values along the line. The latter variation can be thought of as the variation explained by, or attributable to, the regression of y on x, and is referred to as the regression sum of squares.

In practice the analysis of variance is obtained by calculating the total sum of squares

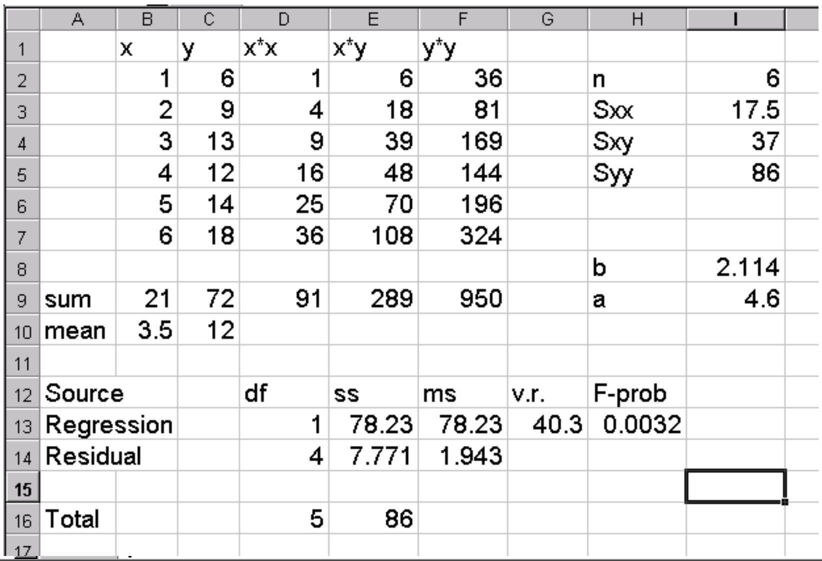
$$S_{yy} = \sum_i \left(y_i - \bar{y}\right)^2 = \sum_i \left(y_i^2\right) - n\bar{y}^2$$

and the regression sum of squares

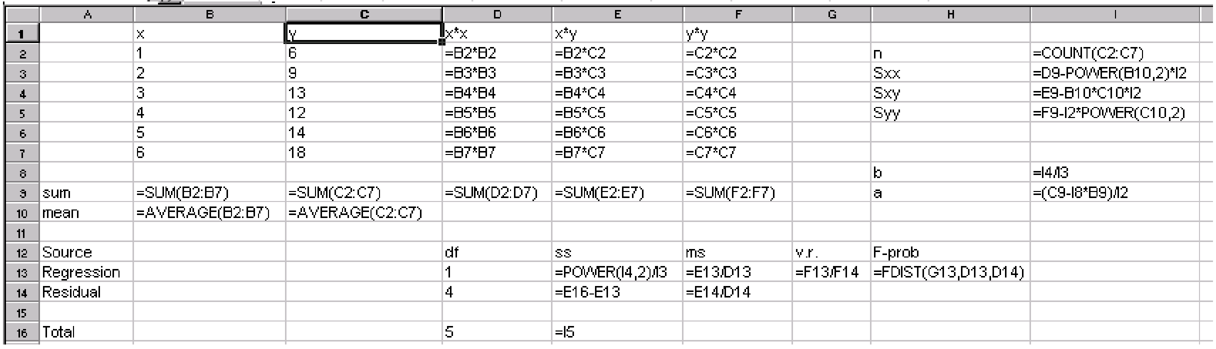
$$= \frac{\left(S_{xy}\right)^2}{S_{xx}} = \frac{\left[\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]^2}{\sum_i \left(x_i - \bar{x}\right)^2}$$

and then obtain the residual sum of squares by subtraction.

The division of the (n – 1) degrees of freedom for the total variation is 1 for the regression sum of squares and (n – 2) for the residual sum of squares. To understand why the residual degrees of freedom is (n – 2), remember that the degrees of freedom for a sample variance is (n – 1) because we are considering n deviations from a single value estimated from the data, the grand mean. For a regression line we a reconsidering n deviations from a set of values on a line, where the line is defined by two parameters, a and b, estimated from the data. As in the analysis of variance for designed experiments, the residual mean square is denoted by s^2 and provides an estimate of the variance of the observations having corrected for the effect of differences in the x-values, **assuming that the relationship between *y* and *x* is linear.** For the data on plant weight gains, the analysis of variance is constructed as shown below:



	A	B	C	D	E	F	G	H	I
1		x	y	x*x	x*y	y*y			
2		1	6	1	6	36		n	6
3		2	9	4	18	81		Sxx	17.5
4		3	13	9	39	169		Sxy	37
5		4	12	16	48	144		Syy	86
6		5	14	25	70	196			
7		6	18	36	108	324			
8								b	2.114
9	sum	21	72	91	289	950		a	4.6
10	mean	3.5	12						
11									
12	Source			df	ss	ms	v.r.	F-prob	
13	Regression			1	78.23	78.23	40.3	0.0032	
14	Residual			4	7.771	1.943			
15									
16	Total			5	86				
17									



	A	B	C	D	E	F	G	H	I
1		x	y	x*x	x*y	y*y			
2		1	6	=B2*B2	=B2*C2	=C2*C2		n	=COUNT(C2:C7)
3		2	9	=B3*B3	=B3*C3	=C3*C3		Sxx	=D9-POWER(B10,2)*I2
4		3	13	=B4*B4	=B4*C4	=C4*C4		Sxy	=E9-B10*C10*I2
5		4	12	=B5*B5	=B5*C5	=C5*C5		Syy	=F9-I2*POWER(C10,2)
6		5	14	=B6*B6	=B6*C6	=C6*C6			
7		6	18	=B7*B7	=B7*C7	=C7*C7			
8								b	=I4/I3
9	sum	=SUM(B2:B7)	=SUM(C2:C7)	=SUM(D2:D7)	=SUM(E2:E7)	=SUM(F2:F7)		a	=(C9-I8*B9)/I2
10	mean	=AVERAGE(B2:B7)	=AVERAGE(C2:C7)						
11									
12	Source			df	ss	ms	v.r.	F-prob	
13	Regression			1	=POWER(I4,2)/I3	=E13/D13	=F13/F14	=FDIST(G13,D13,D14)	
14	Residual			4	=E16-E13	=E14/D14			
15									
16	Total			5	=I5				

The significance of the regression is tested by comparing the ratio of the regression and residual mean squares with the F-distribution having 1 and (n-2) degrees of freedom. The hypothesis being tested is that there is no dependence of *y* on *x*, or, in other words, that the slope of the linear relationship is zero. In this case the variance ratio is very significant, so that there is clearly a trend of *y* increasing with *x*.

A simple summary statistic is the 'percentage variance accounted for', which expresses the reduction in variance due to the regression, and is given by the formula

$$\%\,\text{v.a.f.} = 100 * \frac{(\text{Total m.s.} - \text{Residual m.s.})}{\text{Total m.s.}}$$

It is closely related to the correlation coefficient, *r*, which is equal to the **coefficient of determination**, *R*, when used in a simple linear regression context. *R* is given by the formula

$$R^2 = \frac{(\text{Total s.s.} - \text{Residual s.s.})}{\text{Total s.s.}}$$

Both of these statistics measure the linear association between the response and explanatory variables. As a summary of regression, the percentage variance accounted for (also called the *adjusted* R^2 if divide by 100) is better because it measures variance rather than sums of squares, and because it is more useful in multiple linear regression. Both statistics are **relative** not **absolute**: their size is meaningful only in relation to the set of data being analysed. Very good regression models can have a small percentage variance accounted for if the range of the x-values is small; very bad regression models can have a large percentage variance accounted for if the fitted slope is steep.

Inferences about the slope of a line

The value we have calculated for *b* is our estimate of the rate at which *y* increases for a unit increase in the value of *x*. In some situations we may be directly interested in this rate – in our example we might be interested in whether the weight gain from increasing the level of applied nutrient is economically justifiable. We clearly need to know the precision of the estimate of the rate of increase, or slope. Since *b* is a linear combination of the *y* observations, we can show that the variance of *b* is

$$\mathrm{var}(b) = \frac{\sigma^2}{S_{xx}}$$

where σ^2 is the variance of the observations about the linear relationship between *y* and *x*. We estimate σ^2 by the residual mean square, s^2, and hence have a standard error for *b* given by

$$s.e.(b) = \sqrt{\frac{s^2}{S_{xx}}}$$

The general form of this standard is sensible. Smaller standard errors are obtained when S_{xx} is larger, or in other words, when the spread of *x* values is large. The greater the spread of *x* values, the better the line is defined.

