# Regression & Generalized Linear (Mixed) Models

Mick O’Neill

STatistical Advisory & Training Service Pty Ltd

Last updated August 2010

# Introduction

In recent years a general algorithm, Restricted Maximum Likelihood (REML), has been developed for estimating variance parameters in linear mixed models (LMM). This topic is covered in our manual ANOVA & REML – a guide to linear mixed models in an experimental design context (see www.stats.net.au and Resources).

This manual covers classic statistical techniques of linear and non-linear regression for normally distributed data, and introduces the General Linear Model (GLM) for data that are not normally distributed. When the analysis of non-normal data includes random terms, a General Linear Mixed Model is discussed. It therefore helps to have the basic concepts of REML and deviance for these topics. The statistical package GenStat is used throughout. The current version is 13, although the analyses can generally be performed using the Discovery Edition released in 2010.

In general, data from two familiar text books will be used as examples. The editions we used are the following.

Snedecor, G.W. and Cochran, W.G. (1980). Statistical Methods. Seventh Edition. Ames Iowa: The Iowa State University Press.

Steel, R.G.D. and Torrie, J.H. (1980). Principles and Procedures of Statistics: a Biometrical Approach. Second Edition. New York: McGraw-Hill Kogakusha.

Other sources for data include an example from GenStat's *Statistics* Guide available in its **Help** menu, and an example from each of

Diggle, P.J. (1983). Statistical Analysis of Spatial Point Patterns. London: Academic Press.

Mead, R. and Curnow, R.N. (1990). Statistical methods in agricultural and experimental biology. Chapman and Hall, London.

Ratkowsky, D.A. (1990). Handbook of nonlinear regression models. 102-791-088 (*Last edited on* 2002/02/27 18:18:23 US/Mountain)

The training manual was prepared by Mick O'Neill from the **Statistical Advisory & Training Service Pty Ltd**. Contact details are as follows.

Mick O'Neill mick@stats.net.au
Statistical Advisory & Training Service Pty Ltd
www.stats.net.au

# Table of Contents

# Section 1 - Correlation

Firstly, suppose we have $n$ pairs of observations, $(X_1, Y_1)$, $(X_2, Y_2)$, …, $(X_n, Y_n)$. Both could be *random* variates, or one (say the $X$ variate) could be controlled as part of the experiment (e.g. different set temperature chambers, sowing densities) and is hence a *fixed* variate.

Both correlation and simple linear regression coefficients measure the degree of the *linear relationship* between two variables. To summarise the difference:

*Regression* is used when one is interested in explaining a relationship between the dependent variate Y and the fixed variate X. It may also be used to predict future observations. If X is measured with error, the regression is interpreted as conditional on the X-values observed.

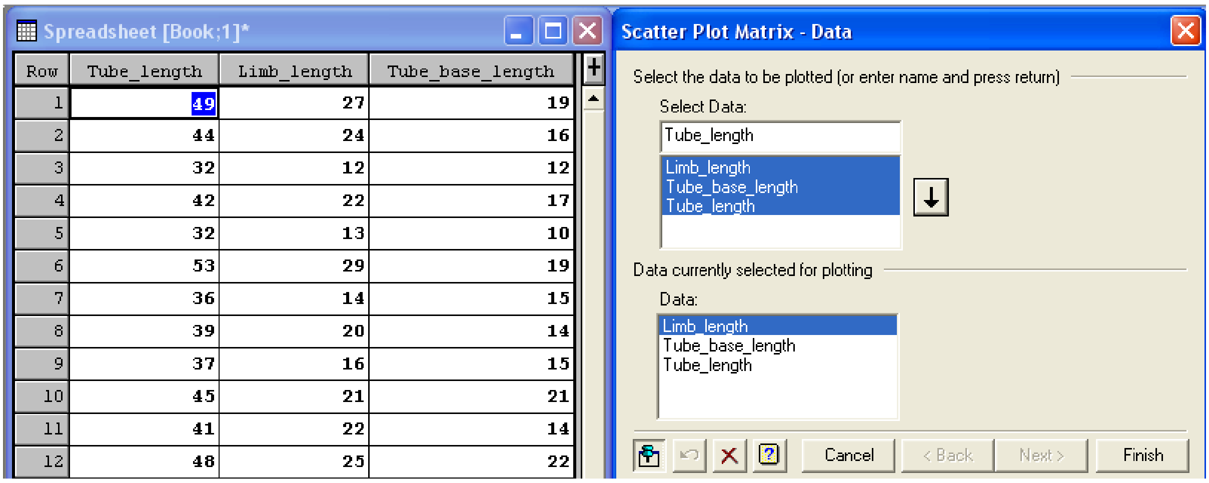
*Correlation* is used when one is simply interested in measuring the co-relation between two variates that appear to vary linearly with each other. Neither X nor Y is more important, they are both variates of interest.

## Correlation

Example 1 Data on flowers of a Nicotiana cross (Steel and Torrie, page 276)

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tube length | 49 | 44 | 32 | 42 | 32 | 53 | 36 | 39 | 37 | 45 | 41 | 48 | 45 | 39 | 40 | 34 | 37 | 35 |
| Limb length | 27 | 24 | 12 | 22 | 13 | 29 | 14 | 20 | 16 | 21 | 22 | 25 | 23 | 18 | 20 | 15 | 20 | 13 |
| Tube base length | 19 | 16 | 12 | 17 | 10 | 19 | 15 | 14 | 15 | 21 | 14 | 22 | 22 | 15 | 14 | 15 | 15 | 16 |

This is clearly when correlation is of interest. GenStat allows all three variates to be plotted against each other. Select **Graphics > Scatter Plot Matrix** and select all three variates into the Data box.



| Row | Tube_length | Limb_length | Tube_base_length |
|---|---|---|---|
| 1 | 49 | 27 | 19 |
| 2 | 44 | 24 | 16 |
| 3 | 32 | 12 | 12 |
| 4 | 42 | 22 | 17 |
| 5 | 32 | 13 | 10 |
| 6 | 53 | 29 | 19 |
| 7 | 36 | 14 | 15 |
| 8 | 39 | 20 | 14 |
| 9 | 37 | 16 | 15 |
| 10 | 45 | 21 | 21 |
| 11 | 41 | 22 | 14 |
| 12 | 48 | 25 | 22 |

Select the data to be plotted (or enter name and press return)

Select Data: Tube_length

Limb_length
Tube_base_length
Tube_length

Data currently selected for plotting

Data:
Limb_length
Tube_base_length
Tube_length

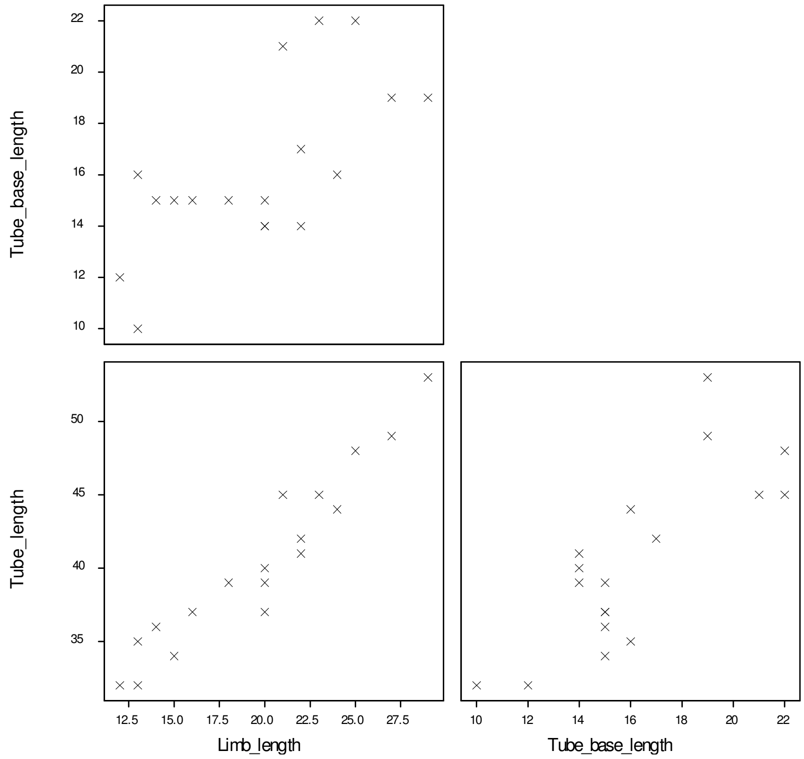
The plot on the following page shows a strong relationship between tube and limb lengths, a relatively strong relationship between tube and tube base lengths, and a slightly weaker linear relationship between tube base and limb lengths. We quantify this strength by the correlation coefficient defined as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})/(n-1)}{s_x s_y} = \frac{\text{covariance}(X,Y)}{\text{sd}(X)sd(Y)} \quad (1)$$

Here we are using $\bar{x}$ and $\bar{y}$ as sample means, *s* as the sample standard deviation, and we introduced the concept of *covariance*. The sample standard deviation (sd) for *Y* is defined as

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{y})^2}{n-1}} \quad (2)$$

The term inside the square root is the sample *variance*. The *covariance* simply replaces the squared term by an equivalent expression in the second variate thereby measuring the co-variance between *Y* and *X*. Covariances are unbounded.

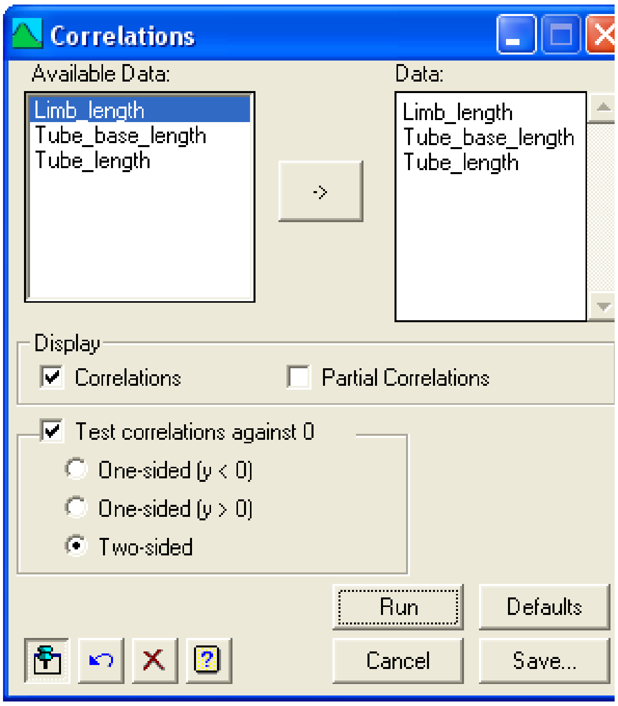


For the data in (1) GenStat returns the correlation coefficients below (use **Stats > Summary Statistics > Correlations** and select the 3 variates):

```
Correlation matrix

         Limb_length          1.000
    Tube_base_length          0.678             1.000
         Tube_length          0.955             0.797               1.000
                        Limb_length  Tube_base_length         Tube_length
```

Correlation coefficients are constrained to lie between -1 and +1. If one variable tends to increase as the other decreases, the correlation coefficient is negative. Conversely, if the two variables tend to increase together the correlation coefficient is positive. A correlation of 1.0

indicates a perfect linear trend with a positive slope. A correlation of 0 indicates no linear trend. If the variates are also *normally distributed*, then a correlation of 0 also indicates that *X* and *Y* are independent.



Correlations

Available Data:
Limb_length
Tube_base_length
Tube_length

Data:
Limb_length
Tube_base_length
Tube_length

Display
☑ Correlations ☐ Partial Correlations

☑ Test correlations against 0
○ One-sided (y < 0)
○ One-sided (y > 0)
● Two-sided

The symbol ρ (rho, Greek r) is usually used for a population correlation coefficient and *r* for a sample coefficient. A special test is available to determine whether variates are uncorrelated, that is, whether ρ = 0. The P-values from GenStat are as follows.

```
Two-sided test of correlations different from zero

Probabilities

Tube_base_length       0.001981
     Tube_length         < 0.001            < 0.001
                     Limb_length   Tube_base_length
```

Clearly, all three variates are strongly linearly related to each other.

Correlated data are extremely common in field experimentation. Sometimes the same plant or plot is measured at various times, and generally observations taken over a short time interval are more strongly correlated than those taken over a long time interval. Similarly, plants grown in a field tend to be more strongly correlated than those grown at distance. Spatial and temporal correlation models have been developed to cater for these common phenomena.

## Calculation in Excel

Suppose the tube length data are named **Tube_length** in Excel and the limb length data **Limb_length**.

**= CORREL(Tube_length,Limb_length)** returns the value 0.9550 (to 4 decimals).

If the **Data Analysis Toolpak** has been added into Excel, the correlation macro produces:

| | *Tube length* | *Limb length* | *Tube base length* |
|---|---|---|---|
| Tube length | 1 | | |
| Limb length | 0.95497792 | 1 | |
| Tube base length | 0.79721422 | 0.678111257 | 1 |

*Warning on calculating covariances in Excel:*
Excel has a sample variance formula **=VAR** and a sample standard deviation formula **=STDEV**. It has a "population" variance formula **=VARP** and a "population"standard deviation formula **=STDEVP**. However, the formula **=COVAR(x,y)** does ***not*** give us what we want. Instead, Excel uses *n* as a divisor instead of *n*-1!

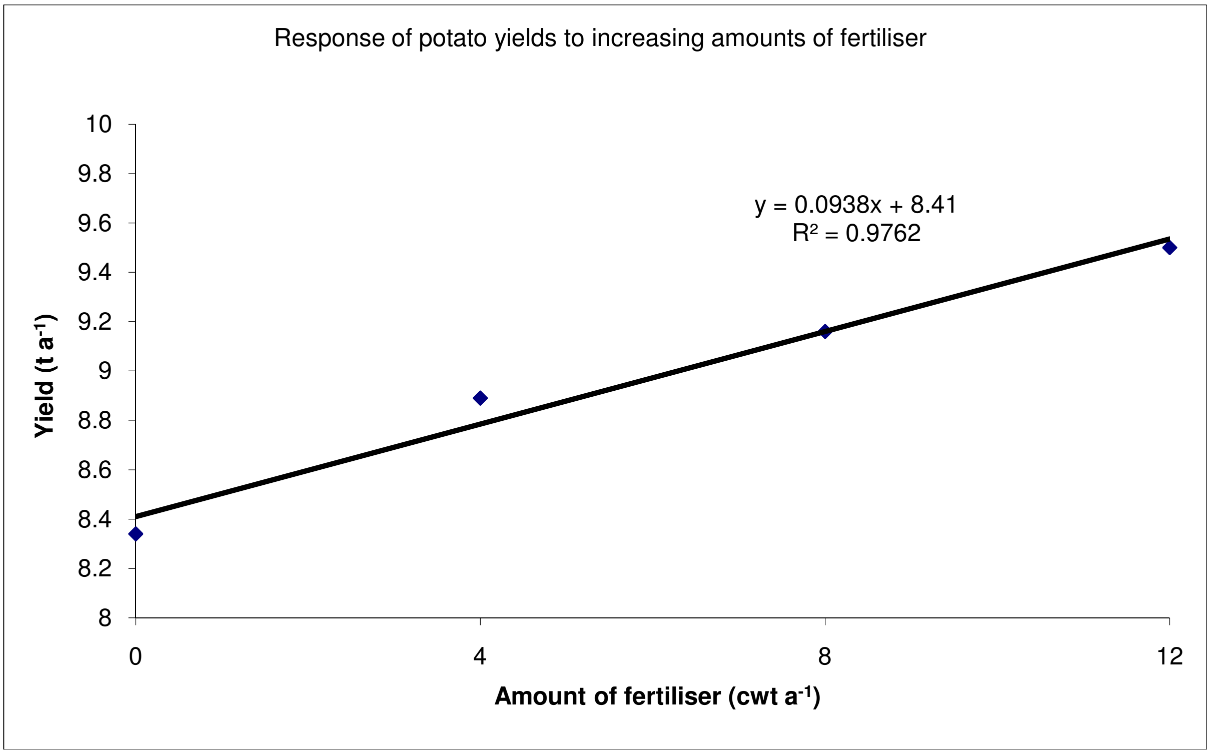
# Section 2 - Regression for normally distributed data

## Simple linear regression

Example 2. .Yields of potatoes receiving amounts of fertilizer (Snedecor and Cochran, page 150).

| Amount | 0 | 4 | 8 | 12 |
|---|---|---|---|---|
| Yield | 8.34 | 8.89 | 9.16 | 9.50 |

This is not a large data set, but a scatter plot in Excel showing a linear trendline indicates a very strong predictive model for yield *over the range of fertiliser levels considered*. We could not use any model generated from these data to predict the yield for more than 12 units of fertiliser.



The line of best fit,

$$Yield = 8.41 + 0.0938\ Fertiliser$$

comes from a procedure known as *least squares*. Drop a perpendicular from each observation to a straight line passing through the points: these are the so-called errors, or residuals. Find the *Residual Sum of Squares*, which is simply the sum of the distances of the errors. Use a mathematical or numerical procedure to minimise the *Residual Sum of Squares*, thereby obtaining a line that goes through the points "as best as possible".

The general form of a simple linear regression line (in one predictor $X_1$) is

$$Y = b_0 + b_1 X_1$$

Here, $b_0$ is the intercept and $b_1$ the slope, that is, the change in $Y$ for a unit increase in $X$. For the potato data, a crop with no fertiliser is predicted to produce 8.41 cwt $a^{-1}$, and for each additional unit of fertiliser added, an increase in yield of 0.09 cwt $a^{-1}$ is predicted.

Furthermore, for simple linear regression, the percentage variation in yield explained by the model is 97.6%. This, in fact, is the square of the correlation coefficient, which turns out to be 0.988. (You can verify that $0.988^2 = 0.976$.)

This model is a special case of a more general linear additive model involving several predictors which we examine now in more detail.

**Multiple linear regression**

The more general multiple linear regression model applies to data taken on a dependent variable $Y$ and a set of $k$ predictor or explanatory variables $X_1, X_2, \ldots, X_k$. We assume we have $n$ sets of data.

With multiple linear regression we explain the variation in the $Y$ values by the following (usually over-simplified) relationship between $Y$ and the set of $X$s

$$Y = (\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k) + error$$

Notice that the linearity refers to the set of parameters $\beta_0$, $\beta_1, \ldots, \beta_k$. Polynomial equations are special cases, with $X_1$, $X_2 = X_1^2$, $X_3 = X_1^3$ and so on. Polynomials in $X$ are still linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$.

The least squares procedure is again used to produce a "line of best fit".

In Example 3 we are interested in predicting burn times using a 3-predictor regression of log(leaf burn) on nitrogen ($N$), chlorine ($Cl$) and potassium ($K$) percentages in tobacco taken from farmers' fields.

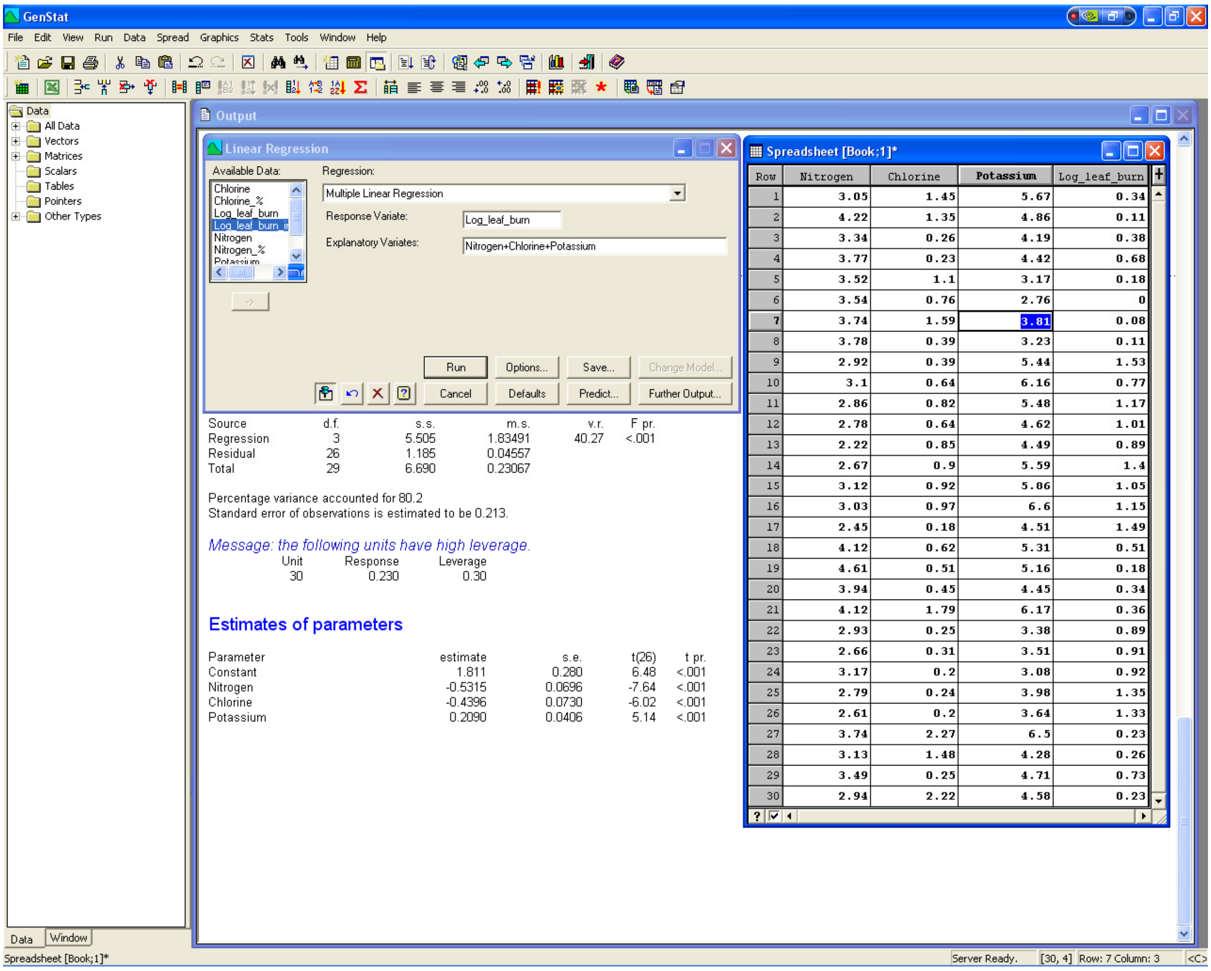
A scatter matrix of the data (see over) shows weak correlations among the dependent variates ($N$, $Cl$ and $K$), as well as negative trends in log(leaf burn) on nitrogen and on chlorine, and a weak positive trend with potassium. These correlations are:

### Correlation matrix

```
Nitrogen         1.000
Chlorine         0.209      1.000
Potassium        0.093      0.407          1.000
Log_leaf_burn   -0.718     -0.500          0.179          1.000
              Nitrogen   Chlorine      Potassium  Log_leaf_burn
```

Example 3 (Steel and Torrie, page 319)

| *N* | *Cl* | *K* | Log(leaf burn) |
|---|---|---|---|
| 3.05 | 1.45 | 5.67 | 0.34 |
| 4.22 | 1.35 | 4.86 | 0.11 |
| 3.34 | 0.26 | 4.19 | 0.38 |
| 3.77 | 0.23 | 4.42 | 0.68 |
| 3.52 | 1.10 | 3.17 | 0.18 |
| 3.54 | 0.76 | 2.76 | 0.00 |
| 3.74 | 1.59 | 3.81 | 0.08 |
| 3.78 | 0.39 | 3.23 | 0.11 |
| 2.92 | 0.39 | 5.44 | 1.53 |
| 3.10 | 0.64 | 6.16 | 0.77 |
| 2.86 | 0.82 | 5.48 | 1.17 |
| 2.78 | 0.64 | 4.62 | 1.01 |
| 2.22 | 0.85 | 4.49 | 0.89 |
| 2.67 | 0.90 | 5.59 | 1.40 |
| 3.12 | 0.92 | 5.86 | 1.05 |
| 3.03 | 0.97 | 6.60 | 1.15 |
| 2.45 | 0.18 | 4.51 | 1.49 |
| 4.12 | 0.62 | 5.31 | 0.51 |
| 4.61 | 0.51 | 5.16 | 0.18 |
| 3.94 | 0.45 | 4.45 | 0.34 |
| 4.12 | 1.79 | 6.17 | 0.36 |
| 2.93 | 0.25 | 3.38 | 0.89 |
| 2.66 | 0.31 | 3.51 | 0.91 |
| 3.17 | 0.20 | 3.08 | 0.92 |
| 2.79 | 0.24 | 3.98 | 1.35 |
| 2.61 | 0.20 | 3.64 | 1.33 |
| 3.74 | 2.27 | 6.50 | 0.23 |
| 3.13 | 1.48 | 4.28 | 0.26 |
| 3.49 | 0.25 | 4.71 | 0.73 |
| 2.94 | 2.22 | 4.58 | 0.23 |