Given our estimate of the slope, *b*, and its standard error, we can calculate a confidence interval for the slope of the population relationship between *y* and *x* in the same way that we can calculate a confidence interval for the population mean, μ, from a sample mean and its standard error. The 95% confidence interval for the slope of the true relationship of *y* on *x* is given by

$$b \pm t_{(0.05,n-2)} * \sqrt{\frac{s^2}{S_{xx}}}$$

where the *t*-value is the 5% point for the *t*-distribution with (*n* – 2) degrees of freedom.

For our example of weight gains and nutrient levels, the standard error of *b* is

$$s.e.(b) = \sqrt{\frac{s^2}{S_{xx}}} = \sqrt{\frac{1.94}{17.5}} = \sqrt{0.1108} = 0.333$$

and the 95% confidence limit for the slope is 2.114 ± 2.78 * 0.333, which is from 1.19 to 3.04.

We can also use the standard error of *b* to test the strength of the relationship between *y* and *x*. To test the hypothesis that there is no dependence of *y* and *x,* we calculate

$$t = \frac{b - 0}{s.e.(b)}$$

and compare this with the *t*-distribution on (*n* – 2) degrees of freedom. This *t*-test is exactly equivalent to the *F*-test calculated from the analysis of variance, the value of *t* (6.34) being the square root of the previously calculated variance ratio.

Predicting using a regression line

Consideration of the standard error of *b* leads to the idea of predicting the value of *y* for a given value of *x* and the precision of such a prediction. The predicted value of *y* for a specific *x*, say x_0, is obviously

$$y_0 = a + b\,x_0$$

So, from our example, the predicted weight gain for an applied nutrient level of 2 is

$$y_0 = a + b\,x_0 = 4.6 + 2.114 * 2 = 8.828$$

and we could calculated predictions for other *x*-values in a similar way. We should, however, be wary of predicting values of *y* for *x*-values much outside the observed range (**extrapolation**), since the precision of such predictions is likely to be poor.

Any error in the predicted value, $a + b\,x_0$, arises entirely from the errors in the estimates of the constant term, *a*, and the slope of the line, *b*. The standard error of the predicted value can be obtained, algebraically, as

$$s.e.(a + b\,x_0) = \sqrt{s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

where $\bar{x}$ is the mean of the *x* values used in the estimation of *a* and *b*. We must be careful to differentiate between this standard error and the standard error of the predicted weight gain of a single plant receiving x_0 units of applied nutrient. The expected weight gain of a single plant is still $a + b\,x_0$ but, in addition to the variation arising from the fact that this predicted value is only an estimate, we must take into account the variation of individual plants in the population about the true, or average, weight gain for the particular value of x_0. The standard error for the predicted value of a single plant is

$$s.e.(a + b\,x_0) = \sqrt{s^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

So, in predicting from our relationship between weight gain and applied nutrient level, the predicted mean weight gain for a large population of plants all being given an applied nutrient level of 2 units is 8.828, and the standard error of this predicted mean value is

$$s.e.(a + b\,x_0) = \sqrt{s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} = \sqrt{1.94 * \left[\frac{1}{6} + \frac{(2 - 3.5)^2}{17.5}\right]} = 0.757$$

However, the predicted weight gain for a single plant given an applied nutrient level of 2 units is also 8.828, but the standard error of this prediction is now

$$s.e.(a+b\,x_0)=\sqrt{s^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}=\sqrt{1.94*\left[1+\frac{1}{6}+\frac{(5-3.5)^2}{17.5}\right]}=1.585$$

From the formulae above, we can see that the standard error of a predicted value obviously depends on the value of *x* and is least when x_0 is equal to the mean of the observed *x*-values. As x_0 moves towards the extremes of the range of observed *x*-values, the standard error of a predicted value increases. Having calculated the standard error associated with any predicted value, we can obviously calculate a 95% confidence interval for the predicted value.

All the errors and standard errors of prediction mentioned so far arise from uncertainty about the values of *a* and *b* in the equation of the line, $y = a + b\,x$, and from the variation of individual values about the line. We have assumed throughout that the relationship is linear, but in practice we should always check this. There are methods available to check this but it always sensible to plot a graph of the data alongside any formal regression analysis.

The dangers of extrapolating a relationship beyond its known range of validity are hopefully well-known, and we have already seen how the standard error of a prediction increases as the value of x_0 moves away from the mean observed value. Doubts about the validity of the relationship at the extremes of the range of the data and beyond are additional reasons for caution in predicting value of *y* in these areas. Predictions for other populations will, of course, only be valid if the same values of *a* and *b* apply to these populations – this needs justification or, better still, verification.

Finally, in regression calculations, the independent variable, *x*, is generally assumed to be measured without observational error and, in some sense, to influence the values taken by the dependent variable, *y*. The values of *x* may be chosen by the experimenter or may, like the *y*-values, be a random sample. As an example of the latter situation, the pairs of *x* and *y* values might be the weights of the roots and shoots of a number of plants. If the values of *x* and *y* can both be treated as random samples representative of some larger population of values, then questions can arise about which, if either, should be treated as the dependent variable and which as the independent variable. The lines representing the regression of *y* on *x* and the regression of *x* on *y* are not the same – they minimize the sums of squares of the deviations about the line in two different directions, the *y*-direction and the *x*-direction respectively. Regression methods, known as functional regression, have been developed for coping with this type of problem.

Regression through the origin

A special case of the use of a simple linear regression model is in a calibration experiment. Here a machine is being set up to measure some attribute of a sample, with reference to standard measurements which are known to be correct. In such cases, it can be appealing to force the fitted regression line to pass through the origin, because the underlying science may make it clear that the two measured variables must be zero together. The model may indeed be fitted in this form, with the following equation:

$$y=b*x$$

with the parameter, *a*, fixed to be zero.

However, if we do fit models with no constant term, we should be aware of potential problems. The most important is that a model of this kind is based on the assumption that the relationship between the two variables is linear not only in the range of the observations, but also right down to the origin. If we have taken measurements only for a range of explanatory values some distance from the origin, we may have no evidence to check the validity of the assumption. There are many scientific processes that may look linear in a restricted range of values, but which behave very differently elsewhere and particularly for small measurements.

A second problem is that the need for a model with no constant inevitably suggests that the variance of small measurements may be much smaller than that of large measurements. This is something that need to be checked carefully before deciding to constrain the regression line in this way.

Lack-of-fit and Pure error

If we have replicated observations at each value of the explanatory variable, remember that we are then able to divide the residual variation into the *lack-of-fit* of the mean response at each value of the explanatory variable about the fitted line, and the pure error of the replicate observations about the mean response at each value of the explanatory variable. Then the regression mean square can be compared with the lack-of-fit term to assess whether there is a real relationship, and the lack-of-fit can be compared with the pure error to assess the goodness of fit. Note that in many packages it is not particularly straightforward to extract these separate components except by hand from two separate analyses.

Linear Regression with Groups (analysis of parallelism)

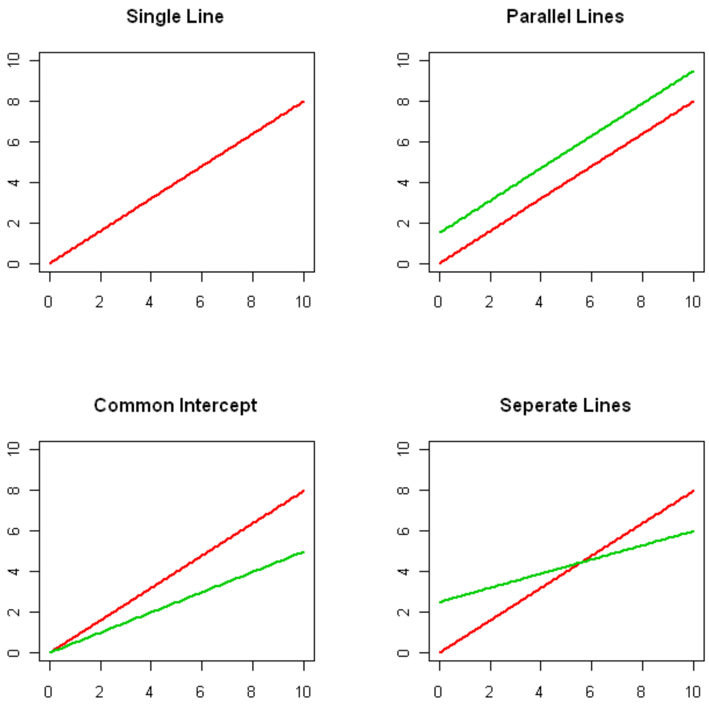
So far we have introduced the method of simple linear regression for modelling the relationship between one continuous response variable and one continuous explanatory variable. However, we may have two or more independent data sets involving the same variables, for each of which we have obtained a simple linear regression model, and be interested in comparing the regression models for the different data sets. Equivalently, in addition to observations on two continuous variables we may have observations on a third, usually discrete, variable. For example we might have sets of observations made in different years, on different sites, for different varieties, or under different treatment regimes. Again, one goal of the investigation may be to discover whether the relationship between the two continuous variables varies between groups. Our interest when analysing such data sets is therefore to test whether or not the same parameter values can be used for each group of observations, and to find the model that gives the simplest yet adequate description of the observed relationships. The method used to achieve this is usually referred to as **linear regression with groups** or **analysis of parallelism**.

Possible Models

The simplest situation for which linear regression with groups is required is where we have one grouping factor with just two levels. In this case four different models are possible, which we can write as fairly simple extensions of the model for simple linear regression given earlier. Descriptions of the four possible models, including their algebraic form, are given below, and the four models are also shown graphically.

(a) Coincident lines

The intercept and slope parameters both take the same values for both groups. Selection of this model to describe the observed data suggests that there is no effect of the grouping factor on the linear relationship between the two continuous variables. The equation for this model looks almost identical to that for simple linear regression, the only changes being the additional *j* subscripts for both the *x* and *y* variables:



$$y_{ij} = a + b\,x_{ij}$$

where *j* = 1 or 2 (denoting which group each observation is from, and *i* = 1...n_j, where n_j is the number of observations in group *j*.

(b) Parallel lines

The slope parameter takes the same value for both groups, but the intercept parameter is different. The interpretation of the parallel lines model is that there is a difference in response due to the grouping factor, but that this difference is unaffected by changes in the explanatory variable. In the equation for this model the parameter *a* gains a *j* subscript denoting that it is different for the two groups, with *i* and *j* defined as for the single line model:

$$y_{ij} = a_j + b\,x_{ij}$$

(c) Concurrent lines

The intercept parameter takes the same value for both groups, but the slope parameter is different. Selection of the concurrent lines model suggests that there is no effect of the grouping factor at the zero level of the explanatory variable, but that the effect of the grouping factor increases as the explanatory variable increases. In the equation for this model the parameter *b* gains a *j* subscript denoting that it is different for the two groups, with *i* and *j* defined as for the single line model:

$$y_{ij} = a + b_j\,x_{ij}$$

(d) General separate lines

Both the intercept and slope parameters take different values for the two groups. If this model is necessary to describe the observed data, then there is both an effect of the grouping factor at the zero level of the explanatory variable and a change in the size of this effect as the explanatory variable changes. In the equation for this model both parameters gain a *j* subscript, with *i* and *j* defined as for the single line model.

$$y_{ij} = a_j + b_j\, x_{ij}$$

These four models form two sequences of nested models. The single line model is a special case of the parallel line model (the intercepts are constrained to be the same) which is, in turn, a special case of the general separate lines model (the slopes are constrained to be the same). The single line model is also a special case of the concurrent lines model (the slopes are constrained to be the same), which is, again, a special case of the general separate lines model (the intercepts are constrained to be the same).

By fitting all four models to a data set and then considering the change in the residual sum of squares when moving from a more complex model (e.g. the parallel lines model) to a simpler model (e.g. the single line model) within each of the above sequences, we can determine which is the least complex model that provides an adequate description of the observed relationship.

For all four models the process of model fitting is as for simple linear regression, the 'best' values of the parameters being chosen using least squares, and the least squares fit being the best fit under the same assumptions as for simple linear regression.

The comparison of the different models within each sequence can be conveniently summarised within an accumulated analysis of variance, showing the variation explained by the simplest, single line, model, and the additional variation explained for each increase in model complexity (separate intercepts and then separate slopes, or vice versa).

Extensions

Whilst we have used a fairly simple example to develop this approach to linear regression with grouped data, the principles can easily be extended. For example, we may have a grouping factor with more than two levels, in which case the sequences of models remain the same but the numbers of parameters per model term increases. Alternatively we may have more than one grouping factor. In this case the number and complexity of model sequences increases, but we may still be able to select a single appropriate sequence to consider. If necessary, however, we can look at all possible sequences, and, starting from the most complex model (usually referred to as the ***full model***), drop unnecessary terms until we find the least complex model which still provides an adequate description of the observed data.

Multiple Linear Regression

We looked at one possible extension of simple linear regression, where we had observations on both the response and explanatory variables for two or more levels of some grouping factor. In our example this grouping factor was related to the presence or absence of Tridemorph. However, if this factor had three or more quantitative levels, such as a range of doses at which a chemical was applied, we might also be interested in the relationship between our response variable and this second possible explanatory variable. One possible approach would be to use the linear regression with groups approach described earlier, and then regress the fitted parameters for the different groups against the values of the grouping factor. An alternative, and more preferable, approach is to turn the grouping factor into a continuous explanatory variable, and regress the response variable against both explanatory variables. This approach can obviously be generalised to the case where we have two genuinely continuous explanatory variables. Linear models like this, with more than one explanatory variable are called ***multiple linear regression models***.

The model described above in which we regress the response variable on two explanatory variables can be thought of as being equivalent to the parallel lines model described in earlier. To produce a model that is similarly equivalent to the general separate lines model described earlier we need to include a term for the interaction between the two explanatory variables. This provides a term which allows the response to one of the explanatory variables to depend on the value of the other explanatory variable. This can be achieved by multiplying the two explanatory variables together to produce a third explanatory variable (in the following the interaction term produced by multiplying together variables x_1 and x_2 will be denoted x_{12}). We now have a multiple regression model with three possible explanatory variables.

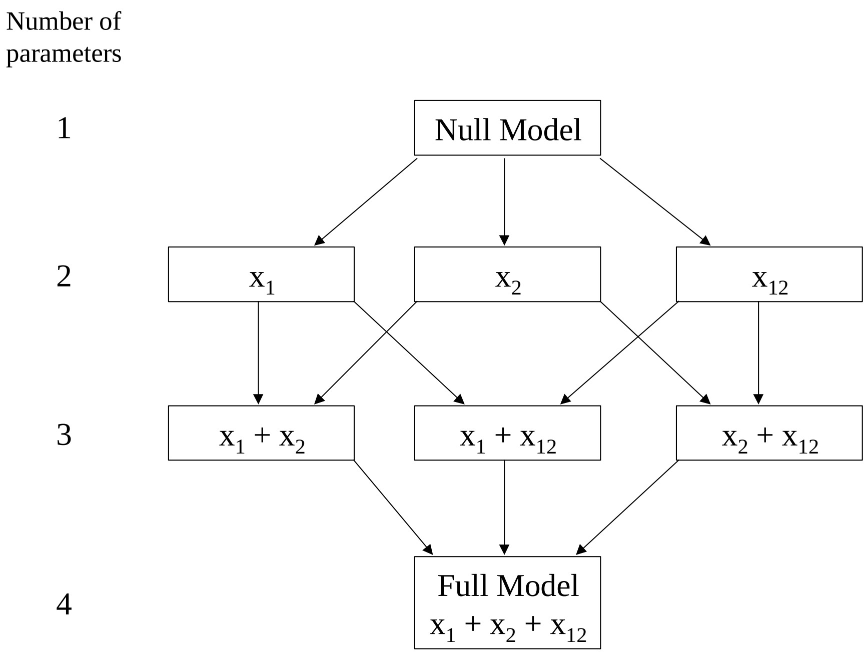
Model Selection

As with the linear regression with groups example we can set up sequences of possible models each starting with the simplest model, usual the ***null model*** in which we only fit the Constant term, and finishing with the most complex model, the ***full model***. To select the best model we could then consider the accumulated analysis of variance for each of the possible sequences. With only two possible explanatory variables this is relatively straightforward since there are only two possible sequences:

Sequence 1	Sequence 2
: Null Model	: Null Model
: x_1	: x_2
: $x_1 + x_2$ (Full Model)	: $x_1 + x_2$ (Full Model)

However, increasing the number of explanatory variables to three (as would be the case by including the interaction between the first two explanatory variables) increases the number of possible model sequences to six. Once the problem gets this big, we need to develop a more efficient approach to finding the best model.