Regression: Multiple Linear Regression

Response Variate: Log_leaf_burn

Explanatory Variates: Nitrogen+Chlorine+Potassium

```
Source          d.f.      s.s.       m.s.     v.r.  F pr.
Regression        3      5.505    1.83491    40.27  <.001
Residual         26      1.185    0.04557
Total            29      6.690    0.23067

Percentage variance accounted for 80.2
Standard error of observations is estimated to be 0.213.

Message: the following units have high leverage.
        Unit    Response    Leverage
          30       0.230        0.30

Estimates of parameters

Parameter              estimate      s.e.    t(26)   t pr.
Constant                  1.811     0.280     6.48   <.001
Nitrogen                -0.5315    0.0696    -7.64   <.001
Chlorine                -0.4396    0.0730    -6.02   <.001
Potassium                0.2090    0.0406     5.14   <.001
```

Spreadsheet [Book;1]*

| Row | Nitrogen | Chlorine | Potassium | Log_leaf_burn |
|---|---|---|---|---|
| 1 | 3.05 | 1.45 | 5.67 | 0.34 |
| 2 | 4.22 | 1.35 | 4.86 | 0.11 |
| 3 | 3.34 | 0.26 | 4.19 | 0.38 |
| 4 | 3.77 | 0.23 | 4.42 | 0.68 |
| 5 | 3.52 | 1.1 | 3.17 | 0.18 |
| 6 | 3.54 | 0.76 | 2.76 | 0 |
| 7 | 3.74 | 1.59 | 3.81 | 0.08 |
| 8 | 3.78 | 0.39 | 3.23 | 0.11 |
| 9 | 2.92 | 0.39 | 5.44 | 1.53 |
| 10 | 3.1 | 0.64 | 6.16 | 0.77 |
| 11 | 2.86 | 0.82 | 5.48 | 1.17 |
| 12 | 2.78 | 0.64 | 4.62 | 1.01 |
| 13 | 2.22 | 0.85 | 4.49 | 0.89 |
| 14 | 2.67 | 0.9 | 5.59 | 1.4 |
| 15 | 3.12 | 0.92 | 5.86 | 1.05 |
| 16 | 3.03 | 0.97 | 6.6 | 1.15 |
| 17 | 2.45 | 0.18 | 4.51 | 1.49 |
| 18 | 4.12 | 0.62 | 5.31 | 0.51 |
| 19 | 4.61 | 0.51 | 5.16 | 0.18 |
| 20 | 3.94 | 0.45 | 4.45 | 0.34 |
| 21 | 4.12 | 1.79 | 6.17 | 0.36 |
| 22 | 2.93 | 0.25 | 3.38 | 0.89 |
| 23 | 2.66 | 0.31 | 3.51 | 0.91 |
| 24 | 3.17 | 0.2 | 3.08 | 0.92 |
| 25 | 2.79 | 0.24 | 3.98 | 1.35 |
| 26 | 2.61 | 0.2 | 3.64 | 1.33 |
| 27 | 3.74 | 2.27 | 6.5 | 0.23 |
| 28 | 3.13 | 1.48 | 4.28 | 0.26 |
| 29 | 3.49 | 0.25 | 4.71 | 0.73 |
| 30 | 2.94 | 2.22 | 4.58 | 0.23 |

## Regression analysis

```
        Response variate:  Log_leaf_burn
            Fitted terms:  Constant, Nitrogen, Chlorine, Potassium
```

## Summary of analysis

```
Source          d.f.      s.s.        m.s.     v.r.   F pr.
Regression         3     5.505     1.83491    40.27   <.001
Residual          26     1.185     0.04557
Total             29     6.690     0.23067

Percentage variance accounted for 80.2
Standard error of observations is estimated to be 0.213.
```

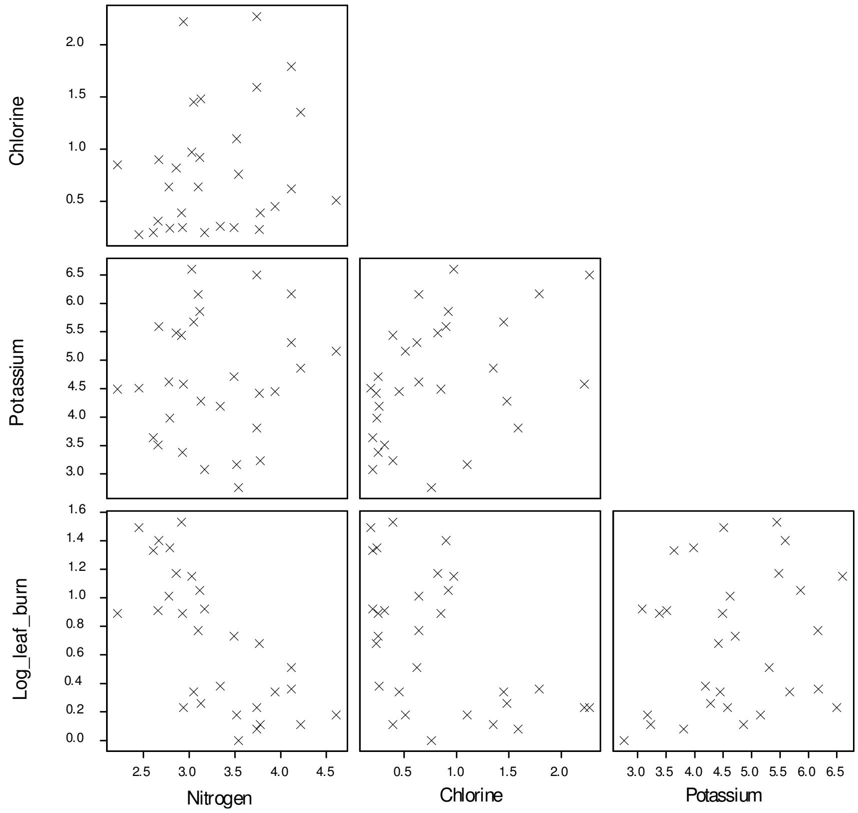
*Message: the following units have high leverage.*

```
          Unit      Response      Leverage
            30         0.230          0.30
```

## Estimates of parameters

```
Parameter             estimate          s.e.      t(26)    t pr.
Constant                 1.811         0.280       6.48    <.001
Nitrogen               -0.5315        0.0696      -7.64    <.001
Chlorine               -0.4396        0.0730      -6.02    <.001
Potassium               0.2090        0.0406       5.14    <.001
```

The largest $R^2$ is associated with nitrogen: if one were interested in a single predictor equation only, then nitrogen would be the best predictor. However, only a fraction over 50% ($-0.718^2 = 0.516$) of the variation in log(leaf burn) data is explained by this simple relationship.



Scatter matrix of nitrogen, chlorine, potassium and log(leaf burn) data of Example 3.

To perform multiple regression in GenStat, choose **Stats > Regression Analysis > Linear Models**.

- **Simple Linear Regression** refers to models with one predictor (including polynomials in one predictor).
- **Multiple Linear Regression** (the other choice, **with Groups**, allows regression equations to be compared across the levels of some factor) refers to models with several predictors.
- **General Linear Regression** allows factors to be included in the model.

The line of best fit is reconstructed from the Estimates of Parameters table in the output:

log(leaf burn) = 1.811 - 0.5315 $N$ - 0.4396 $Cl$+ 0.2090 $K$

## Interpreting regression coefficients

A particular regression coefficient indicates the amount that $Y$ will increase (or decrease) by for a unit rise in that predictor variable, *keeping the other predictor variables fixed*.

For example, for two types of tobacco with the *same* percentage of chlorine and potassium, one with 1% additional nitrogen will burn for -0.5315 fewer log-seconds compared to the other, that is, for only about 30% (= $10^{-0.5315}$) of the time if the transformation used was base10, or about 60% (= $e^{-0.5315}$) of the time if the transformation used was the natural base.

Sometimes it is sensible to interpret the intercept, but it does not always make biological sense to do so. In this case, a value of 1.811 would indicate the log-time that tobacco would burn in the absence of any nitrogen, chlorine and potassium. However, while chlorine is as low as 0.18%, nitrogen and potassium both exceed 2% for all tobacco samples. Interpreting the intercept in this case is like predicting too far away from the experimental data, which is not valid.

In some cases, it might be better to re-write line of best fit by noting the actual solution for the intercept. For line of best fit

$$Y = b_0 + b_1 X_1 + \ldots + b_k X_k$$

using the LS solution

$$b_0 = \bar{y} - b_1\bar{x}_1 - \ldots - b_k\bar{x}_k$$

allows us to write the line as

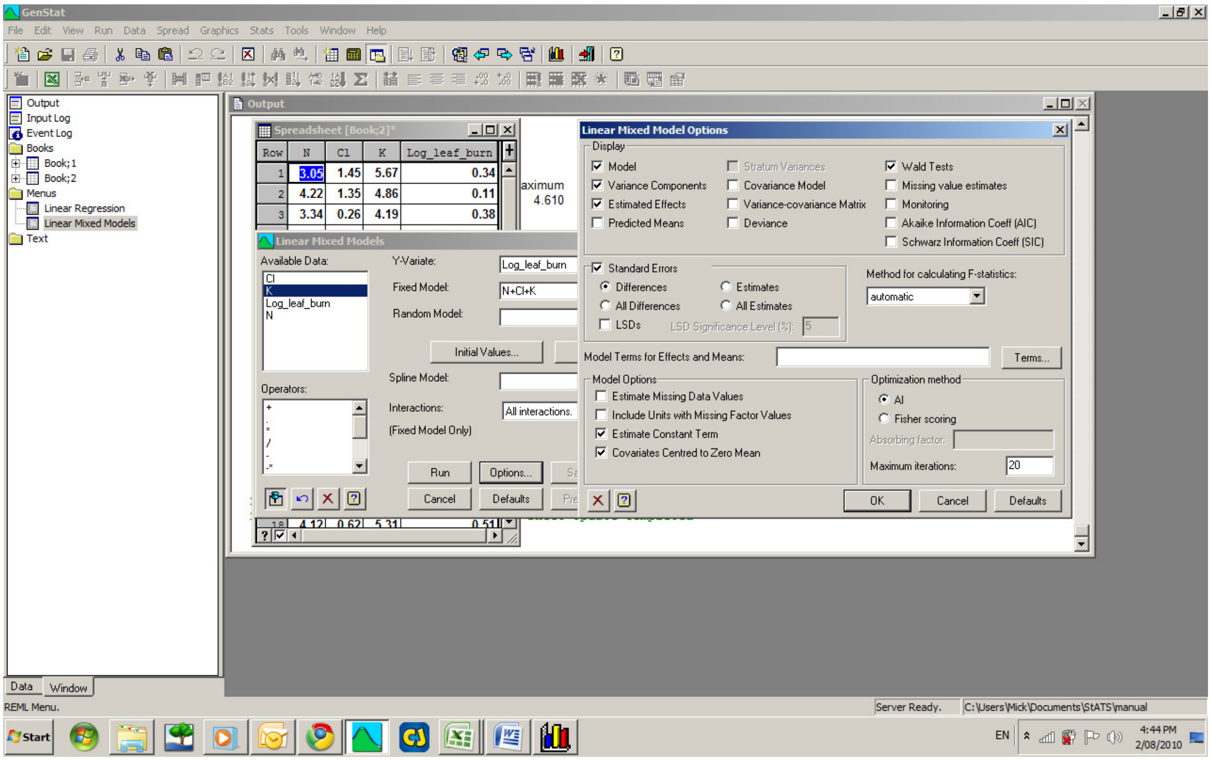
$$Y = \bar{y} - b_1\left(X_1 - \bar{x}_1\right) - \ldots - b_k\left(X_k - \bar{x}_k\right).$$

This simply emphasises predictor variates *centred to their mean*, and is an option in some GenStat procedures (eg Linear Mixed Models). For the current example, the re-arranged model is

log(leaf burn) = 0.686 - 0.5315 ($N$ – 3.2787) - 0.4396 ($Cl$ – 0.8077) + 0.2090 ($K$ – 4.6537)

## LMM (REML) output

REML will produce the centred form of the model as a default. The Fixed Model is the same as used in the regression menu. The Random Model is the Units factor, but since there is just the one residual term in the model it can be omitted.



Spreadsheet [Book;2]*

| Row | N | Cl | K | Log_leaf_burn |
|---|---|---|---|---|
| 1 | 3.05 | 1.45 | 5.67 | 0.34 |
| 2 | 4.22 | 1.35 | 4.86 | 0.11 |
| 3 | 3.34 | 0.26 | 4.19 | 0.38 |
| 18 | 4.12 | 0.62 | 5.31 | 0.51 |

Linear Mixed Models

Available Data: Cl, K, Log_leaf_burn, N

Y-Variate: Log_leaf_burn

Fixed Model: N+Cl+K

Random Model:

Spline Model:

Interactions: All interactions.

(Fixed Model Only)

Linear Mixed Model Options

Display

- ☑ Model
- ☑ Variance Components
- ☑ Estimated Effects
- □ Predicted Means
- □ Stratum Variances
- □ Covariance Model
- □ Variance-covariance Matrix
- □ Deviance
- ☑ Wald Tests
- □ Missing value estimates
- □ Monitoring
- □ Akaike Information Coeff (AIC)
- □ Schwarz Information Coeff (SIC)

☑ Standard Errors

- ● Differences
- ○ All Differences
- ○ Estimates
- ○ All Estimates
- □ LSDs  LSD Significance Level (%): 5

Method for calculating F-statistics: automatic

Model Terms for Effects and Means:

Model Options

- □ Estimate Missing Data Values
- □ Include Units with Missing Factor Values
- ☑ Estimate Constant Term
- ☑ Covariates Centred to Zero Mean

Optimization method

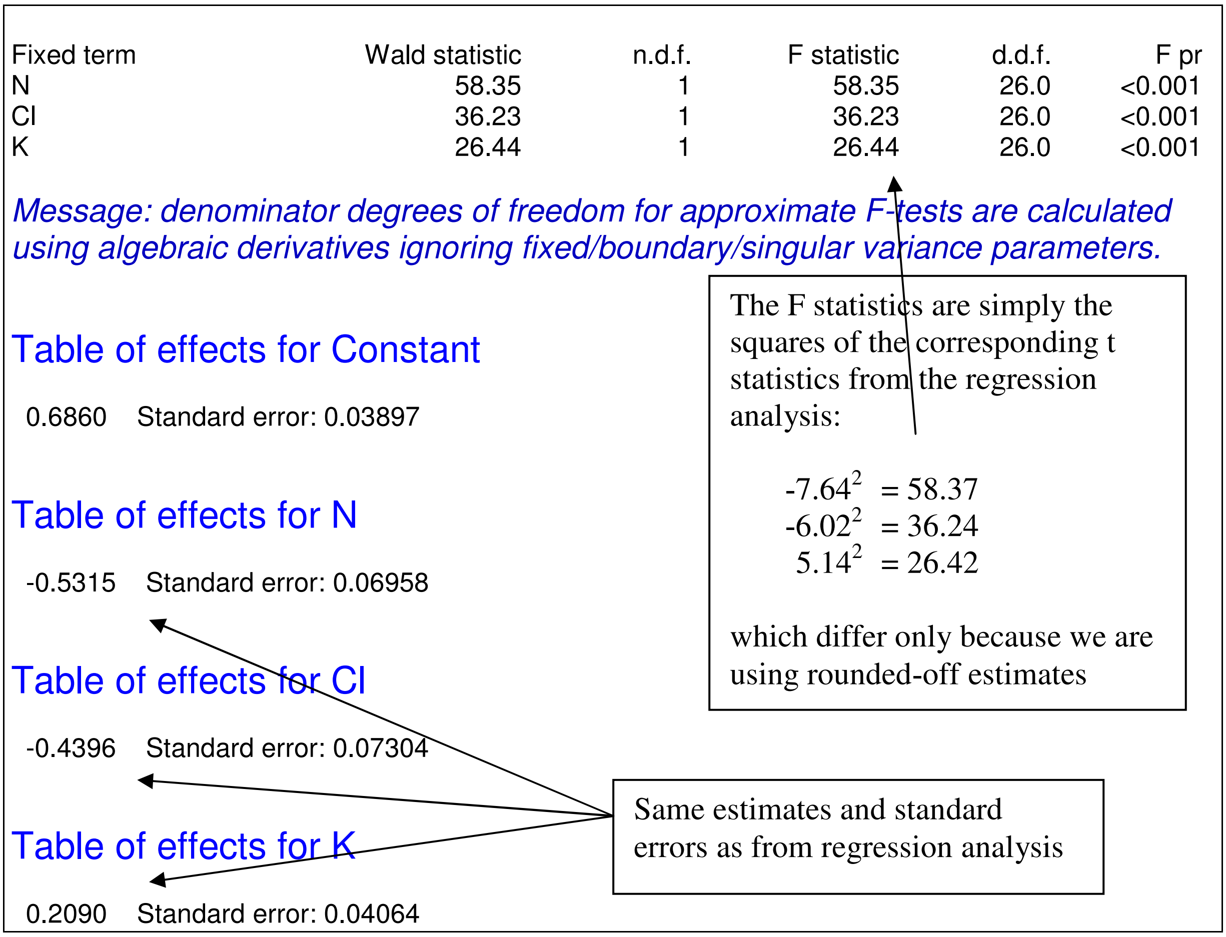
- ● AI
- ○ Fisher scoring

Absorbing factor:

Maximum iterations: 20

```
REML variance components analysis

Response variate:          Log_leaf_burn
Fixed model:               Constant + N + Cl + K
Number of units:           30

Residual term has been added to model

Sparse algorithm with AI optimisation
All covariates centred

Residual variance model

Term                          Factor  Model(order)    Parameter    Estimate        s.e.
Residual                                  Identity       Sigma2      0.0456     0.01264

Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term                Wald statistic    n.d.f.    F statistic      d.d.f.      F pr
N                                  75.62         1          75.62        26.0    <0.001
Cl                                 18.74         1          18.74        26.0    <0.001
K                                  26.44         1          26.44        26.0    <0.001

Dropping individual terms from full fixed model
```

Same estimate of $\sigma^2$ as from the regression ANOVA



```
Fixed term              Wald statistic      n.d.f.      F statistic     d.d.f.      F pr
N                                58.35           1            58.35       26.0    <0.001
Cl                               36.23           1            36.23       26.0    <0.001
K                                26.44           1            26.44       26.0    <0.001

Message: denominator degrees of freedom for approximate F-tests are calculated
using algebraic derivatives ignoring fixed/boundary/singular variance parameters.

Table of effects for Constant

 0.6860    Standard error: 0.03897

Table of effects for N

 -0.5315    Standard error: 0.06958

Table of effects for Cl

 -0.4396    Standard error: 0.07304

Table of effects for K

 0.2090    Standard error: 0.04064
```

The F statistics are simply the squares of the corresponding t statistics from the regression analysis:

$$\begin{aligned} -7.64^2 &= 58.37 \\ -6.02^2 &= 36.24 \\ 5.14^2 &= 26.42 \end{aligned}$$

which differ only because we are using rounded-off estimates

Same estimates and standard errors as from regression analysis

**Checking model assumptions**

Standard practice with any analysis is to check that model assumptions appear satisfactory.

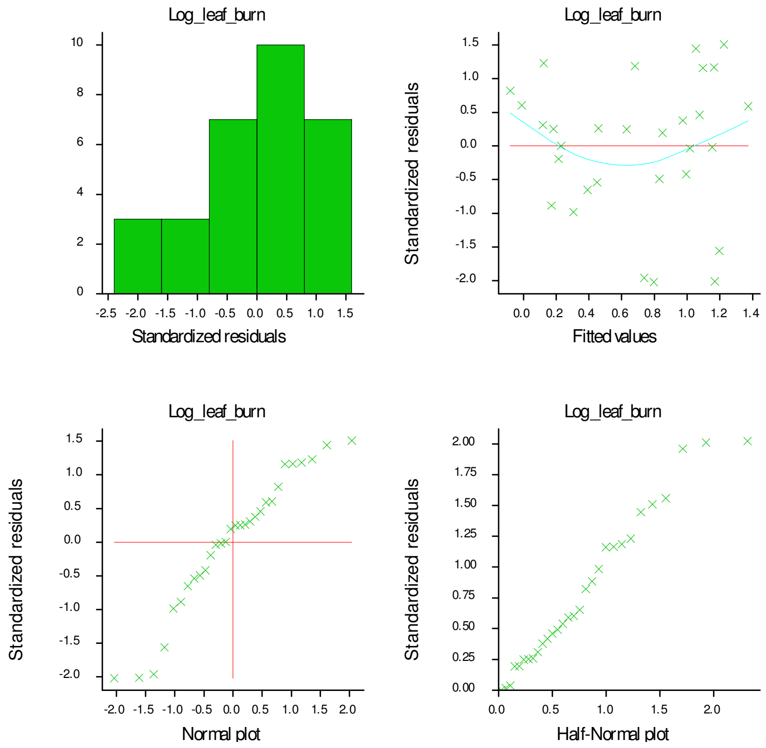
***Normality*.** This assumption is not the most critical assumption, but can be checked in GenStat using histograms or probability plots of residuals. Histograms are not particularly useful for small data sets.

***Constant variance*** A plot of standardised residuals against fitted values is one way to detect a problem with the variance assumption. The plot should show no trend, be randomly scattered around 0, with positive and negative values equally likely at any point. Most of the points should lie within ± 2. Fanning is indicative of data whose variance increases with the mean, and is often corrected by analyzing log-transformed data instead.

The nature of the treatments in an experiment may give rise to the suspicion that the variance may change. For example, a fanning residual plot may be due to the presence of a control treatment: plots untreated may just vary differently to treated plots. A more extreme example arises in say herbicide trials, where an increase in the amount of herbicide leads to a severe reduction in yield, with little variation. Log-transforming will not solve these problems: removing the control data is one solution, using a modern REML analysis with changing variance is preferable.

***Independence*** Lack of independence can be detected spatially by plotting residuals in field position (an option in GenStat's ANOVA menus). If there is a time element to the design, then a plot of residuals over time, or of residuals against lag-1 residuals, is valuable.

For the log(leaf burn) data, the 3-predictor model produces residuals show a very slight trend and possible fanning, but given the size of the dataset, there are no real concerns with any of the model assumptions. The plot is obtained *once the analysis is performed* by clicking on **Further Output > Model Checking**. For linear regression **Deviance** and **Pearson** residuals are standardized.

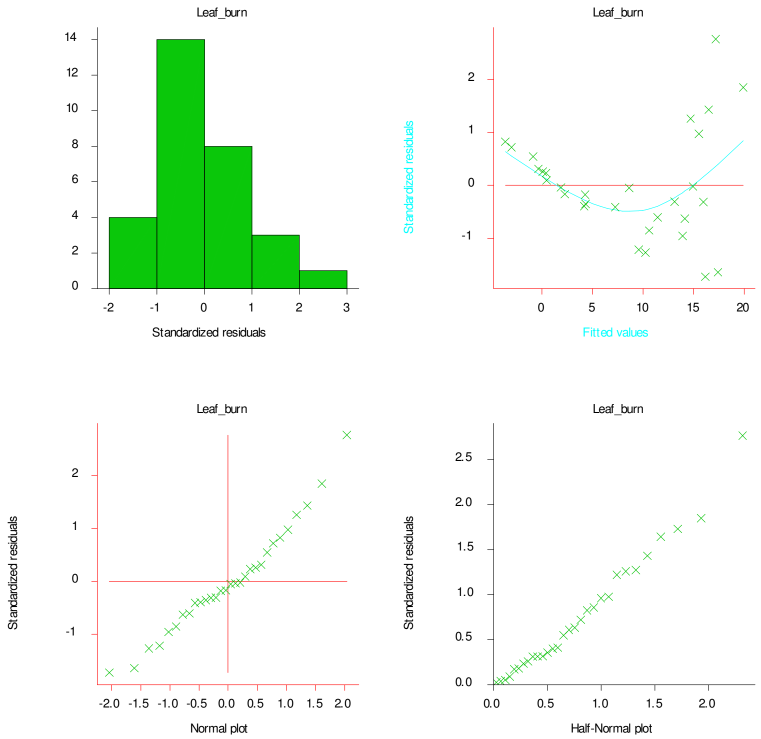


Standardised residual plot for the regression of log(leaf burn) on *N*, *Cl* and *K*.

Compare this residual plot with the following obtained from an analysis of untransformed data. There is a very marked trend and fanning, which led the researchers to transform leaf burn times.

GenStat will flag potential outliers (standardised residuals outside the ± 2 region) and influential points.