With relatively few explanatory variables we can show the relationships between all the possible models diagrammatically, with model complexity (and the number of parameters) increasing from top to bottom, and with each model linked to those which are more complex than it, and of which it is a special case. For example, with three variables, x_1, x_2 and x_{12} as described above, we would get the diagram shown below.



With relatively few possible models, as in this example, it is practical to fit all possible models and calculate the change in the residual sums of squares associated with each link in the diagram. Comparing these differences with the estimate of the random variance obtained as the residual mean square for the full model allows us to test the effect of the additional term associated with each link. For example, the difference for the link between the null model and the model including just x_1 tests the effect of variable x_1 assuming no effect of either x_2 or x_{12}. Similarly, the difference for the link between the model including only x_1 and that including both x_1 and x_2 tests the effect of variable x_2 assuming an effect of x_1 but no effect of x_{12}.

Stepwise Regression Techniques

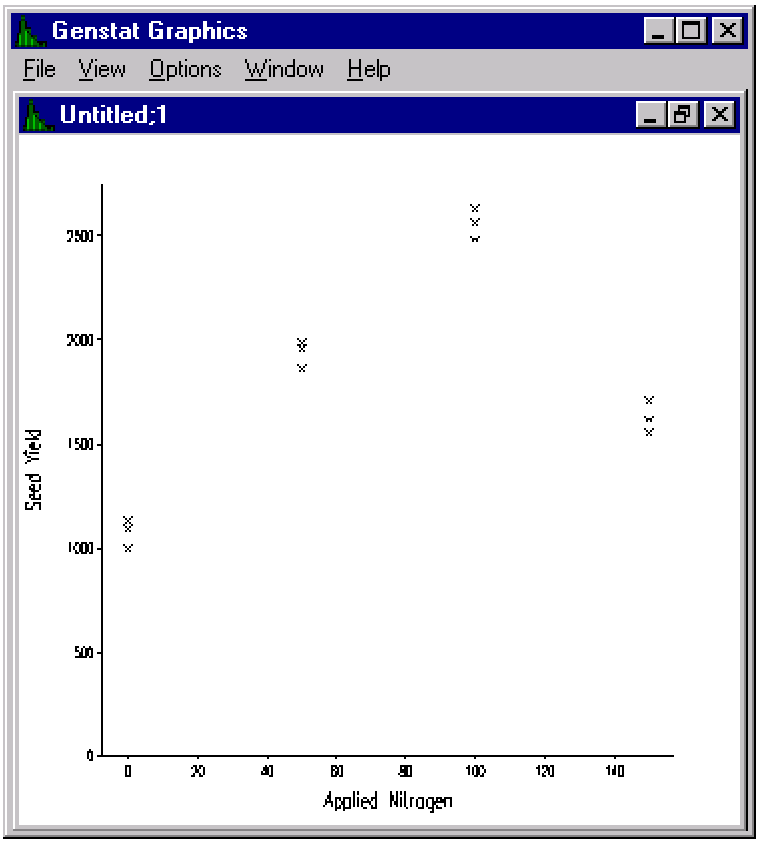
As the number of possible explanatory variables increases the number of possible models quickly becomes too large for it to be practical to fit them all. For example, with six possible explanatory variables there are 64 possible models, and with seven there are 128. For some problems we will have sufficient prior knowledge to be able to select a sub-set of the possible models which it is sensible to consider, in which case we can take a similar approach to that just described. For other problems, usually those for which we are after a predictive model rather than one which provides an understanding of the processes behind the observed relationships, we simply want to find the model which best summarises the observed data. A number of formalised techniques, generally described as ***stepwise regression techniques***, have been developed to tackle this problem.

Forward selection techniques start from the null model and, at each step, add to the current model the variable which has the largest *F*-value, as long as this value is significant. The process stops once there are no variables that are not in the model which have significant *F*-values. Similarly, ***backward elimination*** techniques start from the full model and, at each step, drop from the current model the variable with the smallest *F*-value as long as this value is not significant. Again, the process stops once there are no more variable in the model which have non-significant *F*-values. Combining these two approaches we get a third stepwise regression technique. At each step of the combined technique, the variable not in the model with the largest *F*-value is added if the value is significant, and the variable in the model with the smallest *F*-value is dropped if the value is not significant. The process stops when there are no more variables with significant *F*-values to be added to the model.

In using stepwise regression techniques we should always be aware of the possible dangers associated with using them. Using these techniques is not a substitute for thinking about whether it is sensible for the response variable to be dependant on each of the potential explanatory variables, and the techniques should not be used as a model generating process. However, where we have a large set of related explanatory variables which we believe to have a combined effect on the response variable, as in the example above, stepwise regression techniques provide a useful tool for selecting a subset of the explanatory variables to be used in predicting the response variable.

Polynomial Regression

So far we have only concerned ourselves with linear relationships between the response and explanatory variables, but most biologically relationships are not linear. For example, the figure below shows the curvilinear relationship between yield of sunflower seed and applied nitrogen from an experiment conducted in Nigeria.



Whilst linear regression models are linear in the parameters, they do not have to be linear in the explanatory variables, and we can use the multiple linear regression technique to fit a model such as the quadratic relationship

$$y = a + b\,x + c\,x^2$$

to the yield-nitrogen data simply by calculating an extra explanatory variable equal to the square of our first one.

As with stepwise regression techniques, there are dangers associated with using polynomial regression techniques. By fitting a polynomial of order one less than the number of discrete values of the explanatory variables (i.e. order three in the example below) we will fit the observed mean values exactly. However, the fitted model will not be particularly useful for the prediction of responses for values of the explanatory variable between the observed values, as it will be too curvilinear. Polynomial models will also very rarely provide us with interpretable parameters, unlike the non-linear models we will meet later. Finally, there are a number of constraints imposed by the polynomial form of the model – for example a quadratic is symmetrical about the maximum, and the curvature of a quadratic changes in a fixed way. We should therefore be wary of using the fitted polynomial to predict the response outside the range of the observed explanatory variable (known as ***extrapolating***).

Weighted Regression or Transformed Variables?

One of the assumptions underlying regression techniques, as for the analysis of variance of designed experiments, is that of constant variance for the dependent or response variable. This can be checked graphically by plotting the residuals against fitted values after fitting a model.

If there was evidence that the variance was not constant in the analysis of variance for a designed experiment, we would often consider some form of transformation of the data prior to analysis. Common transformations used include the logarithmic transformation when the variance increases with the mean, the square root transformation for count data, and the angular transformation for proportions based on counts. Similarly transformed data can be analysed within the regression framework, but the analysis of transformed data can introduce additional problems. One possible problem is that whilst the relationship between the response variable and explanatory variable(s) might be a simple linear model for the untransformed response, the relationship is much more complicated when the response variable is on the transformed scale. Alternatively, if the relationship between the transformed variable and explanatory variables is linear, the interpretation of the fitted model can be rather complicated, as we cannot express the original response variable as a simple function of the explanatory variable(s).

For some forms of mean-variance relationship an alternative to analysing a transformed response variable is to use a generalised linear model, in which the model is transformed rather than the response. We will cover this approach in some detail later in the module.

Another possible method of overcoming the problem is to use **weighted regression**. In this approach the contribution of each observation to the fitted regression is weighted so that the influence of more variable observations is reduced and that of less variable observations is increased. The appropriate weighting variate to use depends on how the variability changes in respect of the response variate. For example, the square root and logarithmic transformations correspond to weighting variates of 1/y and $1/y^2$, respectively. Using this approach it is also possible to base the weights on an external estimate of the variance at each value of the explanatory variable.

Non-linear Regression

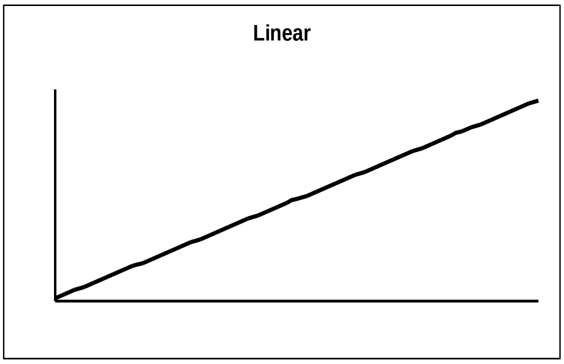
All of the models introduced earlier are in the class of models known as **linear models**, because the parameters in the models occur in a simple linear form. That is, in the three examples repeated below, the response variable, *y*, is a linear combination of the parameters *a*, *b*, *c* and *d*.