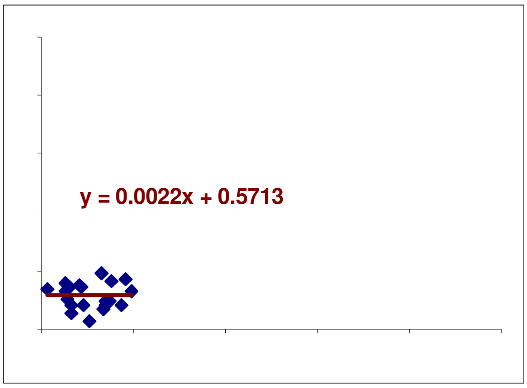
One such influential point was indicated in this analysis. What does it mean?

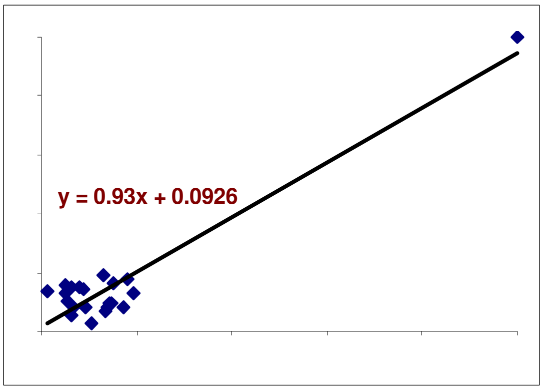


Standardised residual plot for the regression of leaf burn on N, Cl and K.

## Influential points

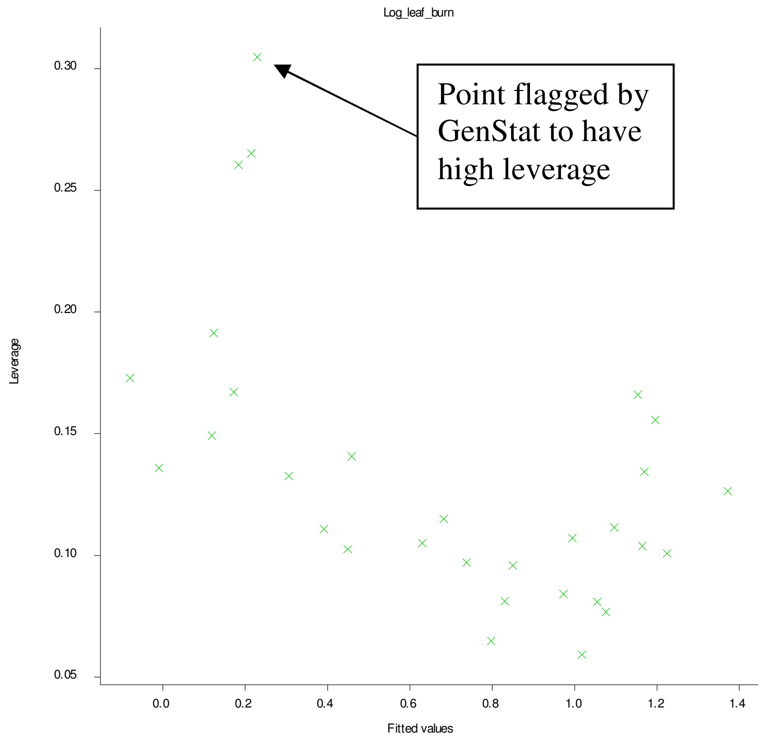
An influential point is one which has a strong influence in the fitting of the line. Consider the following hypothetical example.





The presence of just one point in the right hand diagram has dramatically affected the fitted model: the slope changes from about 0 without the point, to 1. That is not to say the point isn't important: outliers and influential points often tell you more about the system than the rest of the so-called "good data" (the discovery of the hole in the ozone layer being a dramatic example).

You can choose to plot leverages instead of standardised residuals. For the current example, two other data points appear to have high leverage, but obviously not high enough to fail GenStat's leverage test.



Point flagged by GenStat to have high leverage

Plot of leverages for the log(leaf burn) analysis

## The regression ANOVA

The Regression ANOVA actually tests whether *Y* is linearly dependent on the complete set of *X*s. It is not a position that in general we believe scientifically, but is often the starting point to model exploration. To drop all *X*s from the model we set up null and alternative hypotheses as follows:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \qquad \text{vs } H_1\text{: at least one } \beta \text{ parameter} \neq 0$$

## REGRESSION ANOVA for this set of hypotheses

| Source of Variation | *df* | *SS* | *MS* | *F* | *P* |
|---|---|---|---|---|---|
| Regression | *k* | Regression SS | $\frac{\text{Regression SS}}{\text{Regression df}}$ | $\frac{\text{Regression MS}}{\text{Residual MS}}$ | ✔ |
| Residual | *n*–*k*–1 | Residual SS | $\frac{\text{Residual SS}}{\text{Residual df}}$ | | |
| Total | *n*–1 | Total SS | Sample variance of all the data | | |

There are mathematical formulae in any standard text book for these sums of squares (*SS*) and mean squares (*MS*). Note that GenStat uses **v.r.** (variance ratio) for the *F* statistic (for that is what it is, a ratio of two potential estimates of the same variance), and **F.pr.** for the *P* value (since the *P* value is the probability of observing a variance ratio as large as, or larger than, the one observed, assuming an F distribution).

The *Total MS* is simply the *sample variance* of the log(leaf burn) data. Hence the name: *analysis of variance*.

```
Summary of analysis

Source          d.f.        s.s.        m.s.       v.r.    F pr.
Regression         3       5.505     1.83491      40.27    <.001
Residual          26       1.185     0.04557
Total             29       6.690     0.23067

Percentage variance accounted for 80.2
Standard error of observations is estimated to be 0.213.
```

The *Regression SS* is the variation in log(leaf burn) data that the model explains.

The *Residual SS* is the variation in log(leaf burn) data that the model fails to explain. It is exactly what it says. Calculate the residual for each observed value

$$
\begin{aligned}
\textit{Residual} &= \textit{Observed} - \textit{Fitted} \\
&= Y - [b_0 + b_1 X_1 + \ldots + b_k X_k] \\
&= \text{log(leaf burn)} - [1.811 - 0.5315\, N - 0.4396\, Cl + 0.2090\, K]
\end{aligned}
$$

then square each residual and sum the squared residuals.

### The coefficient of determination, $R^2$

Statistical packages generally offer two measures of the success of the model, often called $R^2$ and $R^2$(adjusted). They are fractions, but usually expressed as percentages.

$$\mathrm{R}^2 = \frac{\text{Regression SS}}{\text{Total SS}}$$

and is therefore a measure of the proportion of the total variability (as defined by sum of squares, not variance) explained by the regression model.

An alternative definition arises as follows. Since

Total SS = Regression SS + Residual SS,

$$\mathrm{R}^2 = \frac{\text{Total SS - Residual SS}}{\text{Total SS}} = 1 - \frac{\text{Residual SS}}{\text{Total SS}}$$

Given that the Total MS is the sample *variance*, and the Residual MS is an estimate of $\sigma^2$, the variance of a value of Y given the set of Xs, it is more natural to switch the last definition to *variances* rather than *sums of squares*. When you do this, the resulting statistic is less biased, and a better measure to use when comparing models with different numbers of parameters.

$$R^2_{\text{adj}} = 1 - \frac{\text{Residual MS}}{\text{Total MS}}$$

and is therefore a measure of the proportion of the total *variance* explained by the regression model. In fact, GenStat prefers to use the description **Percentage variance accounted for**, in this case, about 80%.

GenStat also presents $\sqrt{\text{Residual MS}}$ as Standard error of observations is estimated to be 0.213.

### Dropping a single predictor from a model

A general rule in statistics is that, for normally distributed statistics,

$$t_{\text{obs}} = \frac{statistic}{s.e.(statistic)} \sim \text{t variable}$$

and tests that mean value of the statistic = 0.

Hence, under the regression assumptions, dividing each parameter estimate by its standard error tests whether that *the mean of that parameter is zero*.

```
Estimates of parameters

Parameter               estimate          s.e.     t(26)    t pr.
Constant                   1.811         0.280      6.48    <.001
Nitrogen                 -0.5315        0.0696     -7.64    <.001
Chlorine                 -0.4396        0.0730     -6.02    <.001
Potassium                 0.2090        0.0406      5.14    <.001
```

Care must be taken with this table. Consider $H_0$: $\beta_1 = 0$ where $\beta_1$ is the coefficient of *N* in the regression model. This is tested using $t_{\text{obs}} = -0.5315/0.0696 = -7.64$, which is highly significant (*P*<0.001). This says that in a model involving *N*, *Cl* and *K*, *N* cannot be dropped, (the effect of allowing $\beta_1 = 0$ is effectively to drop the variate from the model, providing that *Cl* and *K* remain in the model). Proceeding to ask whether chlorine can be dropped is dangerous: this test assumes *N* and *K* remain.

## Redundant predictor variables

An aliased (or redundant) predictor occurs when a set of variables already included in a model completely explain the values of a new predictor. A simple example is as follows.

Suppose you have in mind a 2-variable model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + error$$

This apparently will explain 2 df. However, suppose that $X_2$ and $X_1$ are linearly related:

$X_2 = a + b\, X_1$

Then the original model only *apparently* involves two independent variates. In fact there is just one:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + error \\ &= \beta_0 + \beta_1 X_1 + \beta_2 (a + b\, X_1) + error \\ &= (\beta_0 + a\, \beta_2) + (\beta_1 + b\, \beta_2 X_1) + error \\ &= \beta_0^* + \beta_1^* X_1 + error \end{aligned}$$

GenStat is helpful, in that it tells you the relationship between the predictor variables in the process of removing redundant predictors.

Allen and Cady (1982) have an example where water samples were taken along a river. A land survey was conducted at each sampling site, and the percentage of land allocated to agriculture, residential, industrial and forest use recorded. A fifth variate, *other*, was included. Thus at each site,

$$agriculture + residential + industrial + forest + other = 100\%.$$

The fifth variate *other* is redundant. If you did include this variate with the other four predictors, GenStat would respond:

```
Message: term Other cannot be included in the model because it is aliased with
terms already in the model.

(Other) = 100.0 - (Agriculture) - (Forest) - (Industrial) - (Residential)
```

The resulting model and analysis has just the first four predictors mentioned:

```
Regression analysis

        Response variate: TOTAL_N
              Fitted terms: Constant + Agriculture + Forest + Industrial + Residential

Summary of analysis
Source           d.f.          s.s.          m.s.       v.r.    F pr.
Regression          4         2.570       0.64246       9.15    <.001
Residual           15         1.053       0.07018
Total              19         3.623       0.19066
```

```
Percentage variance accounted for 63.2
Standard error of observations is estimated to be 0.265.

Estimates of parameters
Parameter              estimate          s.e.     t(15)    t pr.
Constant                   1.72          1.23      1.40    0.183
Agriculture              0.0058        0.0150      0.39    0.705
Forest                  -0.0130        0.0139     -0.93    0.367
Industrial                0.305         0.164      1.86    0.082
Residential             -0.0072        0.0338     -0.21    0.834
```

Another example of a redundant predictor occurs when a factor is included in a model which also contains the overall mean (that is, the Constant in the regression). This will be demonstrated in the section **Regression with groups (factors)**. Take as an example a general regression involving a single factor say Sex (Male/Female). While this factor has 2 "levels", only 1 degree of freedom is available in the regression. For example, if level 1 represents a male, then is the value in the factor column is not a 1, it must be a female. This works no matter how many levels the factor has. In the regression output, the model involving a factor with *t* levels will contain *t*-1 parameters. There will be a model for the "default" level of the factor (which can be changed in Spread > Column > Attributes/Format (F9 is the shortcut) or Spread > Factor > Reference Level.

### LMM (REML) analysis with redundant predictors

GenStat allows all 5 predictors but the final predictor is simply not estimated. In this example we changed the default option by turning off Covariates Centred to zero Mean.

```
REML variance components analysis

Response variate:       TOTAL_N
Fixed model:            Constant + Agriculture + Forest + Industrial + Residential + Other
Number of units:        20

Residual term has been added to model

Sparse algorithm with AI optimisation
Covariates not centred

Residual variance model

Term                          Factor  Model(order)     Parameter    Estimate        s.e.
Residual                                  Identity        Sigma2      0.0702     0.02563

Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term               Wald statistic     n.d.f.   F statistic      d.d.f.        F pr
Agriculture                        8.30          1          8.30        15.0       0.011
Forest                            24.15          1         24.15        15.0      <0.001
Industrial                         4.13          1          4.13        15.0       0.060
Residential                        0.05          1          0.05        15.0       0.834
Other                              0.00          0          0.00        15.0           *
```

```
Dropping individual terms from full fixed model

Fixed term                Wald statistic        n.d.f.        F statistic        d.d.f.        F pr
Agriculture                         0.15             1               0.15          15.0       0.705
Forest                              0.87             1               0.87          15.0       0.367
Industrial                          3.47             1               3.47          15.0       0.082
Residential                         0.05             1               0.05          15.0       0.834

Table of effects for Constant

 1.722    Standard error: 1.2341


Table of effects for Agriculture

 0.005809    Standard error: 0.0150340

Table of effects for Forest

 -0.01297    Standard error: 0.013931

Table of effects for Industrial

 0.3050    Standard error: 0.16382

Table of effects for Residential

 -0.007227    Standard error: 0.0338301

Table of effects for Other

          0.0

Standard error is not available.
```

Use these P values for dropping a single predictor. They are identical to the P values from the regression t tests above.

## Dropping several predictors from a model

A more general test is possible. What if we were to ask whether a *subset* of the predictor variables can be omitted? For example, can we drop both agriculture and residential predictors form the 4-variate predictor model?

This is equivalent to the following in general.

We have a ***maximal model*** involving $k$ conceivable predictor or explanatory variables. These are ordered for convenience only. We are interested in dropping the last $s$ predictors from this model.

$$Y = (\beta_0 + \beta_1 X_1 + \ldots + \beta_{k-s} X_{k-s} + \beta_{k-s+1} X_{k-s+1} + \ldots + \beta_k X_k) + error$$

Dropping the last $s$ predictors (for convenience) gives rise to a ***reduced model***

$$Y = (\beta_0 + \beta_1 X_1 + \ldots + \beta_{k-s} X_{k-s}) \qquad\qquad + error$$

This is equivalent to testing $H_0$: $\beta_{k-s+1} = \ldots = \beta_k = 0$. The way we test this is as follows.

**Step 1.** Fit the *maximal* model and note the *Regression SS* and *Residual MS* in the ANOVA.

**Step 2.** Drop the (potentially) superfluous predictors and fit the *reduced* model.

**Step 3.** Calculate the *change in Residual SS* between the maximal and reduced models, form the *mean square* by diving by the change in degrees of freedom and test this against the *Residual MS* from the full model.

Testing $H_0$: $\beta_{k-s+1} = \ldots = \beta_k = 0$

| Regression Analysis | Source of Variation | *SS* | *df* | *MS* | *F* | *P* |
|---|---|---|---|---|---|---|
| *Maximal* | Using all $k$ Xs | $Reg\ SS_{Full}$ | $k$ | | *ignore* | |
| *Reduced* | Using first ($k$-$s$) Xs | $ReS\ SS_{Reduced}$ | $k$-$s$ | | *ignore* | |
| **Calculate by differencing** | ***Lack of fit*** | ***Diff.*** | $s$ | $\frac{diff.}{s}$ | $\frac{Lack\ of\ fit\ MS}{Res\ MS}$ | ✔ |
| *Maximal* | Residual | *Res SS* | $\boldsymbol{n - k - 1}$ | *Res MS* | | |
| *Maximal* | Total | *Tot SS* | $n - 1$ | | | |

For non-normal data, or for testing random effects in Linear Mixed Models (REML), the equivalent technique is known as *change in deviance*. More about this later.

Firstly, let us illustrate this with the question: can we drop the *Industrial* predictor and leave *Agriculture*, *Forest* and *Residential* predictors. This is equivalent to

$H_0$: $\beta_3 = 0$ vs $H_3$: $\beta_1 \neq 0$ (assuming *Industrial* is the third predictor mentioned in the model)

**Step 1** Fit the *maximal* model with all 4 predictors and note the ANOVA:

| Source | d.f. | s.s. | m.s. | v.r. | F pr. |
|---|---|---|---|---|---|
| Regression | 4 | **2.570** | ignore | ignore | ignore |
| Residual | 15 | 1.053 | **0.07018** | | |
| Total | 19 | 3.623 | 0.19066 | | |

**Step 2** Drop *Industrial*, fit the *reduced* model using the remaining 3 predictors only and note the new Residual SS:

| Source | d.f. | s.s. |
|---|---|---|
| Residual | 16 | 1.296 |

**Step 3** Calculate the change in Residual SS and df, and construct a variance ratio:

| Source | d.f. | s.s. | m.s. | v.r. | F pr. |
|---|---|---|---|---|---|
| Regression using 4 predictors | 4 | 2.57 | ignore | ignore | ignore |
| Residual using 3 predictors | 16 | 1.296 | | | |
| **Change** | **1** | 0.243 | **0.243** | 3.463 | 0.082 |
| Residual using 4 predictors | 15 | 1.053 | **0.07018** | | |
| Total | 29 | 6.690 | 0.23067 | | |

Clearly, there is no statistical evidence ($P$=0.277) to retain the *Industrial* predictor in the model ***providing the other three predictors are retained***.

Note that 0.082 is the P value, and $\sqrt{3.463} = 1.86$, the value we have alongside the parameter estimate in the 4-predictor regression output.

### Estimates of parameters

| Parameter | estimate | s.e. | t(15) | t pr. |
|---|---|---|---|---|
| Constant | 1.72 | 1.23 | 1.40 | 0.183 |
| Agriculture | 0.0058 | 0.0150 | 0.39 | 0.705 |
| Forest | -0.0130 | 0.0139 | -0.93 | 0.367 |
| **Industrial** | **0.305** | **0.164** | **1.86** | **0.082** |
| Residential | -0.0072 | 0.0338 | -0.21 | 0.834 |

Note also that the P value 0.082 and the F value 3.463 are the same as from the Wald statistic from the REML output:

Dropping individual terms from full fixed model

| Fixed term | Wald statistic | n.d.f. | F statistic | d.d.f. | F pr |
|---|---|---|---|---|---|
| Agriculture | 0.15 | 1 | 0.15 | 15.0 | 0.705 |
| Forest | 0.87 | 1 | 0.87 | 15.0 | 0.367 |
| **Industrial** | **3.47** | **1** | **3.47** | **15.0** | **0.082** |
| Residential | 0.05 | 1 | 0.05 | 15.0 | 0.834 |

Hence the general technique for dropping a number (≥1) of predictors gives statistics that are identical to others already produced for this situation.

Next, let us ask the question: can we drop (say) the *Forest* and *Industrial* predictors from the model? This is equivalent to

$H_0$: $\beta_2=\beta_3 = 0$ vs $H_1$: either $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$ (assuming *Forest* is second and *Industrial* third in the list).

**Step 1** Fit the *maximal* model with all 4 predictors and note the ANOVA:

| Source | d.f. | s.s. | m.s. | v.r. | F pr. |
|---|---|---|---|---|---|
| Regression | 4 | **2.570** | ignore | ignore | ignore |
| Residual | 15 | 1.053 | **0.07018** | | |
| Total | 19 | 3.623 | 0.19066 | | |

**Step 2** Drop *Forest* and *Industrial*, fit the *reduced* model using the remaining 2 predictors only and note the new Residual SS:

| Source | d.f. | s.s. |
|---|---|---|
| Residual | 17 | 1.347 |

**Step 3** Calculate the change in Residual SS and df, and construct a variance ratio:

| Source | d.f. | s.s. | m.s. | v.r. | F pr. |
|---|---|---|---|---|---|
| Regression using 4 predictors | 4 | 2.57 | ignore | ignore | ignore |
| Residual using 3 predictors | 17 | 1.347 | | | |
| **Change** | **2** | 0.294 | **0.147** | 2.095 | 0.158 |
| Residual using 4 predictors | 15 | 1.053 | **0.07018** | | |
| | | | | | |
| Total | 29 | 6.690 | 0.23067 | | |

Again, there is no statistical evidence (P=0.158) that to retain both *Forest* and *Industrial* predictors in the model ***providing the other two predictors are retained***.

## Stepwise Regression – automatic selection of predictor variables

In the previous section we examined whether a particular 3-predictor model is statistically as good as the 4-predictor model. We repeated the illustration with particular 3-predictor model. We had no idea what predictors to drop and what predictors to retain.

Stepwise regression is a procedure for *automatic selection of potentially important predictors*. It is especially useful in the early stages of research for forming potential research hypotheses for further esamination.

Firstly, from the **Linear Regression** menu select **General Linear Regression** procedure. Enter the response variable to be analysed, and enter **all potential predictors in the Maximal Model**, separated by **+** or **,**. There are several approaches that GenStat offers. Occasionally a different final model is obtained by the different methods.

1. ***Forward Selection.***
   Start with *no predictor variables* in the model. Use Change Model to sequentially *add* predictors to the model, one at a time, with the most significant predictor going in first. Predictors are added until no further 'significant' predictor is left to be added.

2. ***Backward Elimination***:
   Start with ***all** predictor variables* in the model. Use Change Model to sequentially *drop* predictors from the model, one at a time, with the least significant predictor dropped first. Predictors are removed until no further 'non significant' predictor can be removed. What remains are the significant predictors.

3. ***Stepwise***: Combines aspects of both forwards selection and backwards elimination. Start with *no predictor variables* in the model. Use Change Model to add the best single predictor, then, in steps, sequentially remove existing predictors if a statistically worse model is not produced, or add new predictors if a statistically better model is produced.

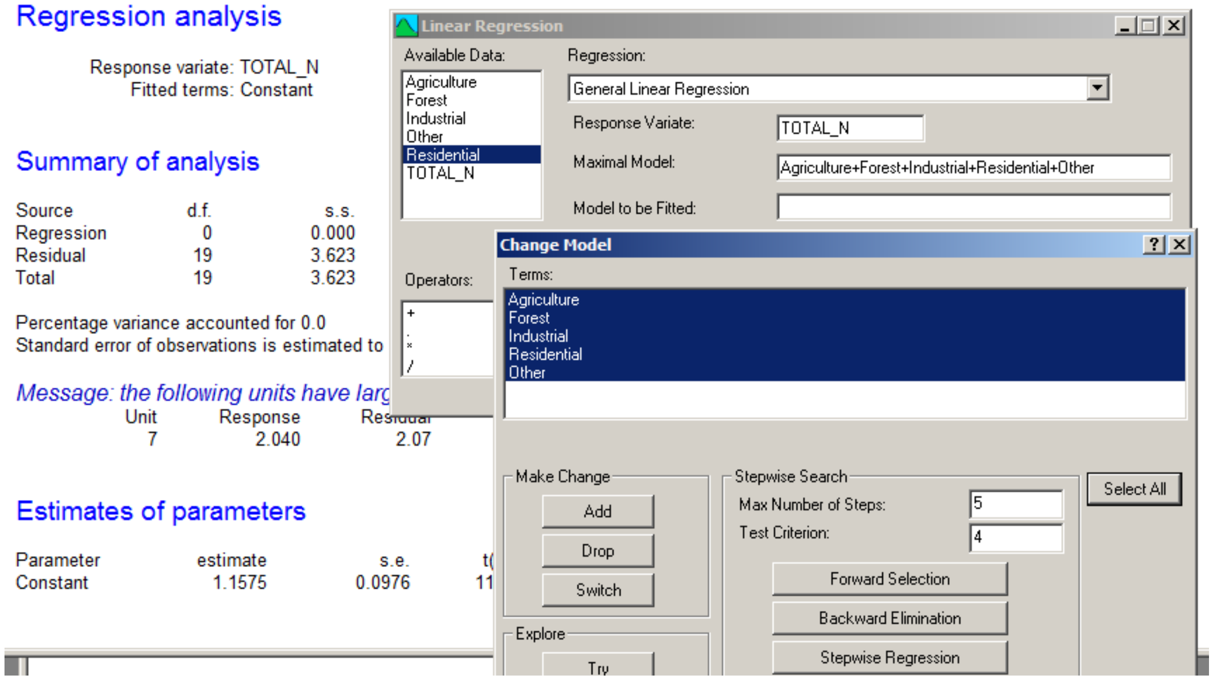
In each case, in the Change Model menu you need to

- select the predictors you wish to explore – typically by clicking **Select All**;

- set the **Max Number of Steps** you wish GenStat to use – typically the same as the number of variates and factors;

- set the **Test Criterion**. Note that the criterion of "significant" terms in the model is somewhat problematic here. Since repeated testing on the same variables is being conducted, the usual significance levels do not really apply here. Frequently, rather than using *P*-values as the testing criteria, fixed critical *F*-values (**Test Criterion**) are used which are not based on the actual *F*-distribution. Typically, criterion values of **4.0** are used for stepwise methods. Why? Dropping or adding a single predictor would lead to a t test with 1 numerator *df*; an $F_{1,v}$ critical value is the same as a (t critical value)$^2$ which tends to $(1.96)^2 \approx (2)^2 = 4$, so that value makes sense. GenStat defaults to a **Test Criterion** of 1. If this is an early stage in your research, that value may be acceptable, but several predictors will be entered that are unlikely to pass the test of time.

The output for the **stepwise regression** is as follows. Here we have used the original 5 predictors which add to 100% in case the predictor *Other* is important (we would then need to think through what other land use this predictor represents).



```
Regression analysis

    Response variate: TOTAL_N
        Fitted terms: Constant

Summary of analysis

Source          d.f.        s.s.
Regression        0        0.000
Residual         19        3.623
Total            19        3.623

Percentage variance accounted for 0.0
Standard error of observations is estimated to

Message: the following units have larg
        Unit      Response      Residual
           7         2.040          2.07

Estimates of parameters

Parameter        estimate        s.e.     t(
Constant           1.1575      0.0976     11
```

Linear Regression

Available Data: Agriculture, Forest, Industrial, Other, Residential, TOTAL_N

Regression: General Linear Regression

Response Variate: TOTAL_N

Maximal Model: Agriculture+Forest+Industrial+Residential+Other

Model to be Fitted:

Operators: + . * /

Change Model

Terms: Agriculture, Forest, Industrial, Residential, Other

Max Number of Steps: 5

Test Criterion: 4

## Step 1: Residual mean squares

```
0.08086   Adding    Forest
0.13668   Adding    Residential
0.14076   Adding    Other
0.14419   Adding    Industrial
0.16890   Adding    Agriculture
0.19066   No change
```

**Chosen action: adding Forest.**

*Forest* is the strongest single predictor of the 5 potential predictors

## Step 2: Residual mean squares

```
0.06531   Adding    Industrial
0.07663   Adding    Residential
0.07915   Adding    Agriculture
0.08086   No change
0.08378   Adding    Other
0.19066   Dropping  Forest
```

**Chosen action: adding Industrial.**

*Industrial* is also a statistically important predictor and adds additional predictive power

## Step 3: Residual mean squares

```
0.06531   No change
0.06600   Adding    Agriculture
0.06645   Adding    Residential
0.06708   Adding    Other
0.08086   Dropping  Industrial
0.14419   Dropping  Forest
```

**Chosen action: no change.**

No dropping or further addition of a predictor leads to an improved model

Notice that GenStat prints out the Res MS from each analysis it trials, tests the change in Res SS between the previous model and the new model, and orders the variates by the significance of the impact of the proposed action. We are told what the chosen action is. To obtain the final model, we return to regression and fit the suggested model

```
Regression analysis

        Response variate: TOTAL_N
              Fitted terms: Constant + Forest + Industrial

Summary of analysis

Source              d.f.          s.s.          m.s.        v.r.    F pr.
Regression             2         2.512       1.25619       19.24    <.001
Residual              17         1.110       0.06531
Total                 19         3.623       0.19066

Percentage variance accounted for 65.7
Standard error of observations is estimated to be 0.256.

Message: the following units have large standardized residuals.
              Unit        Response         Residual
                 7           2.040             2.96
                19           0.660            -2.20

Message: the following units have high leverage.
              Unit        Response         Leverage
                 4           1.000             0.43
                 5           1.990             0.72

Estimates of parameters

Parameter              estimate          s.e.     t(17)     t pr.
Constant                  2.096         0.240      8.72     <.001
Forest                 -0.01648       0.00345     -4.77     <.001
Industrial               0.1877        0.0816      2.30     0.034
```

Thus the fitted model

$$\text{Total N} \;=\; 2.096 - 0.01648\,\textit{Forest} + 0.1877\,\textit{Industrial}$$

explains over 65% of the variation in total N. Keeping the industrial land usage the same and increasing the area dedicated to forests by 1% *lowers* total N by 0.016 units. On the other hand, increasing land use for industrial purposes by 1% and maintaining forest land use *increases* total N by 0.188 units.

### Regression with groups (factors)

One of GenStat's great strengths is its ability to allow any of the predictors to be a *factor*. Remember, a factor is a column whose entries simply identify different conditions. So Variety 1, 2, 3 is a factor; there is no relation necessarily between the 1 and 2, 2 and 3: they could have been labels A, B, C.

Mead and Curnow present the numbers of leaves (labelled Number, averaged from 10 cauliflower plants) in each of two years, and wished to relate cauliflower growth with temperature (labelled DD, measured in day degrees above 32 °F, divided by100).

Example 4. From Mead and Curnow (1990 Page 161)

| 1956/7 season | | 1957/8 season | |
|---|---|---|---|
| DD | Number | DD | Number |
| 4.5 | 3.8 | 4.5 | 6.0 |
| 7.5 | 6.2 | 8.0 | 8.5 |
| 9.5 | 7.2 | 9.5 | 9.1 |
| 10.5 | 8.7 | 11.5 | 12.0 |
| 13.0 | 10.2 | 13.0 | 12.6 |
| 16.0 | 13.5 | 14.0 | 13.3 |
| 18.0 | 15.0 | 16.5 | 15.2 |

Interest lay in which model best describes both years (year = 1, 2) of data:

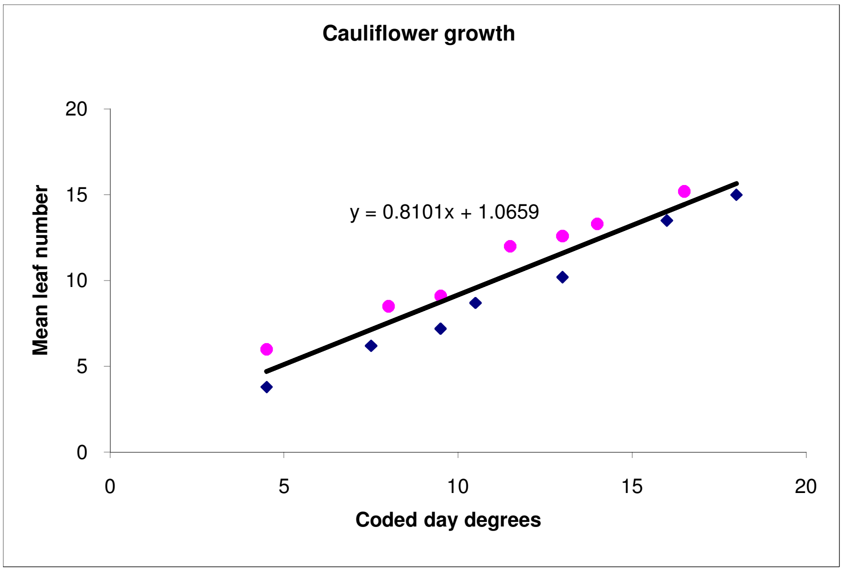
Common line: Mean leaf number = $b_0 + b_1$ DD 1 intercept+1 slope = 2 parameters
Parallel lines: Mean leaf number = $b_{0,\text{year}} + b_1$ DD 2 intercepts+1 slope = 3 parameters
Separate lines: Mean leaf number = $b_{0,\text{year}} + b_{1,\text{year}}$ DD 2 intercepts+2 slopes = 4 parameters

These are simple applications of testing various reduced models in a general linear model framework. The maximal model is the *separate* lines situation.

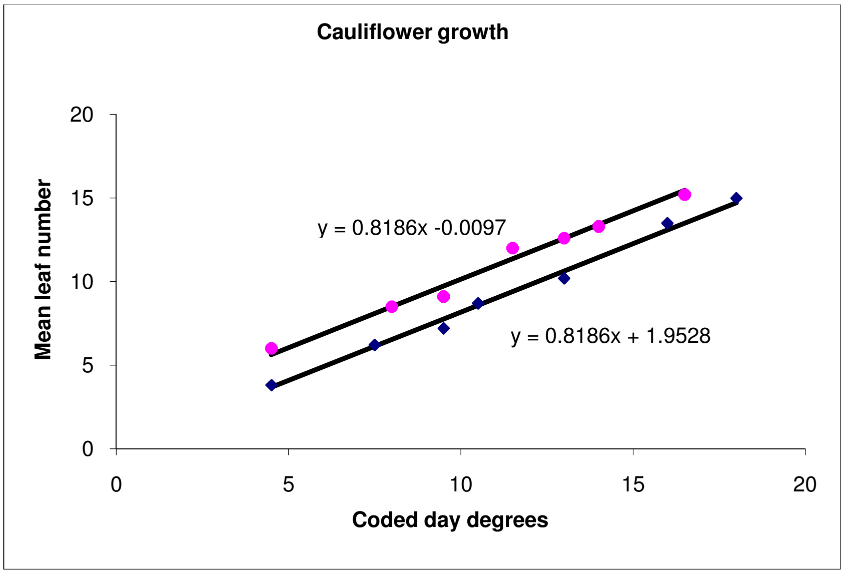
To compare the two regression lines in GenStat the data need to be stacked first, and a factor column created to identify *year*. Note that the data being analysed, average number of leaves, is related to a Poisson distribution. In fact, if the numbers of leaves on one plant is Poisson with mean μ, then the total numbers of leaves on 10 plants is Poisson with mean 10μ. The variance of a Poisson distribution is the same as the mean, so if the mean changes across day degrees or years, so must the variance. Hence, we might anticipate that the residual plots following regression will cast doubt about the constant variance assumption. To overcome this problem, we should analyse the data using log-linear modelling (to be done later).

| DD | Number | Year |
|---|---|---|
| 4.5 | 3.8 | 1956_7 |
| 7.5 | 6.2 | 1956_7 |
| 9.5 | 7.2 | 1956_7 |
| 10.5 | 8.7 | 1956_7 |
| 13.0 | 10.2 | 1956_7 |
| 16.0 | 13.5 | 1956_7 |
| 18.0 | 15.0 | 1956_7 |
| 4.5 | 6.0 | 1957_8 |
| 8.0 | 8.5 | 1957_8 |
| 9.5 | 9.1 | 1957_8 |
| 11.5 | 12.0 | 1957_8 |
| 13.0 | 12.6 | 1957_8 |
| 14.0 | 13.3 | 1957_8 |
| 16.5 | 15.2 | 1957_8 |

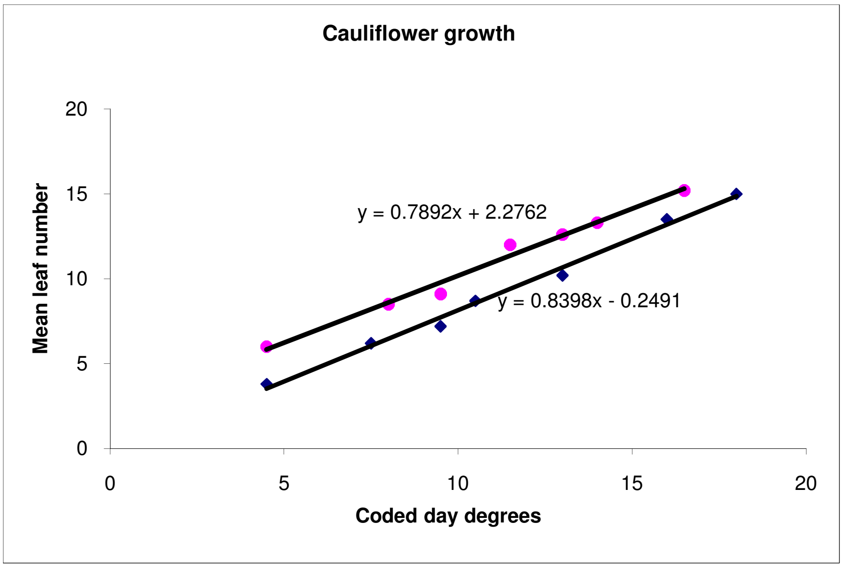
A *common* regression line is obtained by simply analysing the stacked data ignoring the year factor; the model involves just **DD**.



common line



parallel lines



separate lines

*Parallel* regression lines are obtained by adding the factor **Year** to the model to be fitted. GenStat will give output for a reference line, and the regression coefficients allow adjustments to be made for the *intercept* for the other levels of the included factor.

```
Regression analysis – output for parallel lines

        Response variate:  Number
            Fitted terms:  Constant + DD + Year

Summary of analysis
Source              d.f.          s.s.          m.s.        v.r.    F pr.
Regression             2       165.532       82.7660      506.57    <.001
Residual              11         1.797        0.1634
Total                 13       167.329       12.8715

Percentage variance accounted for 98.7
Standard error of observations is estimated to be 0.404.

Estimates of parameters
Parameter                       estimate          s.e.       t(11)    t pr.
Constant                          -0.010         0.337       -0.03    0.978
DD                                0.8186        0.0266       30.81    <.001
Year 1957_8                        1.962         0.216        9.08    <.001

Parameters for factors are differences compared with the reference level:
                   Factor   Reference level
                     Year   1956_7

Accumulated analysis of variance
Change                  d.f.           s.s.           m.s.        v.r.        F pr.
+ DD                       1       152.0694       152.0694      930.74        <.001
+ Year                     1        13.4626        13.4626       82.40        <.001
Residual                  11         1.7972         0.1634
Total                     13       167.3293        12.8715
```

*Separate* regression lines are obtained by adding the factor **DD.Year** to the model to be fitted in addition to **Year**. GenStat will give output for a reference line, and the regression coefficients allow adjustments to be made for the *intercept* and for the *slope* for the other levels of the included factor. Note that you now have a model

**DD + Year + DD.Year**

which can be shortened to **DD*Year**. More of this later in the design section.

Basically, when you have a factor with say *t* levels, GenStat uses *t* columns, each column representing a different level of the factor, with a value +1 for an observation belonging to that level of the factor, and a 0 otherwise.

The model that GenStat prints out is appropriate for the “reference” level it chooses. You can change this reference level if you wish (eg click in the factor column of the spreadsheet and hit **F9**).

There is a procedure which tests whether common, parallel or separate models are better: namely, **Linear Regression with Groups**. There are three parts to the output. There is strong evidence (P<0.001) to conclude that parallel regression lines are necessary, but no significant evidence (P=0.372) that separate lines are needed. On average, there are two extra leaves per cauliflower in the first season, however, growth over the season is similar, with about 82 leaves added for a 100 increase in (coded) day degrees.

Once the comparisons are done, you can choose which model to go with. In fact, you have the choice of re-running (and plotting, in **Further Output**) the chosen analysis so that the actual models are printed out (together with standard errors of all the intercepts and slopes) for the different factor levels, not just as they differ from the reference model.

**Part 1 –common model**

```
         Response variate:  Number
             Fitted terms:  Constant + DD

Summary of analysis
Source              d.f.          s.s.          m.s.       v.r.    F pr.
Regression             1        152.07       152.069     119.58    <.001
Residual              12         15.26         1.272
Total                 13        167.33        12.871

...
Estimates of parameters
Parameter                           estimate          s.e.      t(12)    t pr.
Constant                               1.066         0.879       1.21    0.248
DD                                    0.8101        0.0741      10.94    <.001
```

**Part 2 – parallel models**

```
         Response variate:  Number
             Fitted terms:  Constant + DD + Year

Summary of analysis
Source              d.f.          s.s.          m.s.       v.r.    F pr.
Regression             2       165.532       82.7660     506.57    <.001
Residual              11         1.797        0.1634
Total                 13       167.329       12.8715
Change                -1       -13.463       13.4626      82.40    <.001

...
Estimates of parameters
Parameter                           estimate          s.e.      t(11)    t pr.
Constant                              -0.010         0.337      -0.03    0.978
DD                                    0.8186        0.0266      30.81    <.001
Year 1957_8                            1.962         0.216       9.08    <.001
```

**Part 3 – separate models**

```
            Response variate:  Number
                Fitted terms:  Constant + DD + Year + DD.Year

Summary of analysis
Source               d.f.          s.s.           m.s.       v.r.    F pr.
Regression              3       165.676        55.2255     334.12    <.001
Residual               10         1.653         0.1653
Total                  13       167.329        12.8715
Change                 -1        -0.144         0.1444       0.87    0.372

Estimates of parameters
Parameter                            estimate          s.e.      t(10)    t pr.
Constant                               -0.249         0.425      -0.59    0.570
DD                                     0.8398        0.0351      23.95    <.001
Year 1957_8                             2.525         0.640       3.94    0.003
DD.Year 1957_8                        -0.0506        0.0542      -0.93    0.372
```

**Interpretation**

**Part 1** The common model (same intercept and slope for both years) is

Mean leaf number = 1.066 + 0.8101 DD

**Part 2** The parallel models (different intercepts and same slope) are

Mean leaf number = -0.010 + 0.8186 DD for 1956/7, the reference year,

and

Mean leaf number = (-0.010 + 1.962) + 0.8186 DD.
= 1.952 + 0.8186 DD for 1957/8

This is a statistically superior model compared to a common model (F=82.40, P<0.001).

**Part 3** The separate models (different intercepts and different slopes) are

Mean leaf number = -0.249 + 0.8398 DD for 1956/7,

and

Mean leaf number = (-0. 249 + 2.525) + (0.8186 + 0.0542) DD
= 2.276 + 0.8728 DD for 1957/8

This model is no better statistically than the parallel models (F=0.87, P=0.372).

Re-running the model choosing **Parallel lines, estimate lines** gives the actual two intercepts and common slope to save you having to construct the lines yourself:

```
Estimates of parameters

Parameter                  estimate            s.e.      t(11)     t pr.
DD                           0.8186          0.0266      30.81     <.001
Season 1956_7                -0.010           0.337      -0.03     0.978
Season 1957_8                 1.953           0.330       5.92     <.001
```

## Polynomial regression

A plot of the pasture data of Example 5 shows a strong linear trend with a sigmoid shape typical of plants growing over time. Again ignoring any variance problem, polynomial regression can be used, though a more biologically meaningful model may be available.

Polynomial regression is performed using simple or general linear regression, replacing **time** with a function **pol(time;3)**, where 3 governs the degree of the polynomial. We choose 3 with these data, anticipating the curvature at both ends.

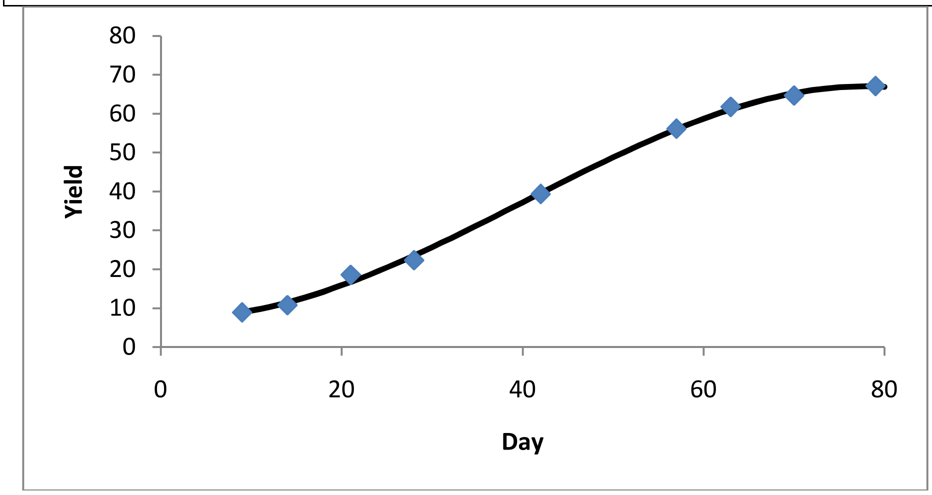
While the model explains 99.78% of the variation in yield, it is still only a mathematical approximation for growth over the period 9 to 79 days. The fitted model (plotted below) is

| time | yield |
|---|---|
| 9 | 8.93 |
| 14 | 10.80 |
| 21 | 18.59 |
| 28 | 22.33 |
| 42 | 39.35 |
| 57 | 56.11 |
| 63 | 61.73 |
| 70 | 64.62 |
| 79 | 67.08 |

Example 5. Pasture data from Ratkowsky (1990)

$$\text{Yield} = 7.8838 - 0.15728\,\textit{time} + 0.03336\,\textit{time}^2 - 0.00028\,\textit{time}^3$$

```
Regression analysis

          Response variate:  yield
              Fitted terms:  Constant + time
                 Submodels:  POL(time; 3)

Summary of analysis
Source              d.f.          s.s.          m.s.       v.r.    F pr.
Regression             3      4641.734      1547.245    1222.20    <.001
Residual               5         6.330         1.266
Total                  8      4648.063       581.008

Percentage variance accounted for 99.8
Standard error of observations is estimated to be 1.13.
...
Estimates of parameters
Parameter                        estimate          s.e.       t(5)    t pr.
Constant                             7.88          2.43       3.25    0.023
time Lin                           -0.157         0.230      -0.68    0.524
time Quad                         0.03336       0.00584       5.71    0.002
time Cub                       -0.0002772     0.0000436      -6.36    0.001
```



## Non-linear regression – standard curves

GenStat has a suite of non-linear standard models, including a logistic equation in the form

$$Y = A + \frac{C}{1 + e^{-B(t-M)}} .$$

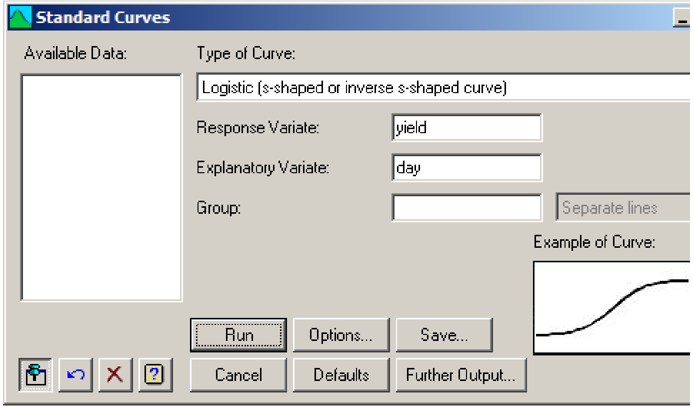
This equation is commonly used for the pasture growth of the previous example, usually with *A* set to 0. The problem with polynomial regression is the absence of biological interpretability of the regression coefficients. The logistic equation with *A*=0 has 3 parameters:

*M* = day that the pasture is growing fastest, having reached a yield of ½ *C*
*C* = the eventual maximum yield, and
*B* = *twice* the relative growth rate on the day the pasture is growing fastest.

| period | RGR |
|---|---|
| day 9 to 14 | 0.038 |
| day 14 to 21 | 0.078 |
| day 21 to 28 | 0.026 |
| day 28 to 42 | 0.040 |
| day 42 to 57 | 0.024 |
| day 57 to 63 | 0.016 |
| day 63 to 70 | 0.007 |
| day 70 to 79 | 0.004 |

Relative growth rate (RGR) is best estimated as *change in log(yield)* divided by *change in time*. It would appear that the *M* ≈ day 35. The RGR then is 0.04, so we would expect *B*≈0.08. Yield appears to be flattening out at *C*≈70. These are fairly good initial estimates. Choose Stats > Regression Analysis > Standard Curves. Select Logistic and turn off the option Estimate Constant Term (which here means setting A to be 0):

```
Nonlinear regression analysis

          Response variate: yield
                Explanatory: day
               Fitted Curve: A + C/(1 + EXP(-B*(X - M)))
                Constraints: A = 0.0


Summary of analysis

Source            d.f.          s.s.         m.s.
Regression           3     18215.364     6071.788
Residual             6         8.057        1.343
Total                9     18223.420     2024.824

Percentage variance accounted for 99.8
Standard error of observations is estimated to be 1.16.


Estimates of parameters

Parameter           estimate          s.e.
B                    0.06736       0.00345
M                      38.87          1.18
C                      72.46          1.73
```


Standard Curves

Available Data:

Type of Curve: Logistic (s-shaped or inverse s-shaped curve)

Response Variate: yield

Explanatory Variate: day

Group:

```
Nonlinear regression analysis

          Response variate:  yield
                Explanatory:  day
               Fitted Curve:  A + C/(1 + EXP(-B*(X - M)))
                Constraints:  A = 0.0
```