(a) Simple linear regression $y = a + b x$

(b) Polynomial regression $y = a + b x + c x^2$

(c) Multiple linear regression $y = a + b x_1 + c x_2 + d x_1 x_2$

It is important in discussing such models to be clear that, for example, the polynomial regression model is a linear model because the parameters occur in a linear fashion, though the relationship between *y* and *x* is definitely not linear, but quadratic.



Linear models have dominated the statistical methods for investigating relationship, not because such models are always the most appropriate, but because the theory of fitting such models to data is very simple. The calculations involved in obtaining estimates of parameters in linear models require only the solution of a set of simple simultaneous equations. These calculations can be done without the aid of computers, though they can be done quicker with computers.

In contrast other, possibly more realistic, forms of model which involve parameters in a non-linear fashion can not be so simply fitted without the use of a computer. Some forms of non-linear model were investigated before the development of modern computers, and complicated methods of fitting them were devised. However, these models inevitably had little appeal to statisticians or research workers because of their complexity, and they were not widely used, with the exception of probit analysis. With the availability of high speed computers the fitting of non-linear models should be no more difficult than that of linear models. The complicated methods of fitting have been replaced by simpler methods, which require a large amount of computation. It is therefore that the research biologist is aware that there should be no difficulties in fitting these models.

To see why non-linear models should be useful, it is necessary to consider why linear models are inadequate to model biological situations. If we are considering a relationship between an explanatory variable, *x*, and a response variable, *y*, then the three simplest forms of linear model are the straight line

$$y = a + b x$$

the quadratic

$$y = a + b x + c x^2$$

and the cubic

$$y = a + b x + c x^2 + d x^3$$

These three models are displayed here.

The straight line is obviously a very restricted relationship. Very few biological relationships are even approximately straight for a reasonable range of values of the explanatory variable, *x*, the most common form of straight line relationship being, perhaps, the allometric relationship between the logarithm of weight of a plant part and the logarithm of the whole plant weight (but note that this is not a straight line but a power relationship $y=Ax^B$ on the natural scale)..

The quadratic model allows for curvature but is restricted in two critical ways. First it is symmetric, the rise in *y* with increasing *x* to a maximum being of exactly the same form as the subsequent decline of *y* with further increase in *x*. The second disadvantage is that the value of *x* must become negative when x is either large or small, and this will usually be biologically unreasonable.

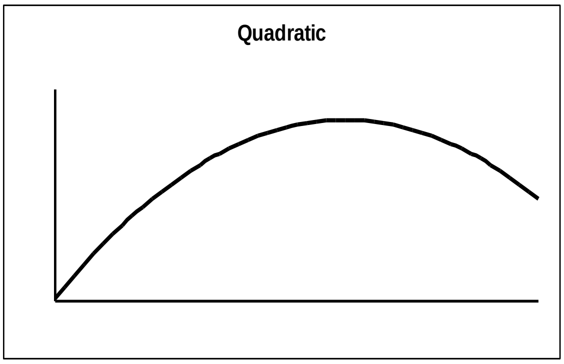
The cubic polynomial, and polynomials of yet higher degree, overcome the disadvantages of symmetry but not those of producing unrealistically negative or very large values of *y* for large or small values of *x*. In addition they have a very rigid structure of maximum and minimum values.

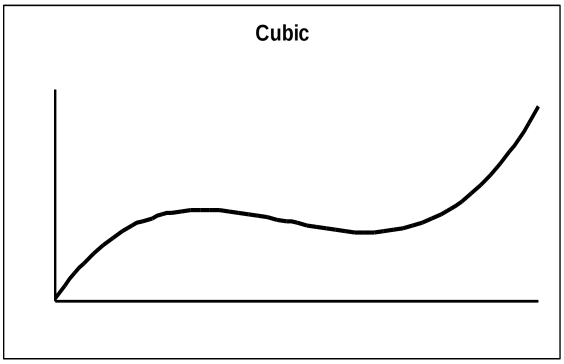
None of the curves in the polynomial family of models allows for a relationship which tends to an asymptotic level of *y* as *x* becomes large, or for relationships where *y* is necessarily positive. In contrast, most of the non-linear models in common use do allow such biologically realistic forms of behaviour. In addition, many of the commonly used non-linear models can be derived from simple biological concepts which, to the extent that the are appropriate, *justify* the use of the non-linear model.

Choosing an appropriate non-linear model

There are two basic approaches to selecting an appropriate non-linear model for a particular set of data. The first is to graph the data and compare the shape of the response with a range of possible models. The one note of caution here is that, for some of the curves that will be presented in this section, considerable flexibility in the shape of response can be achieved by altering the parameter values. The second approach is to consider what models have been used previously for similar types of response, or even to develop some idea of the mechanism behind the observed response.

Both approaches require some knowledge of the range of possible models, and in the next few pages we consider a number of possible "families" of non-linear curves. As well as being possible biologically realistic models for relatively simple relationships, these families of curves can also be considered as the "building blocks" for developing more complex empirical models. The curves described here are those that are provided as standard non-linear curves within GenStat.

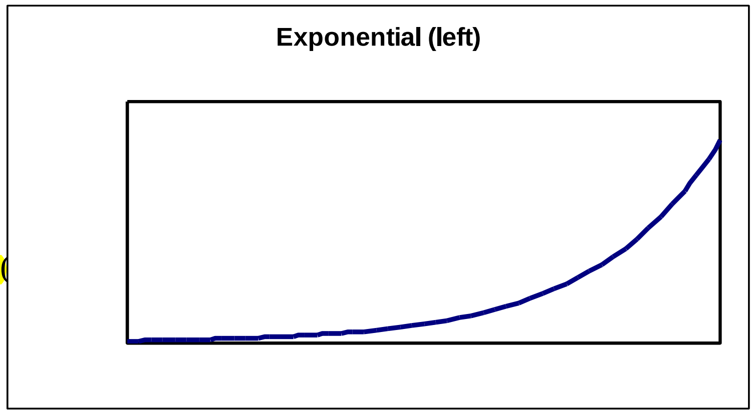




Exponential curves

The simple exponential curve can be useful for a number of different types of relationship. It can be written in two alternative forms

$y = a + br^x$ *or* $y = a + b\exp($



with *r* equal to exp(-*k*).

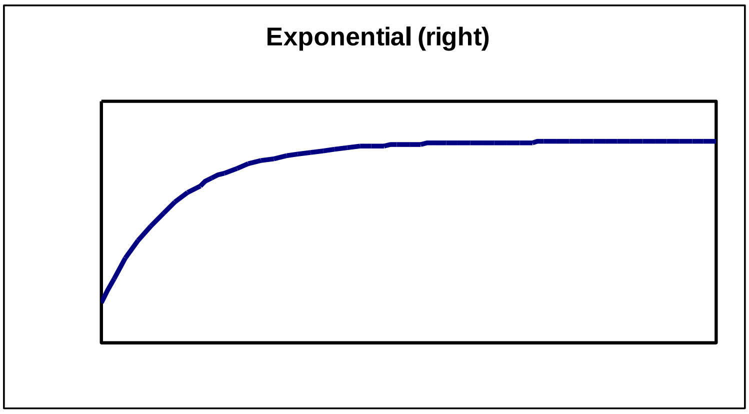
One situation in which this curve is appropriate relates to the decay of some variable with time, the rate of decay being proportional to the value of the variable at any given time. The decay may eventually reach zero, in which case the parameter *a* is zero, or may reach some other, positive, asymptote which will be estimated by *a*. The rate of decay is given by the parameter *k*, and *a* + *b* gives an estimate of the value of the variable at time zero. The half-life (the time for half the activity given by *b* to decay away) is given by *k* * ln(2). Exponential decay curves have been used to model the decay of pesticides in soil, though often something more complex will be required.

Another area in which exponential models have been extensively used is where the response variable either increase with time or as a result of an increasing level of some stimulus variable.

The change in the response is initially fast, but gradually declines until the response reaches an asymptotic level, as shown to the right.

If the initial level of the variable is zero, then parameter *b* is equal to –*a*, and again the parameter *k* estimates the rate of growth, declining as the response approaches the asymptote. In both these first two cases, the parameter *r* (= exp(-*k*)) is less than 1.00.