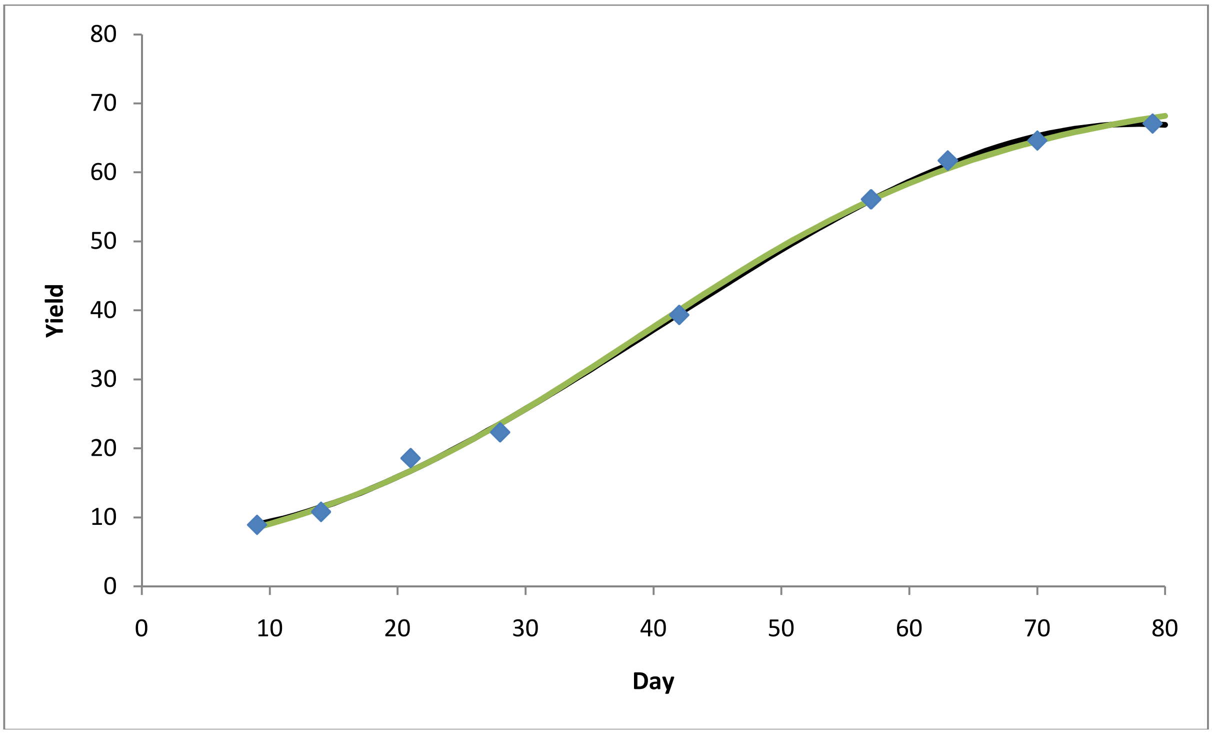
## Summary of analysis

```
Source              d.f.          s.s.          m.s.       v.r.     F pr.
Regression             3     18215.364      6071.788    4521.89     <.001
Residual               6         8.057         1.343
Total                  9     18223.420      2024.824

Percentage variance accounted for 99.8
Standard error of observations is estimated to be 1.16.
```

## Estimates of parameters

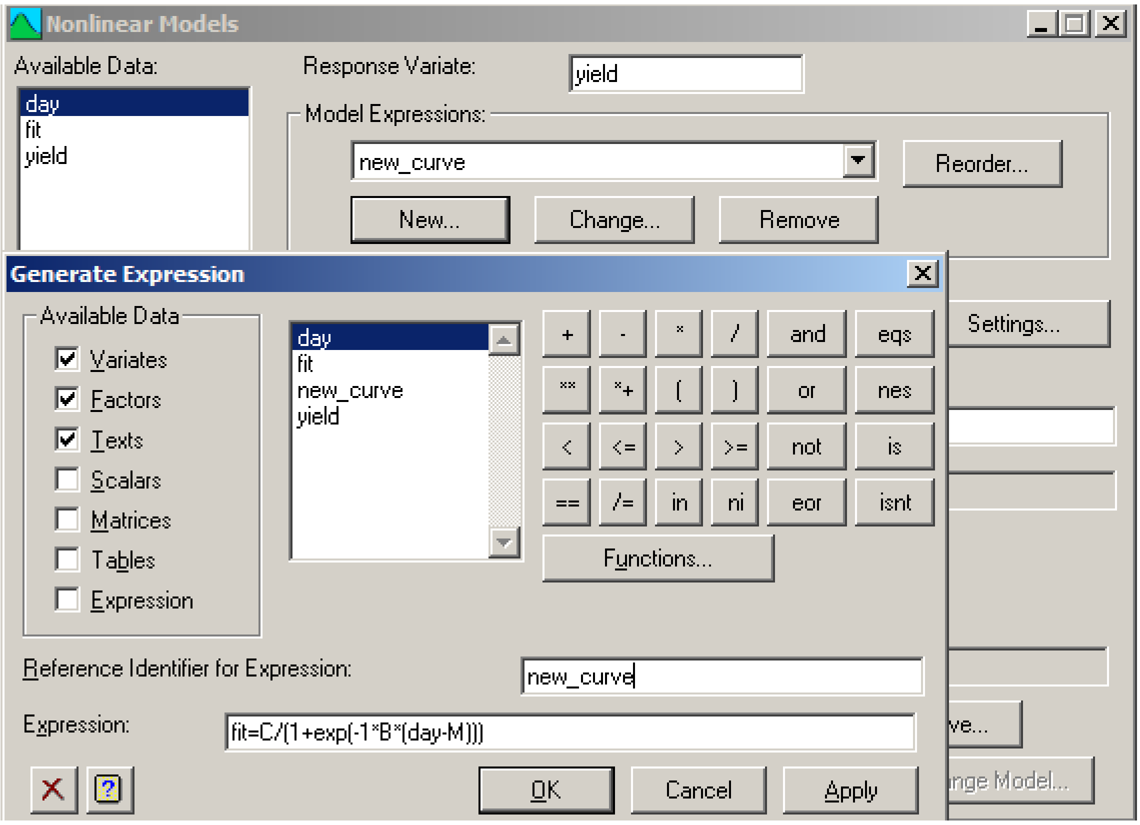
```
Parameter              estimate          s.e.
B                       0.06736       0.00345
M                         38.87          1.18
C                         72.46          1.73
```

There is very little difference visually between the two fitted cubic and logistic curves. Here the logistic (in green) is superimposed on the cubic (in black). The superiority of the logistic is in the ability to attach biological interpretation on the parameters.



## Non-linear regression – user-built functions

GenStat also allows a user to fit their own functions by choosing Stats > Regression Analysis > Nonlinear Models. We will use the previous logistic example to illustrate the method. You need to create a New model. In the sub-menu use any name for your new model (we chose new_curve) and then type the required equation, together with assigning a variate value for each X value (we chose fit).



Nonlinear Models

Available Data: day, fit, yield

Response Variate: yield

Model Expressions: new_curve

Generate Expression

Available Data

☑ Variates
☑ Factors
☑ Texts
☐ Scalars
☐ Matrices
☐ Tables
☐ Expression

day
fit
new_curve
yield

Reference Identifier for Expression: new_curve

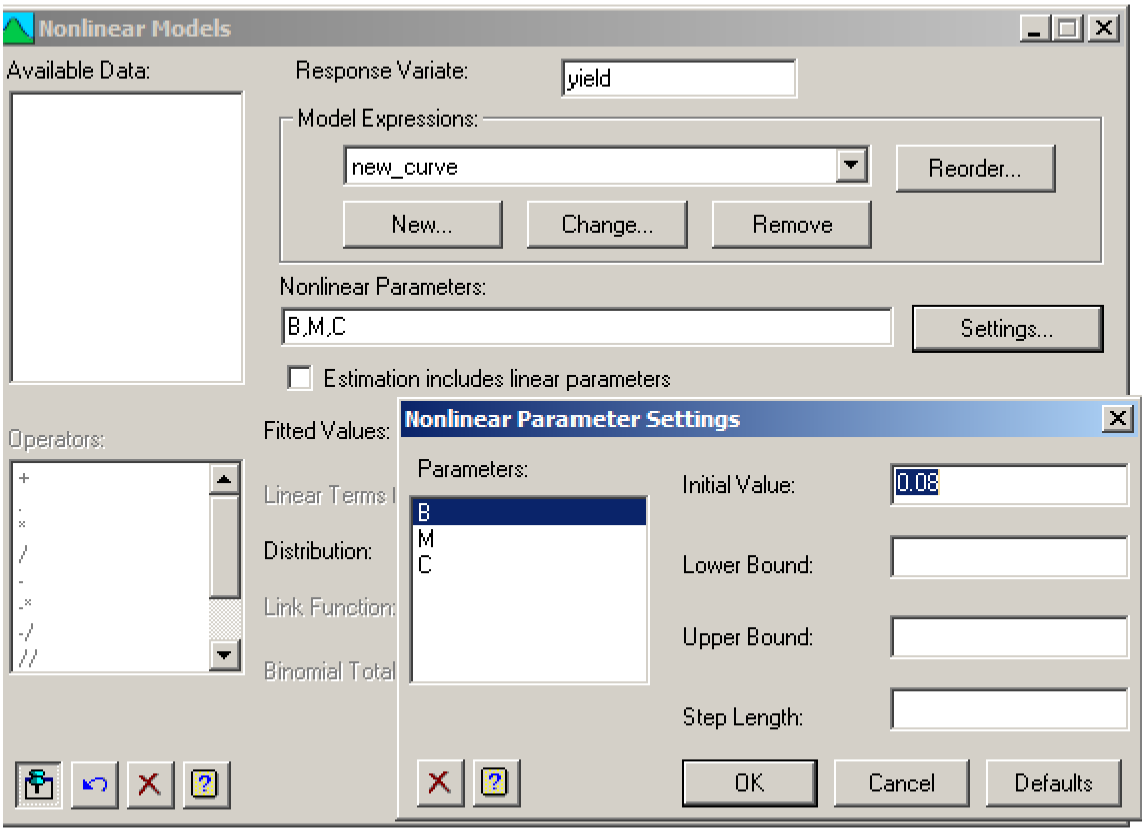
Expression: fit=C/(1+exp(-1*B*(day-M)))

Next, we indicate the list of non-linear parameters in the model (here all three, B, M and C). Iteration may be problematic for some complex models, so it's best to provide sensible and close initial values for these parameters. We'll use our earlier guesses from the biology of the pasture: B≈0.08, M≈35 and C≈70.

Note that this model has no overall mean or any linear part of the model, so we unclick Estimation includes linear parameters.

Finally, we need to provide the variate we used on the LHS of the user-defined expression (remember, we used fit).

Then we run the model. Iteration succeeds since we were close with our initial estimates. The model is the same as obtained by the standard curves menu.



Response Variate: yield

Model Expressions: new_curve

Nonlinear Parameters: B,M,C

□ Estimation includes linear parameters

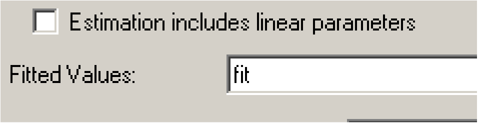
Parameters: B M C

Initial Value: 0.08

Lower Bound:

Upper Bound:

Step Length:



□ Estimation includes linear parameters

Fitted Values: fit

## Nonlinear regression analysis

```
        Response variate: yield
    Nonlinear parameters: B, M, C
      Model calculations: new_curve
```

## Summary of analysis

```
Source          d.f.        s.s.        m.s.      v.r.   F pr.
Regression         3   18215.364    6071.788   4521.89   <.001
Residual           6       8.057       1.343
Total              9   18223.420    2024.824

Percentage variance accounted for 99.8
Standard error of observations is estimated to be 1.16.
```

## Estimates of parameters

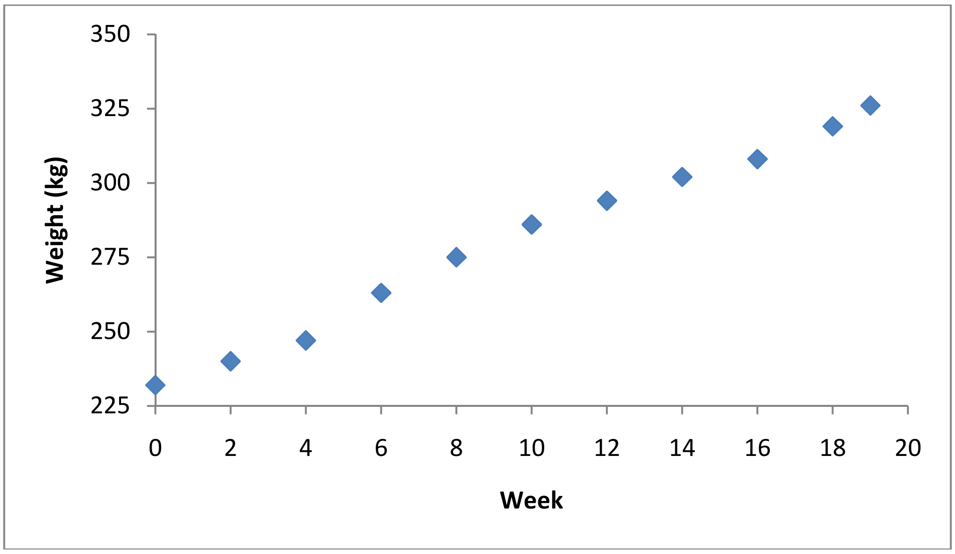
```
Parameter           estimate        s.e.
B                    0.06736     0.00345
M                      38.87        1.18
C                      72.46        1.73
```

## Regression with correlated errors

A typical situation where the errors in regression are correlated is when repeated measurements are made on a single experimental unit.

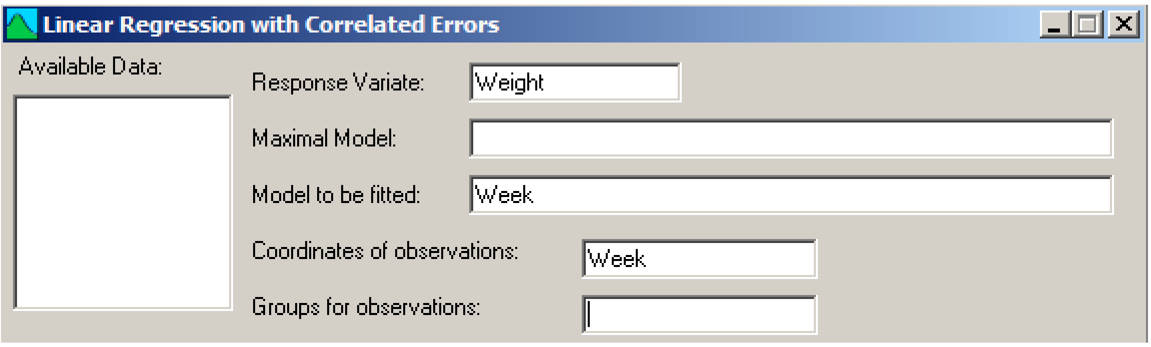
Example 6  Consider the weight of a single animal measured over 19 weeks from birth. A plot of its weight against time suggests a linear trend, although there is some suggestion of a slight slowdown from about week 10 with a spurt at week 16.

| Week | 0 | 2 | 4 | 6 | 8 | 10 | 12 | 14 | 16 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Weight | 232 | 240 | 247 | 263 | 275 | 286 | 294 | 302 | 308 | 319 | 326 |



## Linear regression with correlated errors

GenStat offers a procedure to fit an AR1/power model to the errors of a simple linear regression. Use Stats > Repeated Measurements > Linear Regression with correlated errors. In this case, the time points are the same as the X values in the regression. As an option you can select a ML or a REML algorithm for fitting the correlation between weights one unit of time apart.



Linear Regression with Correlated Errors

Available Data:

Response Variate: Weight

Maximal Model:

Model to be fitted: Week

Coordinates of observations: Week

Groups for observations:

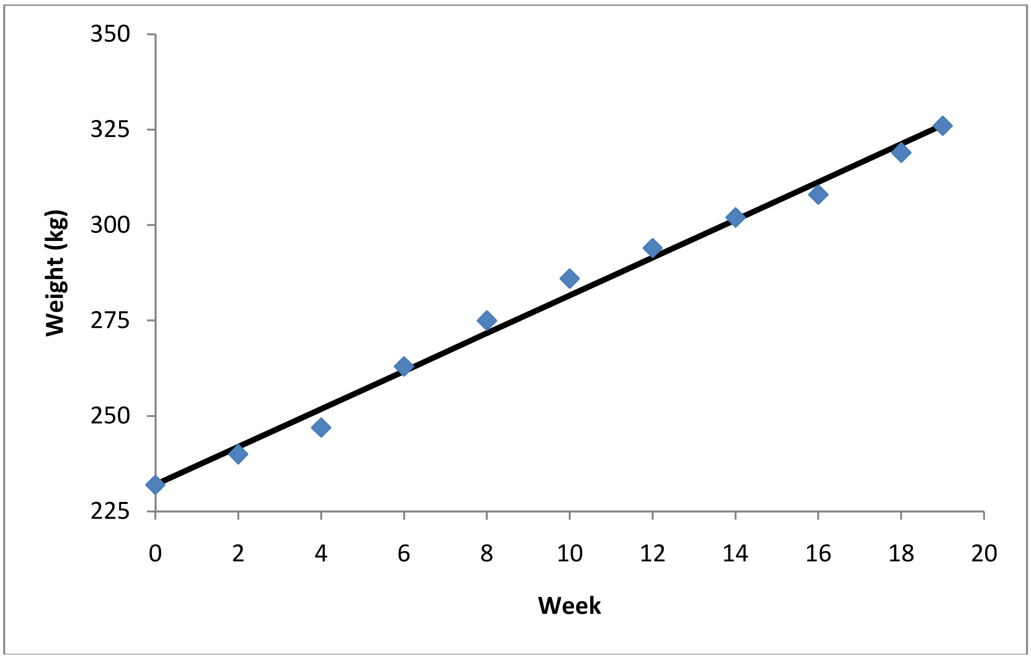
```
Regression analysis

        Response variate: Weight
            Weight matrix: _wgtmat based on power-distance correlation model
              Fitted terms: Constant, Week


Summary of analysis

Source              d.f.          s.s.          m.s.        v.r.    F pr.
Regression             1        4889.6       4889.64      257.20    <.001
Residual               9         171.1         19.01
Total                 10        5060.7        506.07

Percentage variance accounted for 96.2
Standard error of observations is estimated to be 4.36.

Message: the following units have large standardized residuals.
             Unit        Response        Residual
                3          247.00           -2.16

Message: the residuals do not appear to be random; for example, fitted values in the
range 261.73 to 301.37 are consistently smaller than observed values and fitted
values in the range 232.01 to 251.83 are consistently larger than observed values.

Estimates of parameters

Parameter            estimate             s.e.        t(9)     t pr.
Constant               232.01             4.12       56.26     <.001
Week                    4.954            0.309       16.04     <.001

Correlation parameter estimate

Phi: 0.8746
Test for phi non-zero: chi-square 5.073 on 1 d.f., probability 0.024
```

The line of best fit is

Weight = 232 + 4.954×Week

and there is a correlation of 0.8746 between the animal's weight a week apart, and this is significantly different to 0 (P=0.024).

However, the message suggests a systematic trend in the residuals. The actual weights in the middle time points are all above the fitted line, as shown in the following plot. The significant correlation in the errors may well be a result of a poorly chosen model.

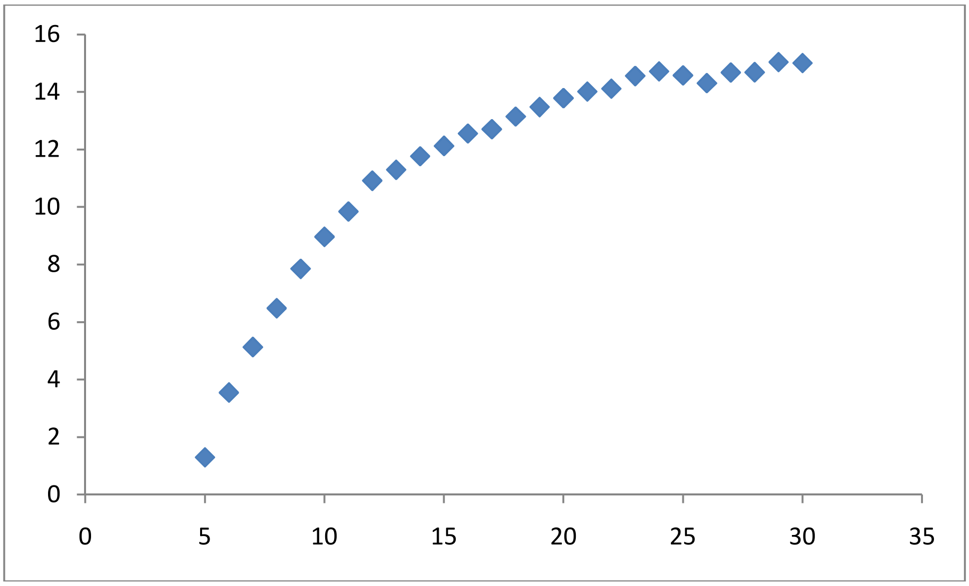


## Standard curves with correlated errors

GenStat also offers a procedure to fit an AR1/power model to the errors of a range of standard curves. Use Stats > Repeated Measurements > Standard curves with correlated errors.

Example 7. Roger Payne has an example in the *Statistics* manual in Help > GenStat Guides (page 1099).

| X | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | 1.3 | 3.55 | 5.13 | 6.48 | 7.85 | 8.96 | 9.84 | 10.91 | 11.29 | 11.76 | 12.12 | 12.55 | 12.7 | 13.14 | 13.47 | 13.78 | 14.01 | 14.11 | 14.55 | 14.71 | 14.57 | 14.3 | 14.67 | 14.68 | 15.03 | 15 |

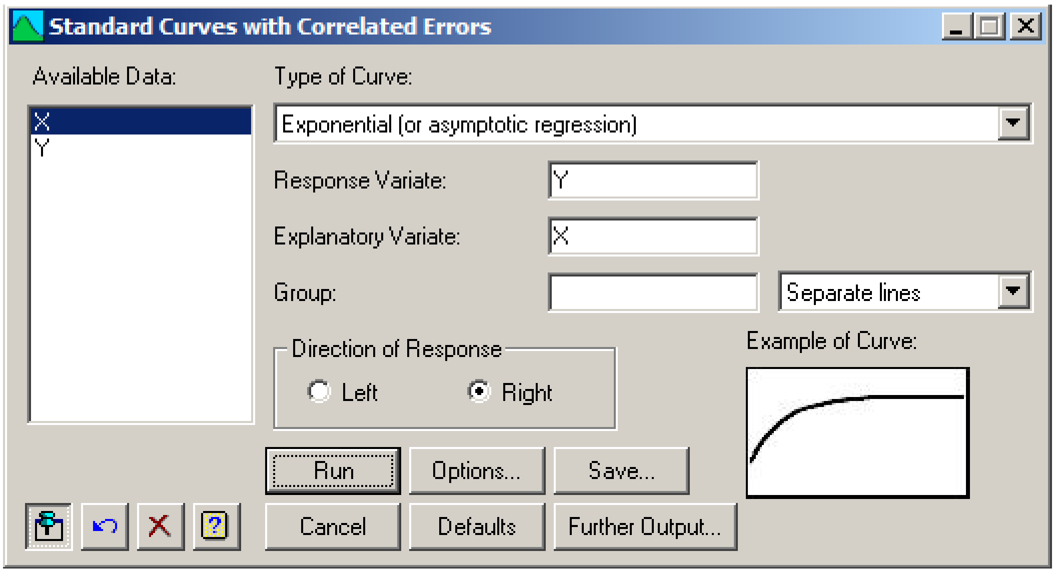


This type of response is an asymptotic regression, or an exponential with an asymptote to the right (*right sensing* in GenStat's terminology). The equation is

$$Y = A + B\,R^{X}$$

The equation that would fit the data would need $B$ to be negative and $0<R<1$. Then $A$ would be the (final) value of $Y$ as $X$ became large (so, from the plot of the data, about 15). With a value $X$=0 in the data set, $B$ would be the difference in $Y$ at X=0 and the asymptotic value of Y (so the amount the unit will eventually increase by over time from time 0). There is not such value in the example, and it is not easy to project the curve backwards to obtain a close idea of the value of $B$ as the plot crosses the $Y$ axis. The parameter $R$ governs the speed of the slowdown in $Y$.

Again, GenStat offers a procedure to fit an AR1/power model to the errors of standard curves where that appears appropriate. Use Stats > Repeated Measurements > Standard Curves with correlated errors.