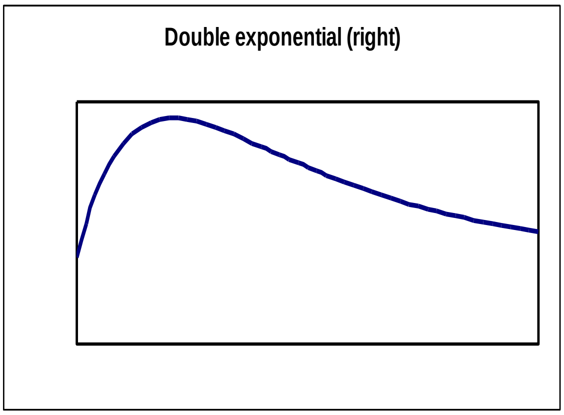
A final area where the simple exponential model is appropriate is for unconstrained growth. The rate of growth, given by parameter *k*, is proportional to the current size of the response, and the parameter *r* will be greater than 1.00. The initial response is given by *a* + *b*, and parameter *b* estimates that part of the initial response contributing to the growth.

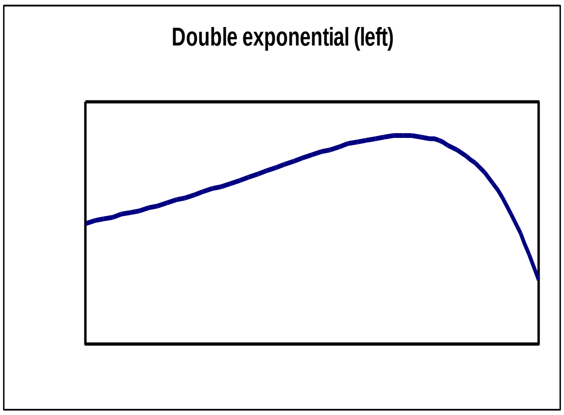


There are a number of extensions to the simple exponential model, the most general represented by the double exponential curve

$$y = a + br^x + cs^x \qquad or \qquad y = a + b\exp(-kt) + c\exp(-lt)$$

which can be extended further by including additional terms of the form *c s*^{x}. This general form relates to the decay of a mixture of two (or more) compounds, with each compound decaying at a

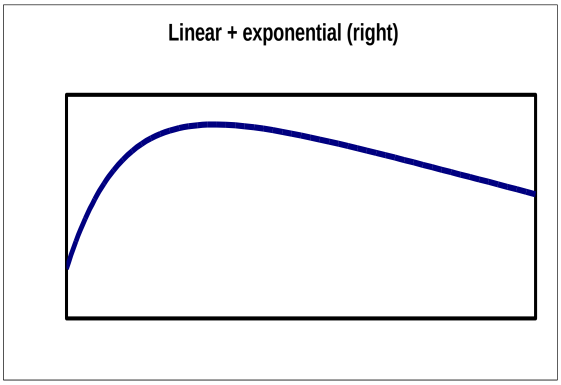


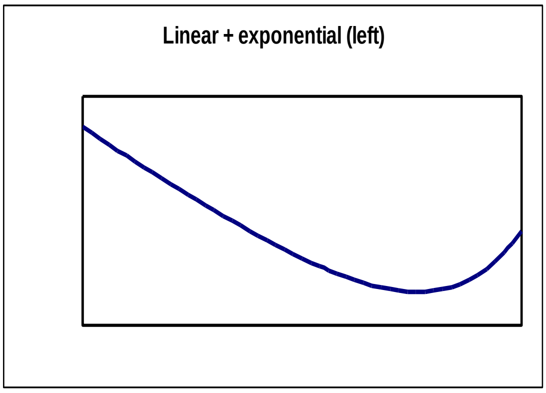


different rate (parameters *r* and *s*) from a different initial level (parameters *b* and *c*). It can similarly be used to model the growth of a mixture of components, with different growth rates and initial amounts of each component. Two possible shapes of response are given below, the first where the parameters *r* and *s* are both less than one, the second where both are greater than one, though more flexibility of shape than this is possible depending on the relative sizes of the different rate parameters.

The double exponential equation can be simplified in a number of ways to give other members of the exponential curve family. A first common example is the "line plus exponential", obtained when one of the decay or growth rates (*r* or *s*) approaches 1.00.

$$y = a + br^{x} + cx \qquad or \qquad y = a + b\exp(-kx) + cx$$





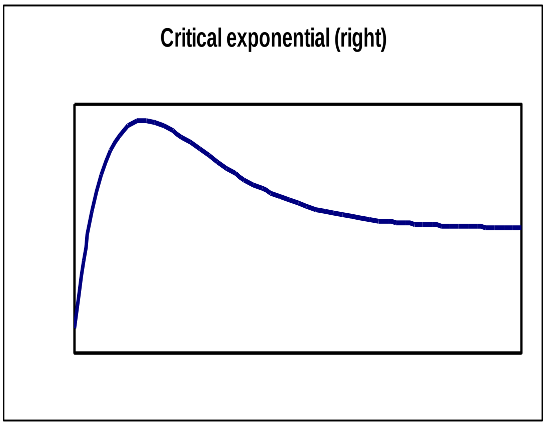
Again, different shapes of response are possible depending on whether parameter *r* is greater or less than 1.00, and depending on the relative sizes of the exponential and linear components. Two examples, the first for *r* less than one, the second for *r* greater than one, are shown below.

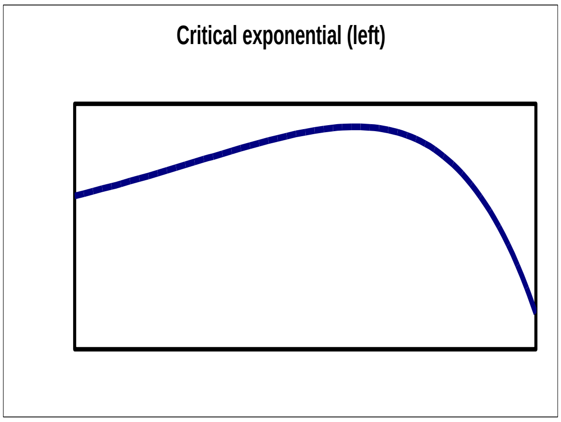
A second common example of a simplification of the double exponential is the "critical exponential,

$$y = a + (b + cx)r^{x} \qquad or \qquad y = a + (b + cx)\exp(-kx)$$

obtained when the two rates are very similar. Again, two forms of the curve are possible depending on whether *r* is greater than or less than 1.00, and considerable flexibility of the shape of the curve is possible depending on the relative sizes of parameters *b* and *c*. Two example curves, the first for *r* less than one, the second for *r* greater than one, are shown below.

For either of these two simplifications, further simplification to the simple exponential curve is possible if parameter *c* is close to zero, and the simple exponential curve simplifies to a straight line if the rate parameter *r* approaches 1.00. Note, however, that a value of exactly 1.00 gives a constant response.





Sigmoid Growth curves

Many biological investigations are concerned with the growth of organisms with time. Extensive studies have been made of the growth of whole plants or individual plant parts, and the growth can usually be described in three phases. Early growth is fairly rapid, and in this initial phase the rate of growth is usually proportional to size (as described by the exponential function introduced above). The second phase tends to be less rapid, being a balance between maintenance and growth, and is often almost linear. In the third phase, the growth rate diminishes with the plant size reaching an asymptotic upper limit, as described by the exponential decline or decay type models introduced earlier.

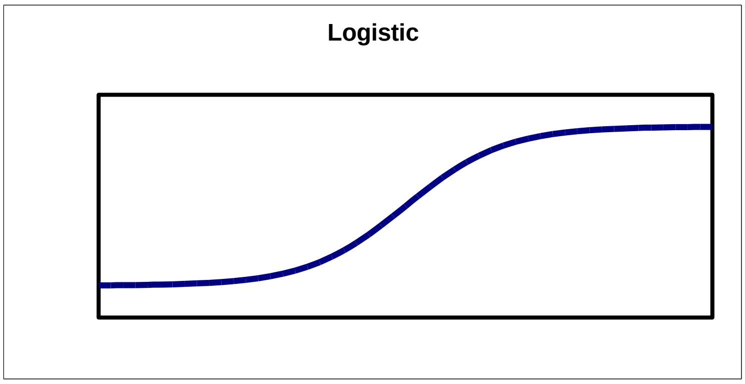
A similar shape of curve can also be used for dose-response studies, for example relating plant dry weight to the log of the applied herbicide dose. When considering a discrete response variable in such studies (such as the number of insects or plants affected), the same shape of response can be considered, but with due account taken of the fact that the data are probably Binomially distributed.