Standard Curves with Correlated Errors

Available Data: X, Y

Type of Curve: Exponential (or asymptotic regression)

Response Variate: Y

Explanatory Variate: X

Group: Separate lines

Direction of Response: ○ Left ● Right

### Nonlinear regression analysis

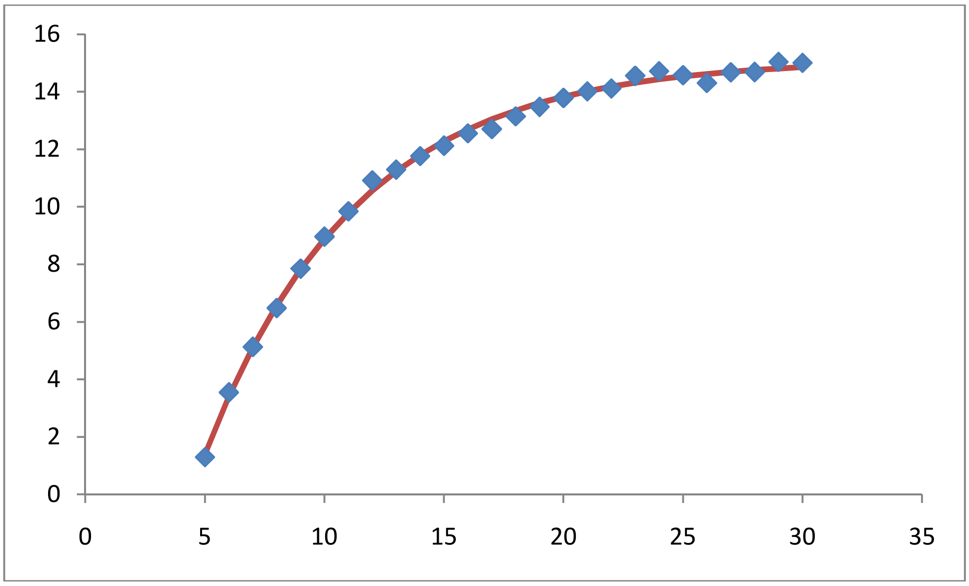
```
       Response variate: Y
          Weight matrix: _wgtmat based on power-distance correlation model
            Explanatory: X
           Fitted Curve: A + B*(R**X)
            Constraints: R < 1
```

### Summary of analysis

```
Source            d.f.         s.s.         m.s.      v.r.   F pr.
Regression           2     196.4891     98.24454   3132.82   <.001
Residual            23       0.7213      0.03136
Total               25     197.2104      7.88841

Percentage variance accounted for 99.6
Standard error of observations is estimated to be 0.177.
```

```
Message: the following units have large standardized residuals.
                Unit        Response        Residual
                   8          10.910            2.46
                  13          12.700           -2.37
                  22          14.300           -2.22

Message: the residuals do not appear to be random; for example, fitted values in the
range 11.793 to 14.177 are consistently larger than observed values and fitted
values in the range 7.808 to 11.226 are consistently smaller than observed values.

Message: the following units have high leverage.
                Unit        Response        Leverage
                   1           1.300            0.50
                   2           3.550            0.28

Estimates of parameters

Parameter                   estimate            s.e.
R                            0.85432         0.00282
B                            -30.166           0.581
A                            15.1216          0.0732

Correlation parameter estimate

Phi: 0.4008
Test for phi non-zero: chi-square 4.313 on 1 d.f., probability 0.038
```

The correlation of 0.4 is significantly different to 0 (P=0.038). The plot settles down at A=15.12 units, and would have increased from a base of -30 at “X=0”. The plot of this model is as follows; 99.6% of the variation in *Y* is explained by the model.



## Section 3 - Analysis of non-normal data

### Link functions

With normally distributed data, the distribution involves a mean parameter $\mu$ and a variance parameter $\sigma^2$. As we have seen in the previous sections, the model for data from one population can be expressed as

$$Y = \mu + Error$$

where *Error* is $N(0, \sigma^2)$. We used maximum likelihood or residual maximum likelihood to estimate $\mu$ and $\sigma^2$.

For non-normal data, it is generally not possible to impose an additive model such as this. For example, if $Y$ is binomial with known $n$ and unknown $p$, we can write down a likelihood expression and maximise it to estimate $p$. If $Y$ is Poisson with unknown mean $\mu$, we can write down a likelihood expression and maximise it to estimate $\mu$.

When we come to many treatments involving binomial or Poisson data, we need to ensure that the maximum likelihood estimates are properly defined, in particular the probability of a success in each treatment must fall in the region (0,1), while for Poisson data each mean must be positive.

Finney was among the first to suggest a method for analysing binomial data for designed experiments involving herbicides, insecticides and so on. The method became known as probit analysis. More often these days, scientists in this area will use logistic regression because it analyses log-odds as we will see.

The modern method of analysing non-normal data is by maximum likelihood, in which the mean is modelled on a scale guaranteed to produce well defined estimates.

For Poisson data, we generally assume that

$$E(Y) = \mu = e^{b_0 + b_1 X_1 + \ldots}$$

where $X_1$, …could be covariates to explain the change in the Poisson mean, or design features (treatment effects and so on). Thus,

$$\log_e(\mu) = b_0 + b_1 X_1 + \ldots$$

We call this a *linear predictor with a log-link*. Estimation of the parameters in the linear function is by ML.

For binomial data, Finney noticed that the percentage of insects dying at low doses was small, increased rapidly as the dose increased and obviously asymptoted to 100% kill with sufficiently high dose. He noted that such a shape is typical of the cumulative distribution function of a normal variable, and proposed that method to estimate the parameters of the binomial. As mentioned these days the logistic distribution is more usual. We allow the probability of a success to depend on linear predictors via the logistic

$$p = \frac{1}{1+e^{-(b_0+b_1X_1+\cdots)}}.$$

This can be transformed as follows. Note that

$$\frac{p}{1-p} = e^{(b_0+b_1X_1+\cdots)}.$$

This ratio, *p*/(1-*p*), is known as the *odds*. If you toss a fair coin you have a 50:50 chance of a head, or an odds of 1. If seeds have about an 80% germination rate, the odds are 0.8:0.2, or 4:1 - an odds of 4.

Taking logs now gives

$$log_e(odds) = log_e\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + \cdots$$

Thus, the link for the binomial is known as the logit link.

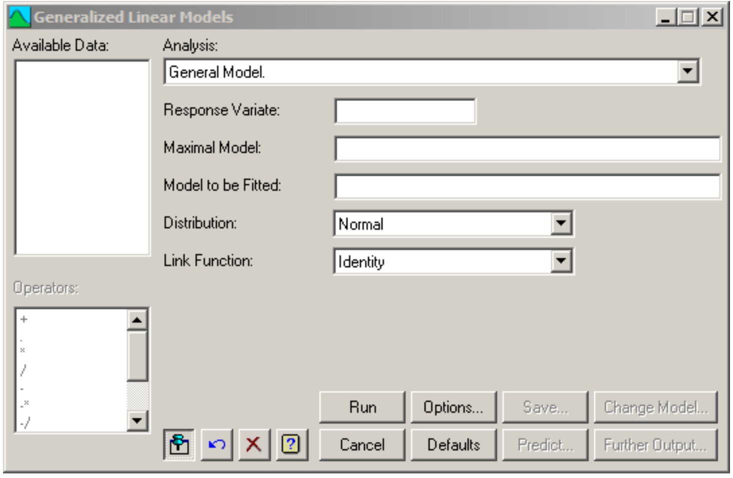
Once you estimate the parameters of this linear predictor, you calculate the odds, then the estimate of the probability:

$$probability = \frac{odds}{1+odds}.$$

To summarise, for non-normal data, for each distribution we have a different linear predictor and link function, use maximum likelihood to estimate the parameters of the linear predictor and use change in deviance to compare models.

**Background to generalized linear models**

This type of regression model is called a *generalized linear model* (GLM). From Version 12 of GenStat you can select a particular menu for your data via Stats > Regression Analysis (choose Logistic Regression for binomial data, Log-Linear Models for Poisson data, Probit Analysis for binomial data, etc), or you can select the Generalized Linear Models general menu and make your selection there:



Analysis: General Model.

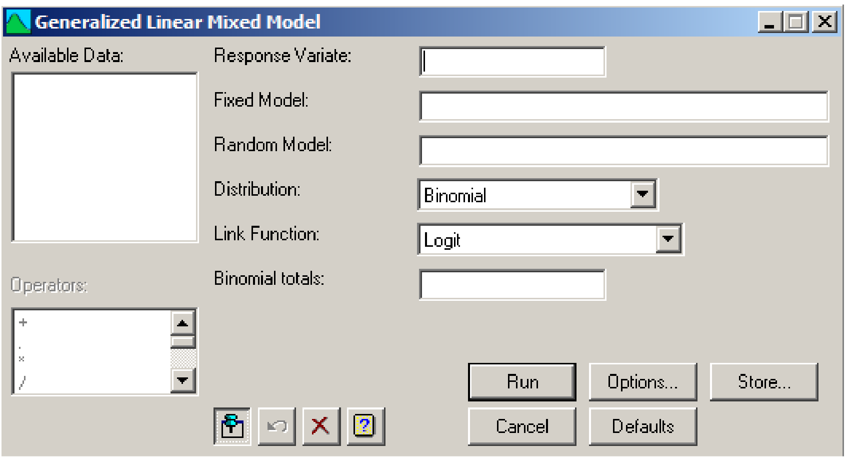
Distribution: Normal

Link Function: Identity

## Background to generalised linear mixed models

In the **Model to be Fitted** of a GLM you should really only enter fixed effects. These correspond to the set of predictors $X_1$, $X_2$, … in the linear predictor. If you do have an experiment involving random effects, then the mathematics is more complex and a different menu is available. We are now dealing with *mixed* models again (fixed and random effects) and a menu for a **Generalized Linear Mixed Model** (GLMM) is available. Choose this type of analysis via Stats > Regression Analysis > Mixed Models or from the dedicated Mixed Models menu.

*The random effects are assumed to be normally distributed for GLMMs.*



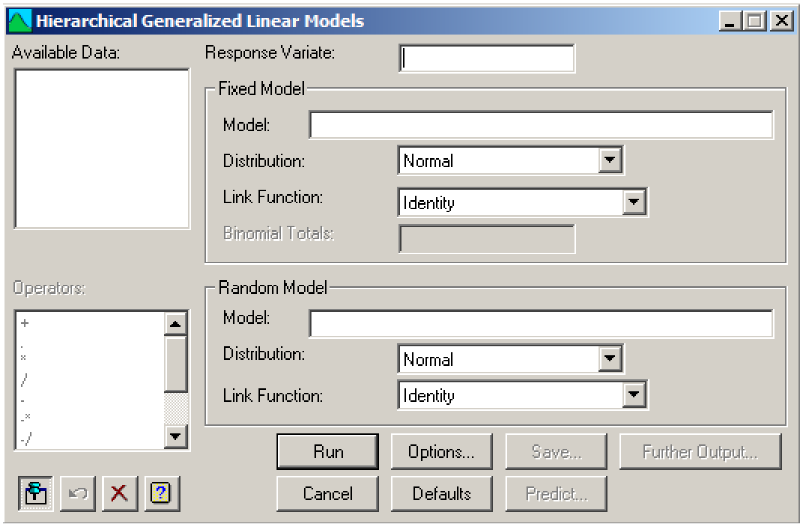
Generalized Linear Mixed Model

Distribution: Binomial

Link Function: Logit

## Background to hierarchical generalised linear mixed models

If you believe that the random effects have a non-normal distribution then the analysis is again more complex. The model is known as hierarchical generalised linear mixed model (HGLMM). Again, choose it from Stats > Regression Analysis > Mixed Models or from the dedicated Mixed Models menu. The HGLMM menu contains a selection of distributions to choose from for the random effects.



Hierarchical Generalized Linear Models

Fixed Model

Distribution: Normal

Link Function: Identity

Random Model

Distribution: Normal

Link Function: Identity

We include one example of GLMMs in this manual, but leave the more complex HGLMMs for another occasion.

## Binary logistic regression

Firstly, let us take the 2×2 contingency table, where the rows represent different treatments. (There are other types of contingency tables, some of which we consider later in this section.)

Example 8. Incidence of rust in Kentucky bluegrass pastures, from Steel and Torrie, page 504

| Pasture field type | Rust | No Rust | Total |
|---|---|---|---|
| 1 | 372 | 24 | 396 |
| 5 | 330 | 48 | 378 |

Readers may be familiar with Pearson's $\chi^2$ goodness of fit statistic used to test whether the probability of rust is the same for the two pasture types.

$$X^2=\sum\frac{(Observed\text{-}Expected)^2}{Expected}$$ asymptotically $\chi^2$ with $d$ = (# *rows*-1)(# *columns*-1).

Under a hypothesis that the probability of rust in field type 1 is the same as that for field type 5, the best (ie ML) estimate of "Rust" is $p$=(372+330)/(396+378) = 0.907. This allows us to work out how many rust-affected clonal isolations are expected for each pasture type. (For example, for pasture type 1, we expect 0.907×396=359.2 to be affected.)

This test is available in GenStat. However, a more common test is now used, the maximum likelihood $\chi^2$ test. It is, in fact, the same as the deviance in a binary logistic analysis of these data.

$$X^2=2\sum Observed\times ln\left(\frac{Observed}{Expected}\right)$$ asymptotically $\chi^2$ with (*rows*-1)(*columns*-1) *df.*

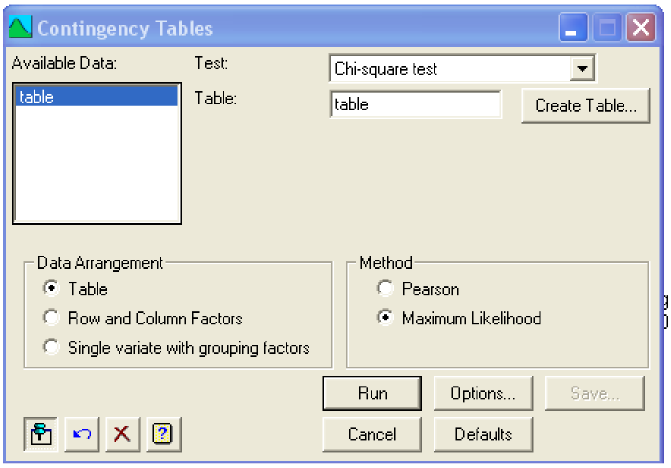
## Analysis as a contingency table

Choose Stats > Statistical Tests > Contingency Tables. If you have not already done so, click Create Table and choose Spreadsheet. Then enter or copy the data to the table, and click back to the menu. Choose the Method (Pearson or Maximum Likelihood) and, in Options, if you wish to see expectations or not.

For these data the test values are virtually identical:

```
Chi-square test for association between C2 and C3

Pearson chi-square value is 10.10 with 1 d.f.
Probability level (under null hypothesis) p = 0.001
```

```
Chi-square test for association between C2 and C3

Likelihood chi-square value is 10.25 with 1 d.f.
Probability level (under null hypothesis) p = 0.001
```



Contingency Tables

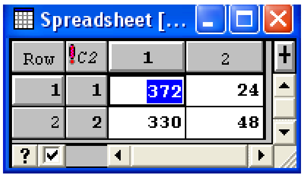
Available Data: table

Test: Chi-square test

Table: table

Data Arrangement: ● Table ○ Row and Column Factors ○ Single variate with grouping factors

Method: ○ Pearson ● Maximum Likelihood



Spreadsheet [...

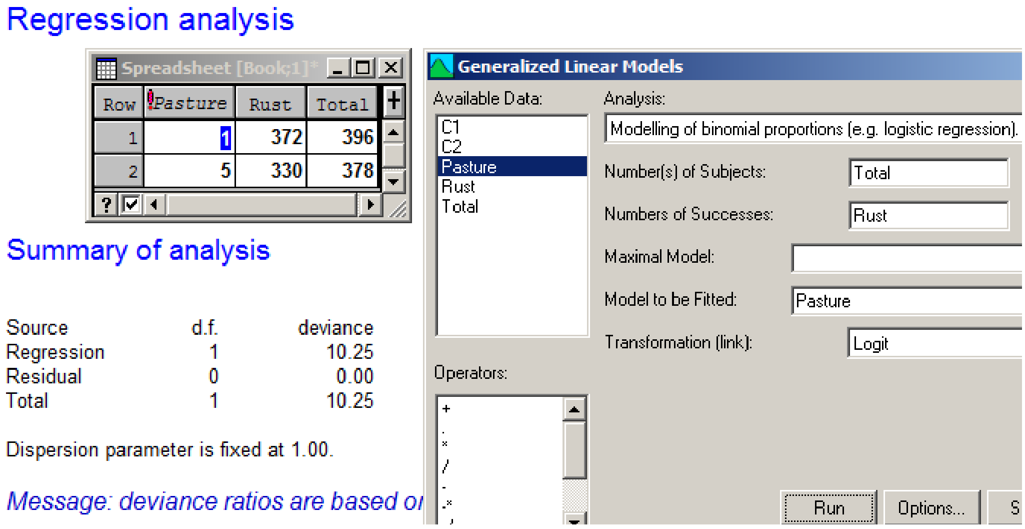
| Row | C2 | 1 | 2 |
|---|---|---|---|
| 1 | 1 | 372 | 24 |
| 2 | 2 | 330 | 48 |

Clearly, there is strong evidence (P=0.001) that individual probability estimates are required for the two pasture types. We would use

Pasture type 1: $p$ = 372/396 = 0.939
Pasture type 2: $p$ = 330/378 = 0.873

### Analysis via logistic regression

Now let us do this in GenStat's Regression > Logistic Regression menu. You need a factor column to identify the two pasture types, a variate of rust numbers and a variate of totals.



Regression analysis

Spreadsheet [Book;1]*

| Row | Pasture | Rust | Total |
|---|---|---|---|
| 1 | 1 | 372 | 396 |
| 2 | 5 | 330 | 378 |

Summary of analysis

```
Source          d.f.     deviance
Regression        1         10.25
Residual          0          0.00
Total             1         10.25

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based o
```

Generalized Linear Models

Available Data: C1, C2, Pasture, Rust, Total

Analysis: Modelling of binomial proportions (e.g. logistic regression).

Number(s) of Subjects: Total

Numbers of Successes: Rust

Maximal Model:

Model to be Fitted: Pasture

Transformation (link): Logit

```
Regression analysis

    Response variate:  Rust
      Binomial totals:  Total
         Distribution:  Binomial
        Link function:  Logit
         Fitted terms:  Constant, Pasture
```

```
Summary of analysis
                                                 mean   deviance  approx
Source            d.f.      deviance         deviance      ratio  chi pr
Regression           1         10.25            10.25      10.25   0.001
Residual             0          0.00                *
Total                1         10.25            10.25

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

Estimates of parameters
                                                                antilog of
Parameter          estimate        s.e.       t(*)    t pr.      estimate
Constant              2.741       0.211      13.01    <.001         15.50
Pasture 5            -0.813       0.261      -3.11    0.002        0.4435

Message: s.e.s are based on dispersion parameter with value 1.

Parameters for factors are differences compared with the reference level:
                    Factor   Reference level
                   Pasture   1
```

This is very similar to a regression output with a factor. Notice:

- The Regression Deviance is 10.25, identical to the ML contingency table $X^2$ statistic.

- The linear predictor on the logit scale is (2.741 – 0.813 $X$), where $X$ takes values 0 (for pasture type 1) and 1 (for pasture type 5). That is, the constant identifies the model for pasture type 1. The model for pasture type 5 is obtained by adding 2.741 and -0.813, obtaining 1.928.

  The *odds* are $e^{2.741}$ = 15.50 (which GenStat produces as the antilog of estimate) for pasture type 1, and $e^{2.741-0.813}$ = 6.876 for pasture type 5. The latter is also available by *multiplying* the default odds, 15.50, by $e^{-0.813}$ (=0.4435, shown as the antilog of the estimate for pasture type 5). Thus, the odds for pasture type 5 are 6.874.

  ***Summary***
  The antilog for the default level of a factor is its odds; the antilogs for the other levels of that factor are *odds ratios*. You multiply the *default odds* by the *odds ratio* to obtain the odds for the other levels of a factor.

  Once the odds are available, the estimated probabilities can be calculated as odds/(1+odds). We obtain 15.50/(1+15.50) = 0.939 for pasture type 1, and 6.876/(1+6.876) = 0.873 for pasture type 5. *These are what we calculated following the contingency table test.*

- If you Save the fitted values, in this case you obtain the actual data. *Thus, if a factor is significant in a model, the fitted values are identical to the actual* ***totals*** *for that factor.*

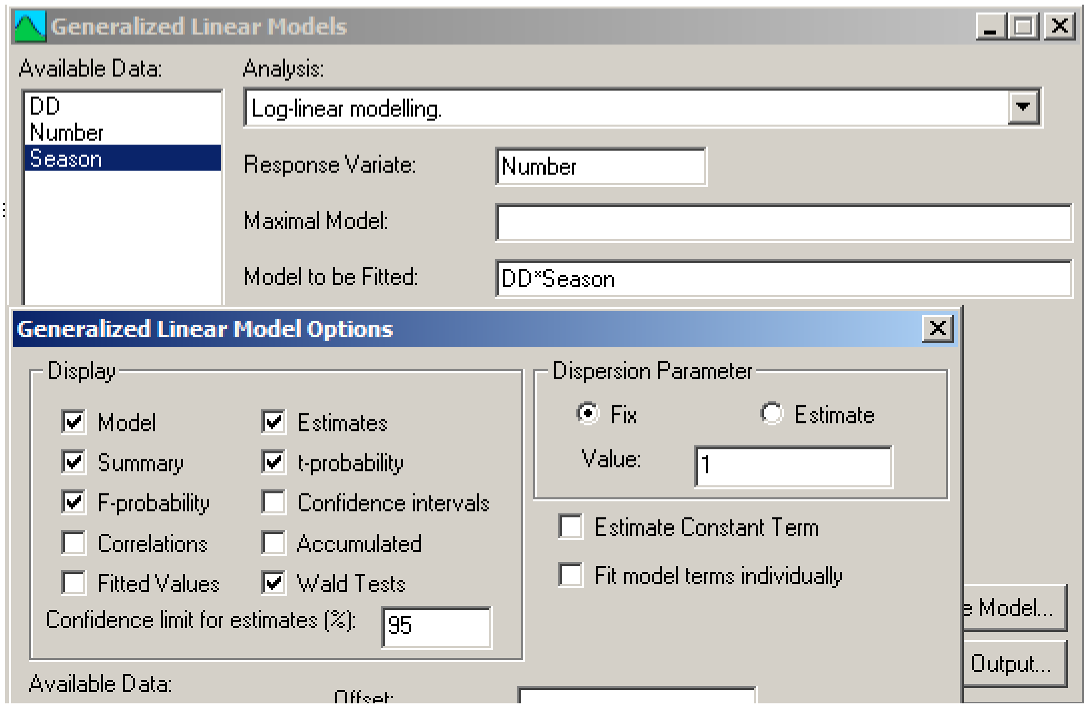
## Poisson Regression

In the parallel regressions section we analysed mean leaf numbers from 10 cauliflower plants over two years using linear regression which assumes normality. If the number of cauliflower leaves is Poisson, the adding the number of leaves from 10 plants is also Poisson. Multiplying by 10 gives the following total leaf numbers. It is true that a Poisson distribution tends to a normal distribution, but we can use the exact distribution for numbers like these, and use a Poisson regression analysis.

| 1956/7 season | | 1957/8 season | |
|---|---|---|---|
| DD | Number | DD | Number |
| 45 | 38 | 45 | 60 |
| 75 | 62 | 80 | 85 |
| 95 | 72 | 95 | 91 |
| 105 | 87 | 115 | 120 |
| 130 | 102 | 130 | 126 |
| 160 | 135 | 140 | 133 |
| 180 | 150 | 165 | 152 |

We need to stack these data, creating a factor column to identify year, a column of total leaf numbers and a column of day degrees (DD), the predictor.

We are interested in whether a single model fits the data, or parallel or separate models over the two years. The Model to be Fitted for separate lines is therefore DD*Years.



Generalized Linear Models

Available Data: DD, Number, Season

Analysis: Log-linear modelling.

Response Variate: Number

Maximal Model:

Model to be Fitted: DD*Season

Generalized Linear Model Options

Display

☑ Model ☑ Estimates
☑ Summary ☑ t-probability
☑ F-probability □ Confidence intervals
□ Correlations □ Accumulated
□ Fitted Values ☑ Wald Tests

Confidence limit for estimates (%): 95

Dispersion Parameter

● Fix ○ Estimate

Value: 1

□ Estimate Constant Term

□ Fit model terms individually

Available Data:

Offset:

Notice that Wald statistics are now available for GLMs. These allow us to test whether factors can be dropped as they are entered last in the analysis. Prior to their introduction, we would select Accumulated and Fit model terms individually, but then have to use all orders for analyzing data such as these,

```
Regression analysis

            Response variate: Number
                Distribution: Poisson
               Link function: Log
                Fitted terms: Constant + DD + Season + DD.Season

Summary of analysis
                                              mean    deviance   approx
Source          d.f.       deviance       deviance       ratio   chi pr
Regression         3        170.693        56.8976       56.90   <.001
Residual          10          4.596         0.4596
Total             13        175.289        13.4838
Change            -1         -1.463         1.4630        1.46   0.226

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

...
Estimates of parameters
                                                                          antilog of
Parameter                  estimate          s.e.       t(*)    t pr.      estimate
Constant                      3.398         0.127      26.71    <.001         29.91
DD                          0.09259       0.00928       9.98    <.001         1.097
Season 1957_8                 0.426         0.181       2.35    0.019         1.531
DD.Season 1957_8            -0.0168        0.0138      -1.21    0.226        0.9834

Message: s.e.s are based on dispersion parameter with value 1.

Parameters for factors are differences compared with the reference level:
                         Factor  Reference level
                         Season  1956_7

Wald tests for dropping terms

       Term        Wald statistic    d.f.      chi. pr.
DD.Season                   1.465       1         0.226
```

Notice the following.