The three growth phases can be combined into a single differential equation relating the growth rate to both the current size of the plant and the difference between the current size and the potential size. In its simplest form the increasing and declining growth phases will have the same rates of change of growth

$$\frac{dy}{dx} = b\,y\,(c - y)$$

Solving this equation leads to the **logistic** function

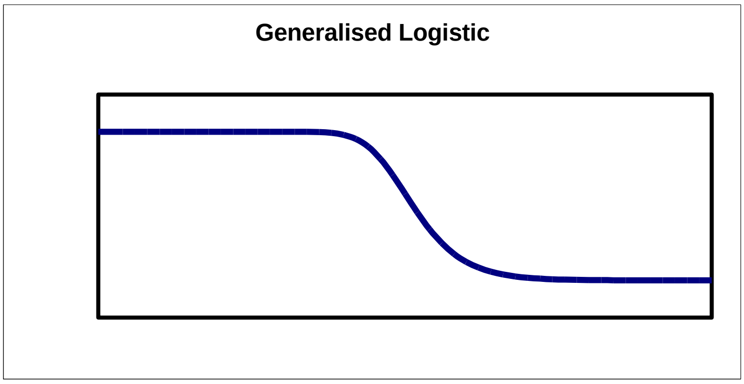
$$y = a + \frac{c}{1 + \exp(-b(x - m))}$$



This is a symmetric curve (as shown), with parameter *a* estimating the initial size and parameter *c* giving the total growth, so that *a* + *c* is the final size. Parameter *m* estimates the time (value of *x*) at which the maximum growth rate is achieved, which is at a size of *a* + (*c*/2). Parameter *b* is related to the maximum growth rate. Obviously, if parameter *b* has the opposite sign, the response will represent a declining rather than growing response

This form of response can be generalised by including an additional parameter, which affects the symmetry of the response. This is known as the **generalised logistic** function,

$$y = a + \frac{c}{\left[1 + t\exp\left(-b\left(x-m\right)\right)\right]^{1/t}}$$



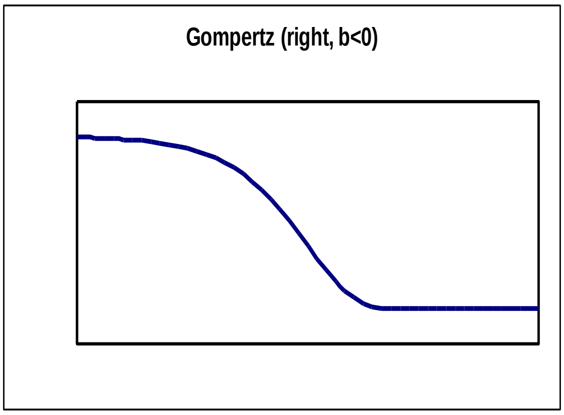
which with parameter *t* set equal to 1.00 simplifies to the logistic.

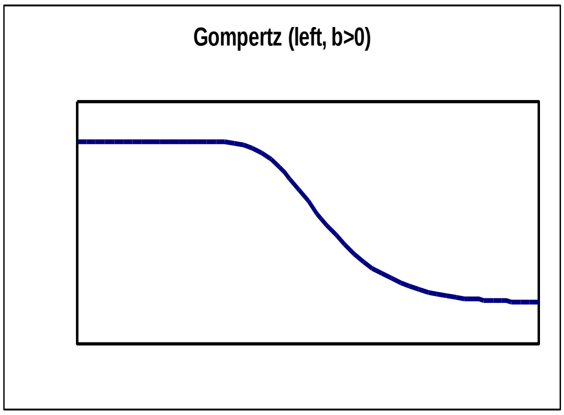
Values of *t* greater than 1.00 lead to the maximum growth/decay rate being closer to the maximum size, those less than 1.00 to it being closer to the minimum size.

Another special case of the generalised logistic is obtained as parameter *t* tends toward zero. This is the **Gompertz** curve.

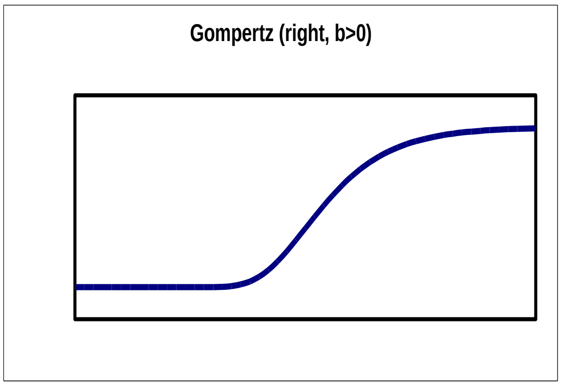
$$y = a + c * \exp\left(-\exp\left(-b\left(x-m\right)\right)\right)$$

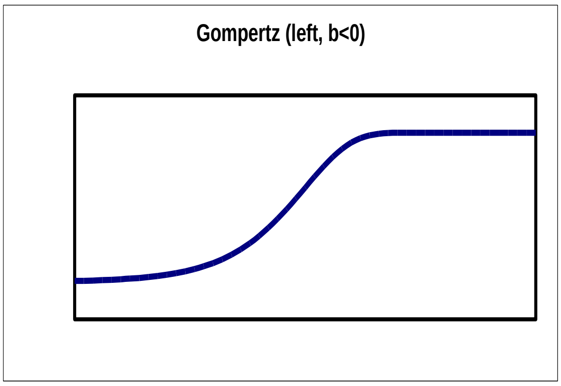
As for the other sigmoid growth curves, the maximum growth rate occurs at time *m*, which occurs at a response that is a proportion (1/e) of the distance between the lower asymptote (*a*) and upper





asymptote (*a* + *c*). Whether this is nearer to the upper or lower asymptote depends on the values of the other parameters, and different combinations can produce a range of different shaped curves, as shown below.





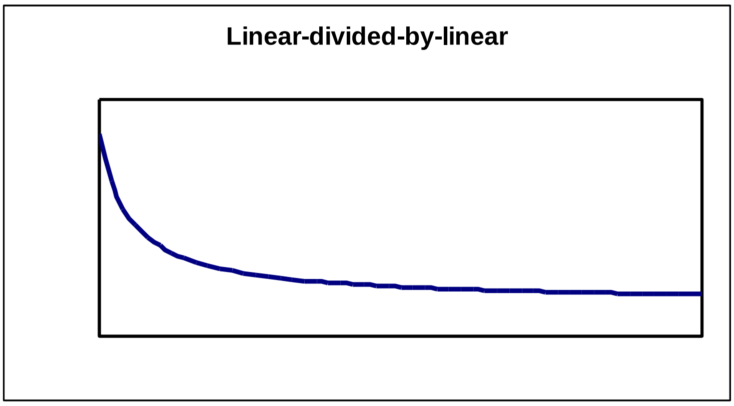
Rational functions

These are ratios of polynomials, though from the formulae this may not be immediately obvious. The linear-divided-by-linear curve is a rectangular hyperbola, which occurs for example as the Michaelis-Menten law of chemical kinetics.

$$y = a + \frac{b}{1 + d\,x}$$

When *x* is zero the response is equal to *a* + *b*, and as *x* increases the response tends towards *a*. An example curve is shown below.

This curve has been used to describe the relationship between, for example, crop yield (*y*) and weed density (*x*). In this case the parameters are interpretable as follows: *a* is the crop yield at large weed densities, *b* is the potential crop loss, and *d* is the weed competitivity (1/*d* is the weed density that will reduce the crop yield by 50% of its potential loss).



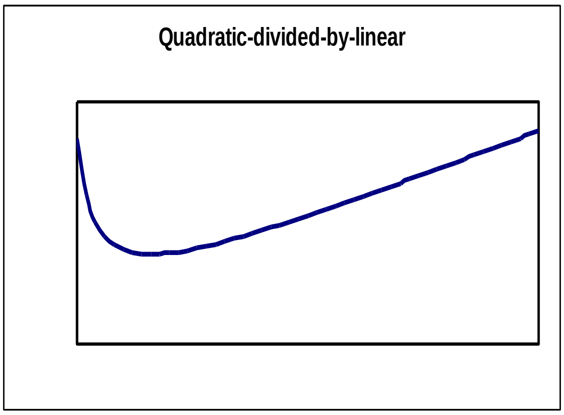
Two extensions of the linear-divided-by-linear curve are the quadratic-divided-by-linear

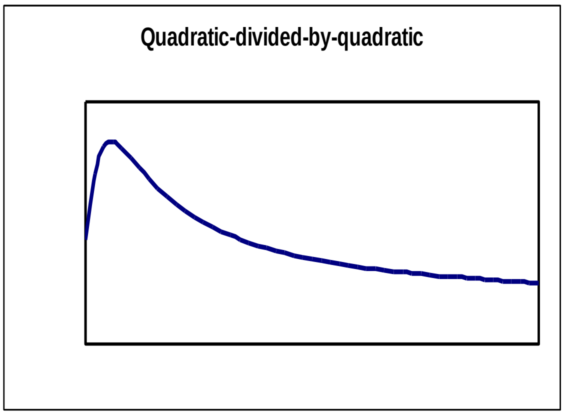
$$y = a + \frac{b}{1 + d\,x} + c\,x$$

And quadratic-divided-by-quadratic

$$y = a + \frac{b + c\,x}{1 + d\,x + e\,x^2}$$

Examples of the shapes of these curves are shown below.