- If we have Poisson data, the "dispersion parameter", which is the mean deviance for the residual term, that is, residual deviance/residual d.f., should be 1. In this case it is under-dispersed, with a of dispersion parameter 0.4596. Is this a problem? What we do is test whether the deviance of 4.596 is likely to have come by chance from a $\chi^2$ distribution with (in this case) a low 10 *df*. A lower critical probability is 0.0835. Before we do the experiment, there is no reason why the deviance will be greater than or less than what is expected, by chance. Hence the *P* value we would quote for this is 0.167. We would not reject a hypothesis that the data are Poisson.

DD is a variate and the term DD.Season basically is in the model to allow *different slopes* for different years. Naturally the P value for the Wald test of this effect is the same as the P value for the final coefficient in the regression. This parameter estimate is not significant (P=0.226), and hence parallel regressions are indicated (unless the response to day degrees is itself not significant). The term DD.Season can be dropped from the model. When this is done, Wald tests of DD and of Season are given, each adjusted for the other.

```
Wald tests for dropping terms

   Term          Wald statistic     d.f.       chi. pr.
     DD                  152.57        1         <0.001
 Season                   16.33        1         <0.001
```

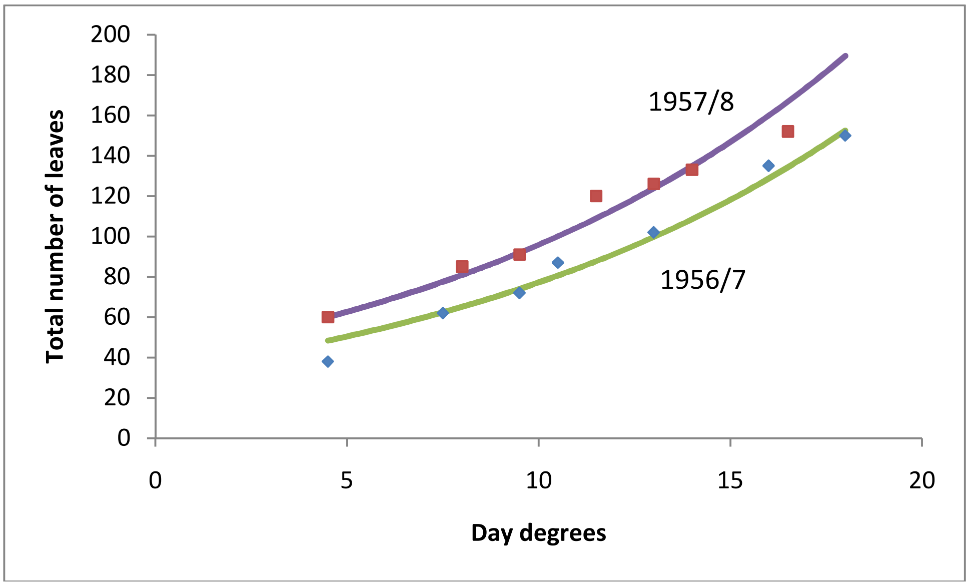
The parameter Season 1957_8 is in the model to allow a different *intercept* for the second year, and this is significant (P<0.001). Hence there is a shift in the mean number of leaves over the two seasons.

Similarly, DD is significant (P<0.001), indicating a strong linear relationship between leaf number and temperature as measured by day degrees.

In the present analysis, however, linearity is on the log-scale:

```
Estimates of parameters
                                                                     antilog of
Parameter                    estimate         s.e.        t(*)   t pr.    estimate
Constant                       3.4949       0.0972       35.95   <.001       32.95
DD                            0.08513      0.00689       12.35   <.001       1.089
Season 1957_8                  0.2169       0.0537        4.04   <.001       1.242
```

Thus, the model for 1956/7 is $32.95 \times 1.089^{DD}$ and for 1957/8 it is $1.242 \times 32.95 \times 1.089^{DD}$. When plotted on the data, the fitted line looks unacceptable. The original concept was a simple linear increase in leaf numbers with increasing temperatures, and the fitted model is decidedly exponential:



If it seems that negative means are not likely to be obtained in the estimation process (ass here), it is quite acceptable to choose a different link function. In this case, the Identity link simply instructs the algorithm to model on the leaf number scale.

Using the Identity link function results in a dispersion parameter of only 0.1752. This is significantly under-dispersed (P=0.004) for a Poisson distribution (remember, we expect 1). The usual step to take in this situation is to *estimate* the dispersion parameter rather than *fix* it at 1, which we do in the options menu. The effect is to change the underlying distribution of the Wald statistic from $\chi^2$ to F. We then find:

- Separate regressions are unnecessary (P= 0.495);
- Season and DD are both strongly significant (P<0.001).

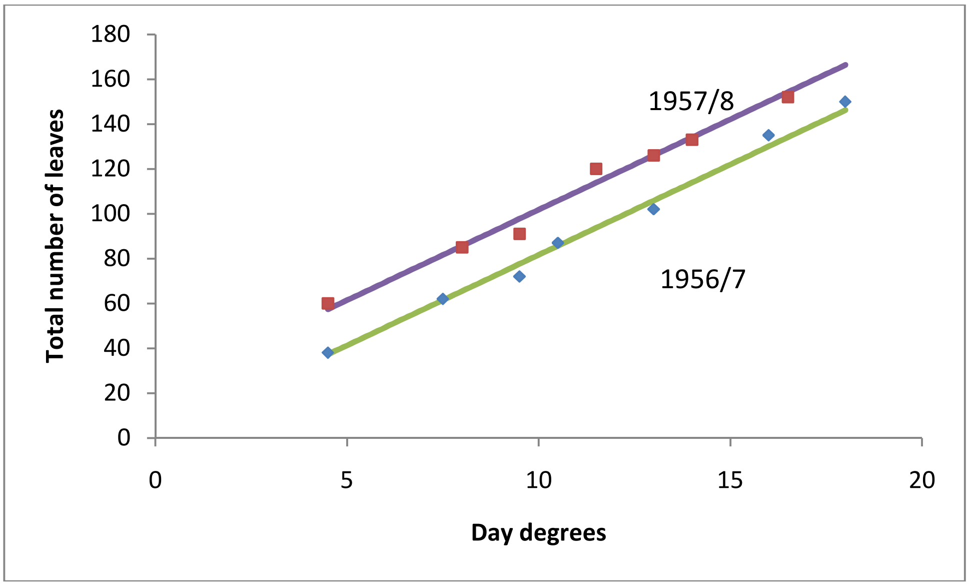
The new models for total leaf number are constructed from:

```
Estimates of parameters

Parameter                 estimate          s.e.      t(11)     t pr.
Constant                      0.96          2.68       0.36     0.727
DD                           8.068         0.247      32.63     <.001
Season 1957_8                20.15          2.06       9.80     <.001
```

Thus, we obtain

Total leaf numbers = 0.96 + 8.068×DD for 1956/7, and
Total leaf numbers = 21.11 + 8.068×DD for 1957/8.

These are very similar to those from the regression analysis, though for that analysis we used means rather than totals (and hence the current parameter values will be 10 times those from the regression).



## Log-linear modelling

This general analysis is used for more complex contingency tables. It turns out that binomial data can be treated as Poisson data *conditional on the totals being fixed*. Thus, provided we use terms in the model to fix the totals, we should obtain the same analysis using log-linear modelling as we do from logistic regression. Log-linear modelling, of course, is more general - one can have any numbers of outcomes, not just two (success/failure).

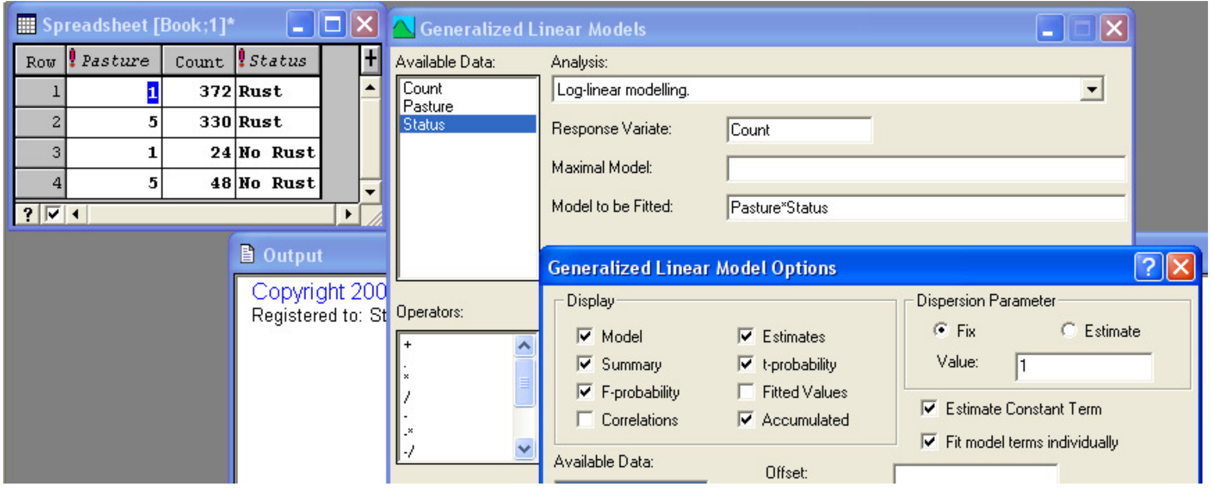
Firstly, consider the incidence of rust in Kentucky bluegrass pastures again. This time we stack the successes and failures, and provide a factor column to identify each. Thus, the information in the table will need to be prepared as shown:

| Pasture field type | Rust | No Rust |
|---|---|---|
| 1 | 372 | 24 |
| 5 | 330 | 48 |

| Pasture | Count | Status |
|---|---|---|
| 1 | 372 | Rust |
| 5 | 330 | Rust |
| 1 | 24 | No Rust |
| 5 | 48 | No Rust |

The **Model to be Fitted** is Pasture*Status.

We need to keep the pasture totals fixed, so Pasture must be present in the model simply to fix these. Status alone tests whether the counts are equal, and is of no interest. The only factor of interest in this experiment is the apparent interaction Pasture.Status. It assesses whether the Rust:No Rust ratio of counts is equal for the two pasture types.



Spreadsheet [Book;1]*

| Row | Pasture | Count | Status |
|---|---|---|---|
| 1 | 1 | 372 | Rust |
| 2 | 5 | 330 | Rust |
| 3 | 1 | 24 | No Rust |
| 4 | 5 | 48 | No Rust |

Output

Copyright 200

Registered to: St

Generalized Linear Models

Available Data: Count, Pasture, Status

Analysis: Log-linear modelling

Response Variate: Count

Maximal Model:

Model to be Fitted: Pasture*Status

Generalized Linear Model Options

Display

☑ Model ☑ Estimates

☑ Summary ☑ t-probability

☑ F-probability □ Fitted Values

□ Correlations ☑ Accumulated

Dispersion Parameter

● Fix ○ Estimate

Value: 1

☑ Estimate Constant Term

☑ Fit model terms individually

Available Data:

Offset:

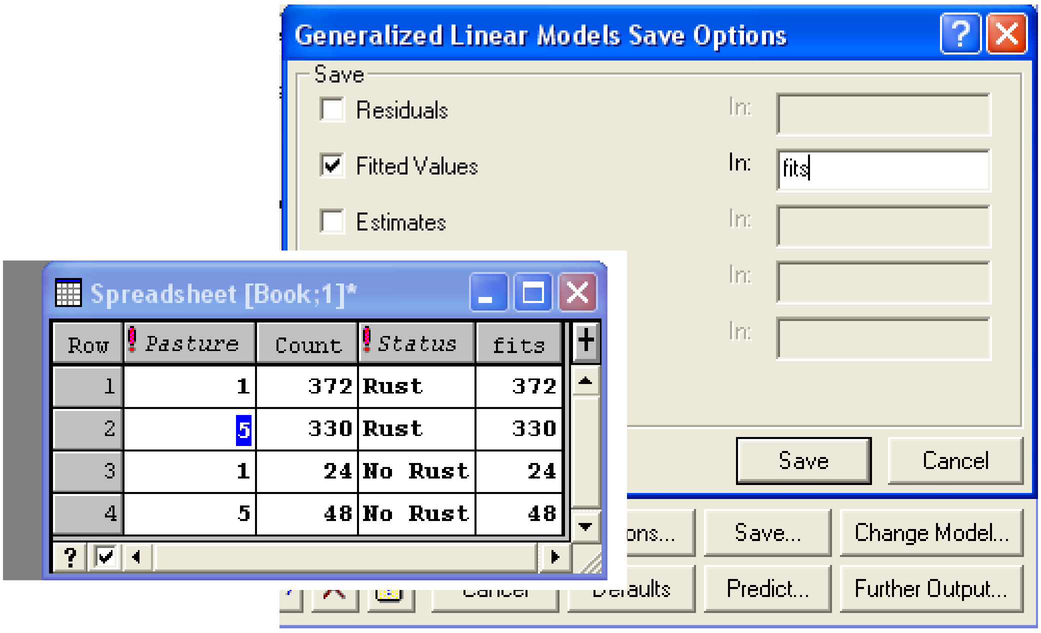
```
Regression analysis

        Response variate:  Count
            Distribution:  Poisson
           Link function:  Log
           Fitted terms:   Constant + Pasture + Status + Pasture.Status

Summary of analysis

                                              mean    deviance   approx
Source             d.f.     deviance      deviance       ratio   chi pr
Regression            3        604.6         201.5      201.53    <.001
Residual              0          0.0             *
Total                 3        604.6         201.5
Change               -1        -10.3          10.3       10.25    0.001

Dispersion parameter is fixed at 1.00.
```

```
Message: deviance ratios are based on dispersion parameter with value 1.

Estimates of parameters

                                                                                antilog of
Parameter                 estimate        s.e.        t(*)         t pr.       estimate
Constant                     3.178       0.204       15.57         <.001          24.00
Pasture 5                    0.693       0.250        2.77         0.006          2.000
Status Rust                  2.741       0.211       13.02         <.001          15.50
Pasture 5 .Status Rust      -0.813       0.261       -3.11         0.002         0.4435

Message: s.e.s are based on dispersion parameter with value 1.

Parameters for factors are differences compared with the reference level:
                   Factor   Reference level
                  Pasture   1
                   Status   No Rust

Accumulated analysis of deviance

                                                      mean      deviance        approx
Change                       d.f.     deviance    deviance         ratio        chi pr
+ Pasture                       1         0.42        0.42          0.42         0.518
+ Status                        1       593.92      593.92        593.92         <.001
Residual                        1        10.25       10.25
+ Pasture.Status                1        10.25       10.25         10.25         0.001
Total                           3       604.59      201.53
```

Notice:

- The Residual deviance is 0 and has 0 *df*. This is because there are only 4 cells in the table, hence 3 df*;* and we are modelling the data with 3 terms each of which has just 1 df. This is known as a *saturated model.*

  **The fitted values for a saturated model are just the original data values.**



Generalized Linear Models Save Options

Save

☐ Residuals In:

☑ Fitted Values In: fits

☐ Estimates In:

Spreadsheet [Book;1]*

| Row | Pasture | Count | Status | fits |
|---|---|---|---|---|
| 1 | 1 | 372 | Rust | 372 |
| 2 | 5 | 330 | Rust | 330 |
| 3 | 1 | 24 | No Rust | 24 |
| 4 | 5 | 48 | No Rust | 48 |

- The Pasture.Status component (10.25) of the ML $\chi^2$ is identical to that obtained using logistic regression.

```
Estimates of parameters

                                                                              antilog of
Parameter                 estimate            s.e.           t(*)          t pr.      estimate
Constant                     3.178           0.204          15.57          <.001         24.00
Pasture 5                    0.693           0.250           2.77          0.006         2.000
Status Rust                  2.741           0.211          13.02          <.001         15.50
Pasture 5 .Status Rust      -0.813           0.261          -3.11          0.002        0.4435
```

- The fitted model is referenced to pasture type 1, no rust. The saturated model gives an antilog of 24, namely the actual count for that combination. Pasture type 5 then has a fitted count of 2×24 = 48, again the actual count. The fitted count for pasture type 1, rust is 24×15.5 = 372, again the actual count.

  This is another illustration of the rule that the presence of any main effect or interaction in a generalized linear model induces the fitted counts to be identical to the observed counts for the table concerned.

- The Wald statistic in this analysis is very slightly different to that of the ML $\chi^2$ statistic for the interaction. It is based on a different approach, just as the Pearson $\chi^2$ statistic is slightly different to the ML $\chi^2$ statistic. Choose either test.

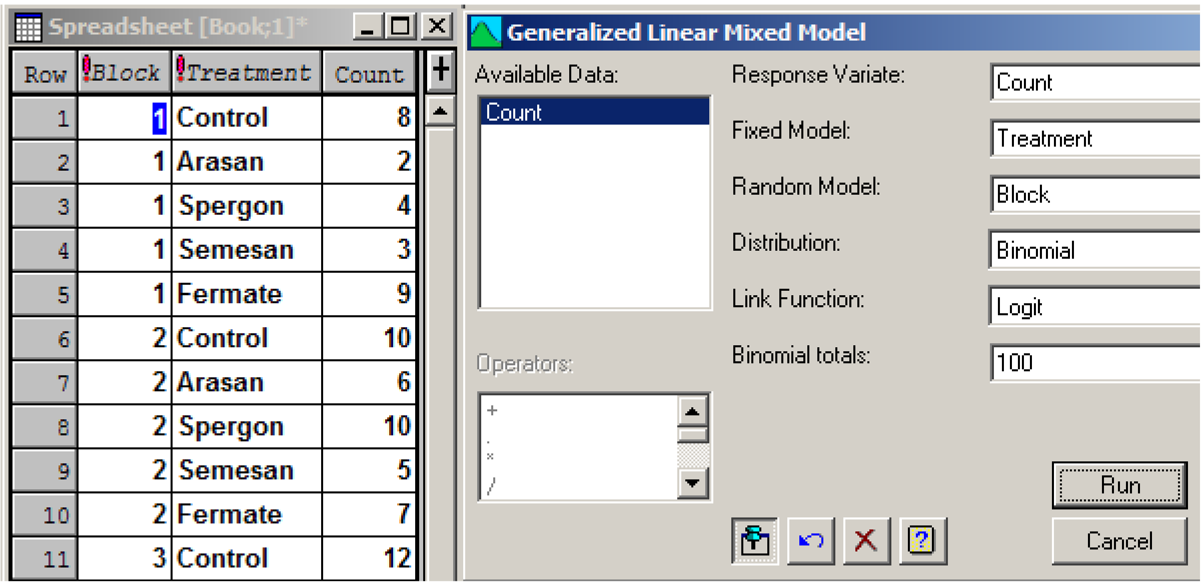
```
Wald tests for dropping terms

           Term        Wald statistic     d.f.      chi. pr.
 Pasture.Status                 9.689        1         0.002
```

## Generalized Linear Mixed Models

Example 9. The number of soybean plants that failed to emerge (each out of 100 plants) using seeds that had one of four treatments or no treatment, from Snedecor and Cochran, page 256.

| Treatment | Block 1 | Block 2 | Block 3 | Block 4 | Block 5 |
|---|---|---|---|---|---|
| Control | 8 | 10 | 12 | 13 | 11 |
| Arasan | 2 | 6 | 7 | 11 | 5 |
| Spergon | 4 | 10 | 9 | 8 | 10 |
| Semesan, Jr. | 3 | 5 | 9 | 10 | 6 |
| Fermate | 9 | 7 | 5 | 5 | 3 |

If these were normally distributed data the analysis would be a standard RCB – in fact this was the analysis used by Snedecor and Cochran. However, the data are binomial counts in a block design. Blocks are usually regarded as random. Hence we need to use a GLMM.



| Row | Block | Treatment | Count |
|---|---|---|---|
| 1 | 1 | Control | 8 |
| 2 | 1 | Arasan | 2 |
| 3 | 1 | Spergon | 4 |
| 4 | 1 | Semesan | 3 |
| 5 | 1 | Fermate | 9 |
| 6 | 2 | Control | 10 |
| 7 | 2 | Arasan | 6 |
| 8 | 2 | Spergon | 10 |
| 9 | 2 | Semesan | 5 |
| 10 | 2 | Fermate | 7 |
| 11 | 3 | Control | 12 |

Available Data: Count

Response Variate: Count

Fixed Model: Treatment

Random Model: Block

Distribution: Binomial

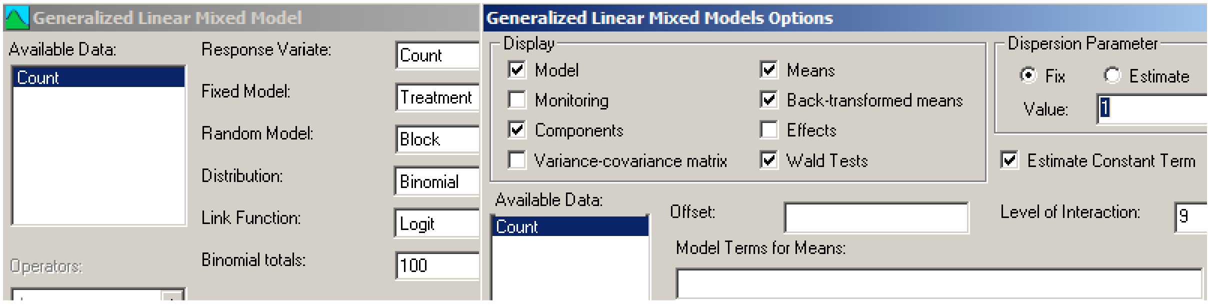
Link Function: Logit

Binomial totals: 100

We have no reason to suspect that the distribution would be over- or under-dispersed. If we estimate the dispersion parameter, we obtain:

```
Residual variance model

Term          Factor        Model(order)        Parameter        Estimate        s.e.
Dispersn                    Identity            Sigma2              0.932      0.3305
```

An estimate of 0.932 with a standard error of 0.33 clearly suggests fixing the value at 1, the expected dispersion parameter under the model. The analysis is as follows:



Generalized Linear Mixed Model

Available Data: Count

Response Variate: Count

Fixed Model: Treatment

Random Model: Block

Distribution: Binomial

Link Function: Logit

Binomial totals: 100

Operators:

Generalized Linear Mixed Models Options

Display

☑ Model ☑ Means
☐ Monitoring ☑ Back-transformed means
☑ Components ☐ Effects
☐ Variance-covariance matrix ☑ Wald Tests

Dispersion Parameter

● Fix ○ Estimate

Value: 1

☑ Estimate Constant Term

Available Data: Count

Offset:

Level of Interaction: 9

Model Terms for Means:

## Generalized linear mixed model analysis

```
Method:                 c.f. Schall (1991) Biometrika
Response variate:       Count
Binomial totals:        100
Distribution:           binomial
Link function:          logit
Random model:           Block
Fixed model:            Constant + Treatment

Dispersion parameter fixed at value 1.000
```

## Estimated variance components

```
Random term                                component          s.e.
Block                                          0.024         0.038
```

## Residual variance model

```
Term          Factor        Model(order)     Parameter        Estimate      s.e.
Dispersn                    Identity         Sigma2              1.000     fixed
```

## Tests for fixed effects

```
Fixed term    Wald statistic   n.d.f.        F statistic          d.d.f.      F pr
Treatment     11.81            4             2.95                   18.0     0.049
```

## Tables of means with standard errors

```
     Treatment
        Arasan       -2.721
       Control       -2.115
       Fermate       -2.792
       Semesan       -2.654
       Spergon       -2.420

   Estimated variance-covariance matrix
           Arasan       0.03736
          Control       0.00528      0.02465
          Fermate       0.00528      0.00528      0.03943
          Semesan       0.00528      0.00528      0.00528      0.03555
          Spergon       0.00528      0.00528      0.00528      0.00528      0.03007
                         Arasan      Control      Fermate      Semesan      Spergon
```

Standard errors of means on the logit scale are obtained as the square root of the variances, the diagonal elements of the variance-covariance matrix.

```
Standard errors of differences between pairs

   Treatment Arasan   1       *
  Treatment Control   2   0.235       *
  Treatment Fermate   3   0.267   0.240       *
  Treatment Semesan   4   0.259   0.231   0.263       *
  Treatment Spergon   5   0.247   0.218   0.251   0.243       *
                              1       2       3       4       5

Back-transformed Means (on the original scale)

     Treatment
        Arasan      6.178
       Control     10.767
       Fermate      5.779
       Semesan      6.577
       Spergon      8.173
```

There is evidence that the probability of failure to emerge differs across treatments (P=0.049). To estimate the probabilities of failure to emerge, divide the back-transformed means (which are mean numbers of failures from 100 seeds) by *n* = 100. We obtain: 0.108 (Control), 0.062 (Arasan), 0.058 (Fermate), 0.066 (Semesan) and 0.082 (Spergon).

Standard errors of back-transformed means and standard errors of differences of back-transformed means are not available because the analysis was done on the logit scale. Confidence intervals for the means on the logit scale can be calculated and the end-points back-transformed (exponentiating to obtain the *odds*, and using *odds*/(1+*odds*) to obtain the probabilities) to provide confidence intervals for individual back-transformed means. The F test for treatments had a denominator df of 18, and we would therefore use a critical t value (2.101) based on 18 df in the calculation of confidence intervals.