The advantage of these curves is that they are extremely flexible, though they have the disadvantage that the parameters are not always easy to interpret. All of these rational functions have been used to model the relationship between crop yield and applied nutrient levels, and they are related to the inverse polynomial models develop by Nelder for such relationships.

Fourier curves

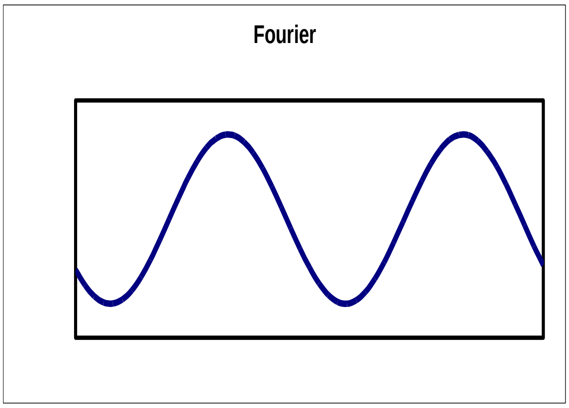
Fourier curves are trigonometric functions, using either the sine or cosine function to model periodic behaviour. Two such curves are a single Fourier curve

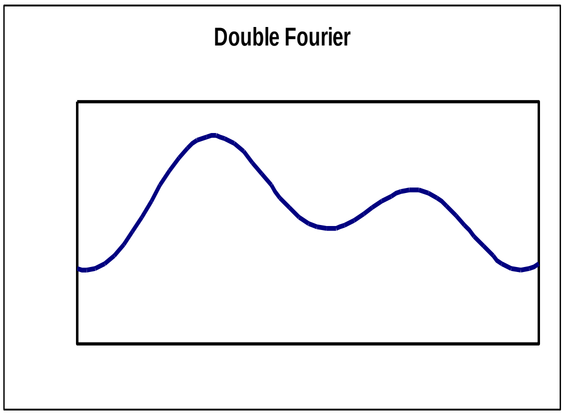
$$y = a + b\sin\left(\frac{2\pi(x-e)}{w}\right)$$

and a double Fourier curve

$$y = a + b\sin\left(\frac{2\pi(x-e)}{w}\right) + c\sin\left(\frac{4\pi(x-f)}{w}\right)$$

Parameter w defines the wavelength or period of the response, with parameters e and f defining how the response is shifted along the x-axis. Parameters b and c define the amplitude of the (components of the) response, with parameter a defining the mean response (the response at $x = 0$ if parameters e and f are zero). Example curves are shown below.





Gaussian curves

The Gaussian curve is bell-shaped curve like the Normal probability density function. Two such curves are a single Gaussian curve

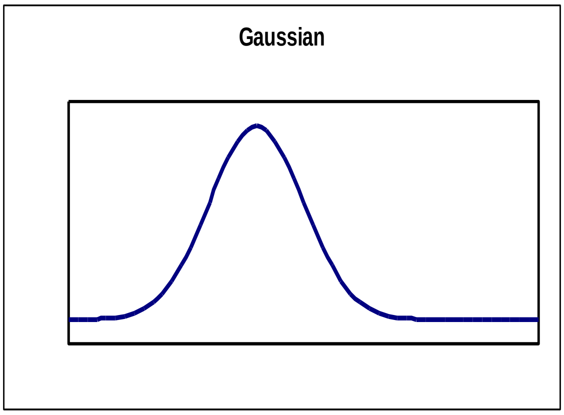
$$y = a + \frac{b}{\sqrt{2\pi}} \exp\left(\frac{-(x-m)^2}{2s^2} \right)$$

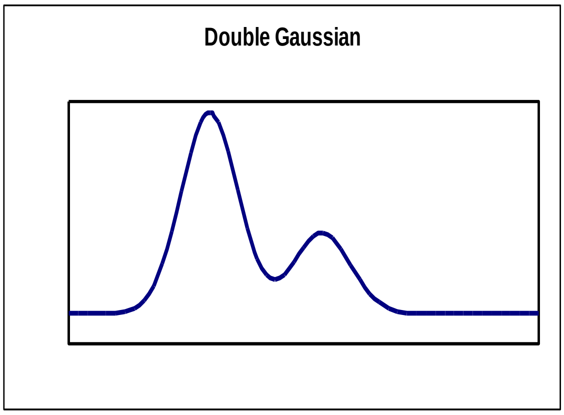
and a double Gaussian curve

$$y = a + \frac{b}{\sqrt{2\pi}} \exp\left(\frac{-(x-m_1)^2}{2s^2} \right) + \frac{c}{\sqrt{2\pi}} \exp\left(\frac{-(x-m_2)^2}{2s^2} \right)$$

The double Gaussian curve is simply a sum of two overlapping Gaussian curves, and occurs, for example, in spectography. Parameter *a* is usually referred to as the background and parameters *b* and *c* defining the strength of each component in the mixture. Parameters m_1 and m_2 define the positions of the peaks, with parameter *s* defining the spread of both components. Note that this form constrains the spread to be the same for both components

Example curves are shown below.





Parallel non-linear curve analysis

Following the approach developed in the linear regression framework for determining whether or not separate regression lines were needed for the responses at different levels of some classifying factor, we can develop the idea of parallel curve analysis. Of course, since we have more parameters involved, and in a more complicated manner, such an analysis for a non-linear curve is more complicated than that for any simple linear regression model.

Other modelling approaches

"Broken stick" regression (also referred to as "Split-line regression")

Sometimes we will have a response that consists of two distinct phases, both of which can be modelled using a straight line relationship. Our interest here might be to estimate the slopes of the responses in each phase, together with the value of the explanatory variable at which the transition occurs. A simplistic approach to fitting such a model is to divide the dataset into two groups, fit linear regressions to each group separately and combine the residual sum of squares and degrees of freedom. Repeating this division for each possible division of the data into two contiguous groups, we can then identify the division which leads to the minimum residual sum of squares, and the parameters from the fitted lines for this analysis will give us the required information. Alternatively we can treat the "break-point" as a non-linear parameter, and fit the combined model using a general non-linear regression approach.

Cubic smoothing splines

Rather than initially choose a particular parametric model (straight line, quadratic, exponential, logistic, etc.) to describe our response, a recently developed and generally exploratory approach is to fit what might be thought of as a non-parametric curve to the data. This is usually referred to as a smoothing spline and consists of a number of separate segments of cubic polynomials fitted between the distinct values of the explanatory variate, and constrained to be "smooth" at the joins. These smoothing splines have a rather complicated parameterization and so are of relatively little value for interpretation or predictions. Finding the best fitting model is a question of balancing how well the fitted response agrees with the data and the smoothness of the fitted response. At one extreme is a simple straight line relationship, and at the other is a curve which goes through every data point.

Locally weighted regression

This is another approach which avoids selecting a particular parametric model, fitting linear or quadratic polynomials locally around each data point, with the regression weighted so that observations further away from the point of interest make less of a contribution to the fit. You may see this approach referred to as LOESS regression.

General non-linear regression and function minimization

Whilst the collection of standard non-linear response curves described above will include a suitable model form for many biological responses, you may find that your particular observed response has a shape that cannot be described by any of them. However, if you can write down an algebraic equation to describe the shape of response, preferably based on knowledge of the biological mechanisms driving the observed response, then, probably working with a friendly consultant statistician, it should be possible to estimate the parameters of the response function using a maximum likelihood modelling approach. With such non-linear models it will often be necessary to try different parameterisations of the same model to find one which is easy to fit, and, as the fitting methodology uses an iterative approach, it will usually be essential to be able to provide good initial estimates for each of the parameters. Minimising the number of parameters fitted in the model is also sensible.

General non-linear regression also provides a way of generalising models where, for example, a standard curve has been fitted to the response at each level of a second explanatory variable, and interest is in modelling the parameters of the standard non-linear curve as functions of the second explanatory variable. Here we should be able to use the standard non-linear model parameterisation, and then include extra models, possibly linear, to describe how the parameters change with the second explanatory variable.