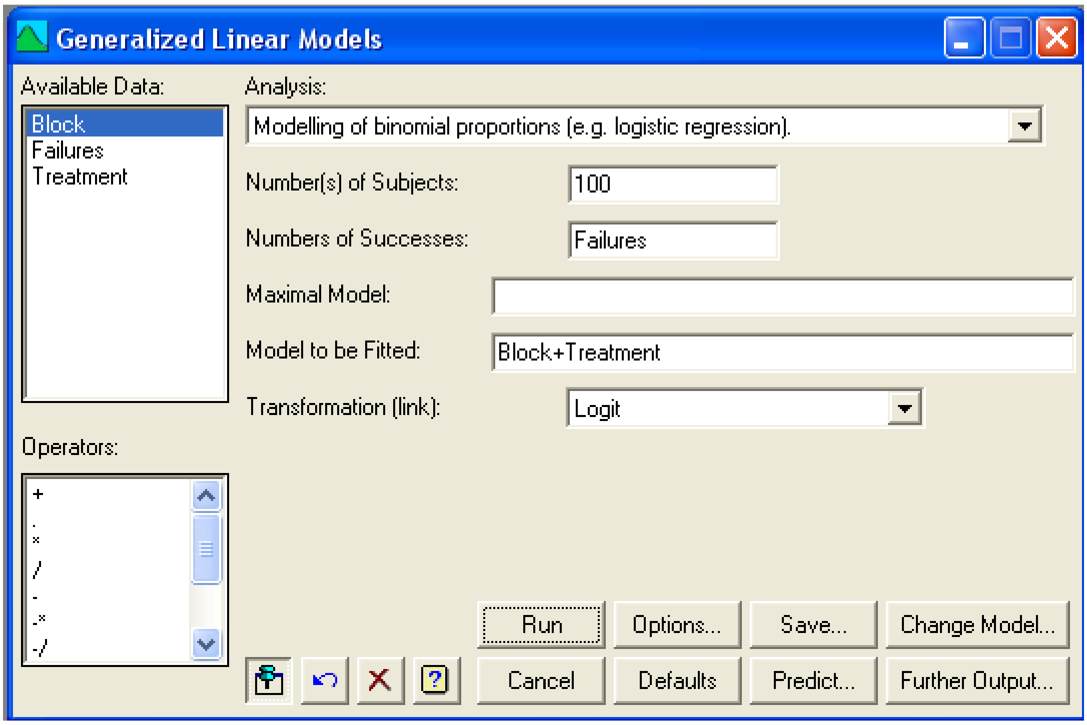
| | logit scale | | | | *odds* scale | | | *odds*(1+*odds*) scale | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 95% CI | | | 95% CI | | | 95% CI | |
| | mean | s.e. | lower | upper | *odds* | lower | upper | *prob* | lower | upper |
| Arasan | -2.721 | 0.193 | -3.127 | -2.315 | 0.066 | 0.044 | 0.099 | **0.062** | 0.042 | 0.090 |
| Control | -2.115 | 0.157 | -2.445 | -1.786 | 0.121 | 0.087 | 0.168 | **0.108** | 0.080 | 0.144 |
| Fermate | -2.792 | 0.199 | -3.209 | -2.375 | 0.061 | 0.040 | 0.093 | 0.058 | 0.039 | 0.085 |
| Semesan | -2.654 | 0.189 | -3.050 | -2.258 | 0.070 | 0.047 | 0.105 | 0.066 | 0.045 | 0.095 |
| Spergon | -2.420 | 0.173 | -2.784 | -2.055 | 0.089 | 0.062 | 0.128 | 0.082 | 0.058 | 0.114 |

Differences in means are tested on the logit scale, and differences and confidence intervals back-transformed as above. *However, differences become* ***odds-ratios*** *when back-transformed.* For example, **Aresan** has a significantly lower failure-to-emerge probability compared to the **Control**, as evidenced by:

Difference on logit scale = 0.606, s.e.d. = 0.235,
t = 0.606/0.235 = 2.58 (18 df, P = 0.019).
*odds-ratio* = $e^{0.606}$ = 1.833
95% CI on logit scale = 0.060 ± 2.101×0.235 = (0.112, 1.099)
95% CI of *odds-ratio* = (1.12, 3.00)

Hence the odds of failing to emerge for the Control is 1.833 times that of Aresan (but this could be as low as 1.12 or as high as 3.00)

The data could also be analysed via a basic generalized linear model if we are prepared to assume blocks are fixed. We turned on Accumulated and Fit model terms individually to allow the contribution from blocks to be measured as well as the effect of treatments.



Generalized Linear Models

Available Data: Block, Failures, Treatment

Analysis: Modelling of binomial proportions (e.g. logistic regression).

Number(s) of Subjects: 100

Numbers of Successes: Failures

Maximal Model:

Model to be Fitted: Block+Treatment

Transformation (link): Logit

```
Regression analysis

      Response variate:  Failures
      Binomial totals:   100
          Distribution:  Binomial
         Link function:  Logit
          Fitted terms:  Constant, Block, Treatment

Summary of analysis

                                              mean    deviance    approx
Source          d.f.      deviance        deviance       ratio    chi pr
Regression         8         18.91          2.3639        2.36     0.015
Residual          16         15.01          0.9380
Total             24         33.92          1.4133
Change            -4        -11.50          2.8741        2.87     0.022

Dispersion parameter is fixed at 1.00.

Message: deviance ratios are based on dispersion parameter with value 1.

Estimates of parameters

                                                                      antilog of
Parameter            estimate          s.e.       t(*)     t pr.      estimate
Constant               -3.113         0.263     -11.82     <.001       0.04448
Block 2                 0.407         0.263       1.55     0.122         1.502
Block 3                 0.516         0.258       2.00     0.046         1.676
Block 4                 0.640         0.253       2.53     0.011         1.897
Block 5                 0.318         0.267       1.19     0.234         1.374
```

A similar *estimated* dispersion parameter as was found using a GLMM. A deviance of 15.01 can be tested using a $\chi^2$ distribution with 16 df.

```
Treatment Control          0.607         0.235       2.58   0.010      1.835
Treatment Fermate         -0.071         0.266      -0.27   0.790     0.9314
Treatment Semesan, Jr.     0.067         0.259       0.26   0.796      1.069
Treatment Spergon          0.302         0.247       1.22   0.222      1.353

Message: s.e.s are based on dispersion parameter with value 1.

Parameters for factors are differences compared with the reference level:
                    Factor   Reference level
                     Block   1
                 Treatment   Arasan

Accumulated analysis of deviance

                                                mean   deviance     approx
Change                  d.f.     deviance   deviance      ratio    chi pr
+ Block                    4       7.4145     1.8536       1.85     0.116
+ Treatment                4      11.4964     2.8741       2.87     0.022
Residual                  16      15.0087     0.9380
Total                     24      33.9196     1.4133
```

The GLMM gave a Wald F statistic of 2.95, consistent with the above. The model was referenced to Arasan, and we can use the P value of 0.010 (from the Treatment Control line in the Estimates of parameters part of the analysis) to conclude that the probability of failure for untreated seeds is different to that for Arasan-treated seeds. Calculations following the GLMM gave a P value of 0.019.) Furthermore, no other seed treatment was significant when compared to Arasan.

## Ordinal logistic regression

Occasionally scientists will only be able to score plants or plots and special analyses need to be used for score data. In this section, we introduce ordinal logistic regression for ordered scores.

Example 10
We will take the data from Snedecor and Cochran page 205. Their data involves ordered scores for improvement in health of leprosy sufferers. They analysed the data as a *t* test. We will use the same data, but imagine them to come from the following plant pathology experiment. Suppose *Sclerotinia sclerotiorum* was tested as a biological control of the noxious weed bitoubush (*Chrysanthemoides monilifera* ssp. *Rotundata*). Two isolates were assessed for pathogenicity, and varying numbers of plants were assessed per isolate. We will use the following 5-point ordered scale.

1 = no reaction
2 = lesions confined to <20% of leaves
3 = lesions confined to 20% to 50% of leaves
4 = lesions confined to 50% to 70% of leaves
5 = lesions confined to >70% of leaves

The data are the same as in Snedecor and Cochran. Here we present it in two ways. Firstly, we have 144 random plants with isolate 1 with varying scores, followed by 52 random plants.

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Isolate 1 | 2 | 3 | 2 | 2 | 3 | 4 | 2 | 2 | 5 | 2 | 2 | 4 | 2 | 1 | 3 | 2 | 2 | 3 | 2 | 4 |
| | 2 | 3 | 2 | 3 | 4 | 4 | 2 | 3 | 3 | 2 | 3 | 3 | 5 | 2 | 4 | 1 | 2 | 3 | 4 | 2 |
| | 3 | 3 | 2 | 2 | 3 | 2 | 5 | 2 | 3 | 3 | 3 | 1 | 4 | 2 | 4 | 2 | 3 | 1 | 3 | 4 |
| | 3 | 1 | 5 | 4 | 3 | 2 | 3 | 5 | 3 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 | 4 | 2 | 3 |
| | 4 | 2 | 2 | 4 | 3 | 2 | 2 | 2 | 4 | 2 | 2 | 2 | 4 | 4 | 2 | 4 | 2 | 5 | 2 | 3 |
| | 1 | 2 | 4 | 4 | 3 | 5 | 2 | 2 | 2 | 5 | 3 | 2 | 2 | 1 | 4 | 2 | 3 | 3 | 3 | 4 |
| | 3 | 4 | 2 | 2 | 2 | 3 | 3 | 5 | 2 | 3 | 4 | 5 | 3 | 2 | 5 | 4 | 1 | 2 | 2 | 3 |
| | 2 | 4 | 3 | 1 | | | | | | | | | | | | | | | | |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Isolate 2 | 2 | 3 | 3 | 2 | 4 | 3 | 5 | 5 | 4 | 2 | 2 | 3 | 3 | 2 | 3 | 4 | 4 | 4 | 2 | 5 |
| | 3 | 5 | 2 | 3 | 2 | 5 | 2 | 4 | 3 | 5 | 4 | 4 | 4 | 1 | 3 | 3 | 2 | 4 | 2 | 4 |
| | 2 | 5 | 4 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 2 | | | | | | | | |

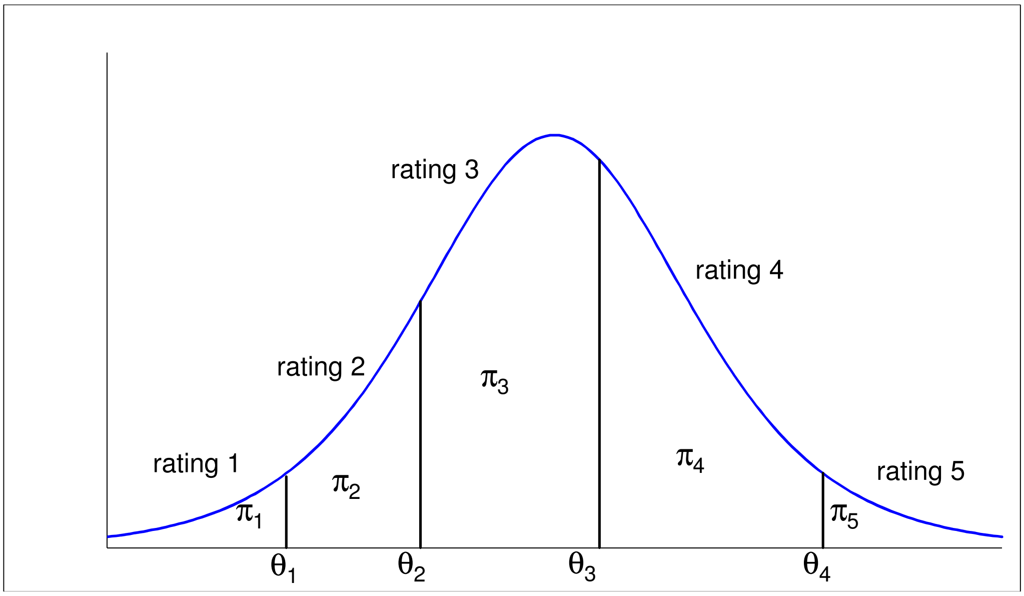
GenStat would need the scores in a single ***factor*** column as well as a factor column to identify the isolate for each plant.

Alternatively, we could supply the data in frequency form. The scores would need to be in five ***variate*** columns, and we would need a factor to identify the isolates for each row of frequencies.

| Isolate | Score1 | Score2 | Score3 | Score4 | Score5 |
|---|---|---|---|---|---|
| 1 | 11 | 53 | 42 | 27 | 11 |
| 2 | 1 | 13 | 16 | 15 | 7 |

Now plants do not suddenly jump in discrete steps from one score to the next. Rather, there is a continuous change in the severity of damage of the plant. A severity score of 1 represents undamaged plants, although damage may be slowly taking place, perhaps unseen. These plants have a score while the damage is confined to a point we call the first cut-point, $\theta_1$. A score of 2 then represents plants with damage from $\theta_1$ to some new cut-point $\theta_2$.

So, generally some underlying continuous distribution is assumed, such as logistic. For the 5-point rating scale under discussion, this would appear as follows, assuming an underlying logistic distribution.



We do not say that the scores necessarily represent equal spacings on this continuous scale. To quantify the discussion to date, suppose we use $Y$ for the continuous damage variable. Then a rating of 1 represents plants whose damage value on the continuous scale is any $Y < \theta_1$. The underlying probability of obtaining a plant with this rating is $\pi_1 = \mathrm{P}(Y < \theta_1)$. We do not know $\theta_1$ and we don't know $\pi_1$.

Similarly, a rating 2 represents a plant whose damage value on the continuous scale is anything between $\theta_1$ and $\theta_2$. The underlying probability of obtaining a plant with this rating is $\pi_2 = \mathrm{P}(\theta_1 < Y < \theta_2)$. We don't know $\theta_2$ and we don't know $\pi_2$. And so on.

It is actually simpler to model the *cumulative* probabilities $\mathrm{P}(Y < \theta_1) = \pi_1$, $\mathrm{P}(Y < \theta_2) = \pi_1+\pi_2$, $\mathrm{P}(Y < \theta_3) = \pi_1+\pi_2+\pi_3,\ldots$. For notation we will define

$\gamma_1 = \mathrm{P}(Y < \theta_1) = \pi_1$,<br>
$\gamma_2 = \mathrm{P}(Y < \theta_2) = \pi_1+\pi_2$,<br>
$\gamma_3 = \mathrm{P}(Y < \theta_3) = \pi_1+\pi_2+\pi_3$ and so on. (The last value must be 1.)

Thus, for a 5-point scale we need to estimate four cut-points $\theta_1$ to $\theta_4$ and four probabilities $\pi_1$ to $\pi_4$ (since the 5th rating is a value larger than $\theta_4$ and $\pi_5$ is $1 - \pi_1 - \pi_2 - \pi_3 - \pi_4$) and hence $\gamma_1$ to $\gamma_4$.

Now we propose a set of logistic regression equations for the cumulative probabilities:

$$\gamma_i = \frac{1}{1+e^{-(\theta_i - b_1 X_1 - \ldots)}}$$

where $\{\theta_i\}$ are the cut-points for the ordered scale and $X_1$, …, are covariates, or, as in the case of a designed experiment, the usual design features. On the logit scale this becomes

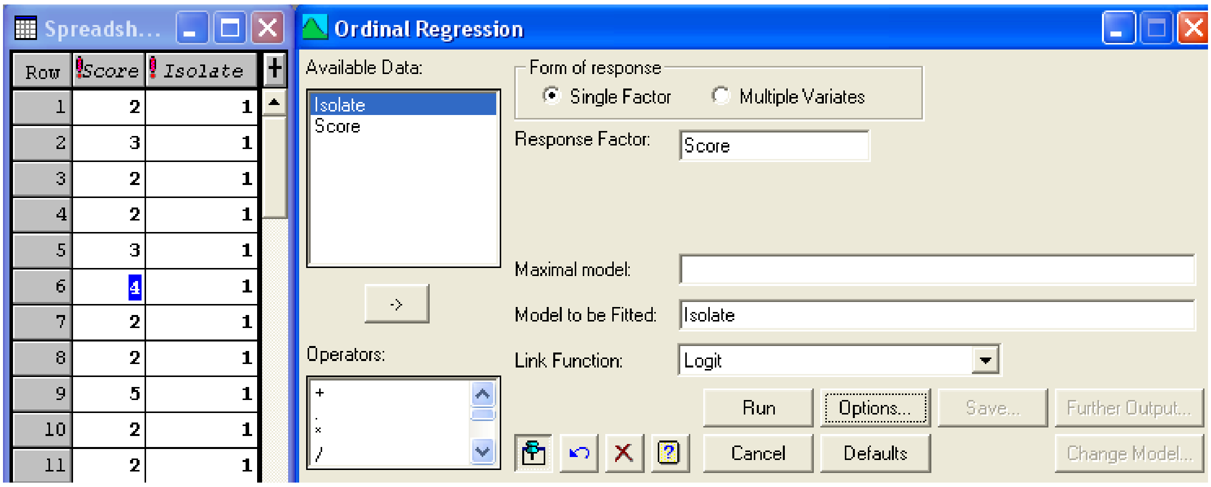
$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right) = \theta_i - b_1 X_1 - \ldots \qquad \text{for ratings } i = 1, 2, \ldots$$

Note that this is sometimes referred to as the *proportional odds model*, because, for a given state (score), the ratio of the odds does not depend on the state. Notice that the *log-odds* value, and hence the odds ratio, are calculated on the *cumulative* scale (ie using the $\gamma_i$, not the $\pi_i$). By difference, once we have estimated the cumulative probabilities we can calculate the individual probabilities.

The parameters are estimated by maximum likelihood as with ordinary logistic regression.

Choose Stats > Regression Analysis > Ordinal Regression. Then use either one of the following methods.

Method 1 Individual plant scores (as a factor) with a treatment factor column:



| Row | Score | Isolate |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 3 | 1 |
| 3 | 2 | 1 |
| 4 | 2 | 1 |
| 5 | 3 | 1 |
| 6 | 4 | 1 |
| 7 | 2 | 1 |
| 8 | 2 | 1 |
| 9 | 5 | 1 |
| 10 | 2 | 1 |
| 11 | 2 | 1 |

Available Data: Isolate, Score

Form of response: ● Single Factor ○ Multiple Variates

Response Factor: Score

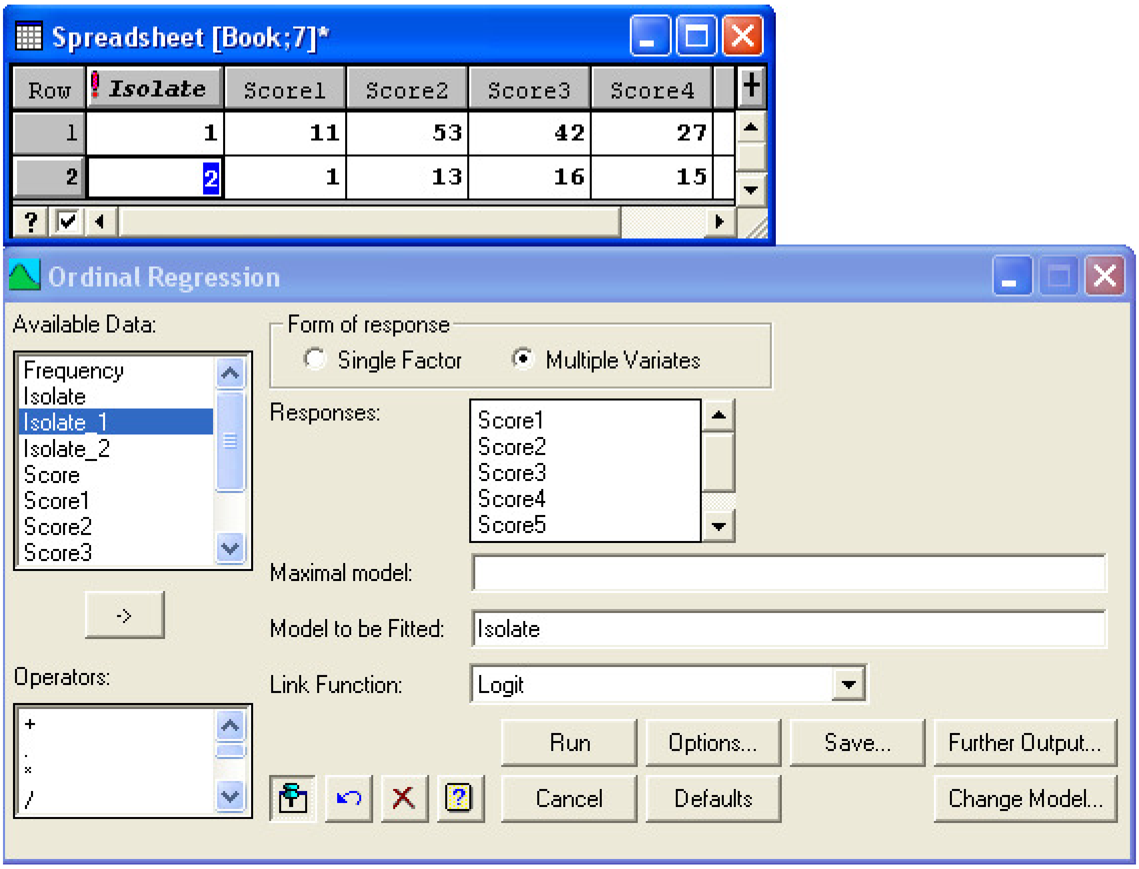
Maximal model:

Model to be Fitted: Isolate

Link Function: Logit

The same analysis is obtained by the following method.

Method 2 Individual score variates of frequencies with a treatment factor column:



Spreadsheet [Book;7]*

| Row | Isolate | Score1 | Score2 | Score3 | Score4 |
|---|---|---|---|---|---|
| 1 | 1 | 11 | 53 | 42 | 27 |
| 2 | 2 | 1 | 13 | 16 | 15 |

Ordinal Regression

Available Data: Frequency, Isolate, Isolate_1, Isolate_2, Score, Score1, Score2, Score3

Form of response: ○ Single Factor ● Multiple Variates

Responses: Score1, Score2, Score3, Score4, Score5

Maximal model:

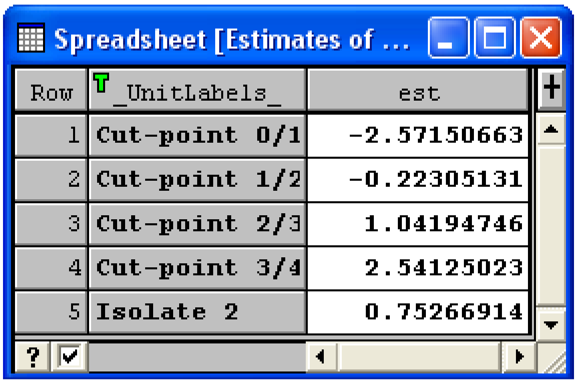
Model to be Fitted: Isolate

Link Function: Logit

```
         Response variates:  ordinal model for categories defined by
                            Score1, Score2, Score3, Score4, Score5
              Distribution:  Multinomial
             Link function:  Logit
              Fitted terms:  Isolate

Summary of analysis

                                                 mean    deviance    approx
Source           d.f.        deviance        deviance       ratio    chi pr
Regression          1             6.7           6.679        6.68     0.010
Residual          191           560.6           2.935
Total             192           567.3           2.955

Dispersion parameter is fixed at 1.00.

Estimates of parameters

                                                                      antilog of
Parameter              estimate          s.e.       t(*)     t pr.      estimate
Cut-point 0/1            -2.572         0.303      -8.48     <.001       0.07642
Cut-point 1/2            -0.223         0.164      -1.36     0.173        0.8001
Cut-point 2/3             1.042         0.180       5.78     <.001         2.835
Cut-point 3/4             2.541         0.270       9.41     <.001         12.70
Isolate 2                 0.753         0.295       2.55     0.011         2.123

Parameters for factors are differences compared with the reference level:
                   Factor   Reference level
                  Isolate   1
```

The two isolates have significantly different (*P*=0.010) probability distributions of the five scores.

Before estimating the individual probability distributions for the two isolates, it is wise to save the estimates so they can be opened in Excel with full accuracy.



Spreadsheet [Estimates of ...

| Row | _UnitLabels_ | est |
|---|---|---|
| 1 | Cut-point 0/1 | -2.57150663 |
| 2 | Cut-point 1/2 | -0.22305131 |
| 3 | Cut-point 2/3 | 1.04194746 |
| 4 | Cut-point 3/4 | 2.54125023 |
| 5 | Isolate 2 | 0.75266914 |

We can interpret the model as follows.

The model is referenced to isolate 1. With a treatment factor with only two levels (isolate 1 and 2) we have only one predictor in the model. Hence $X_1$ = 1 for isolate 2 and 0 otherwise.

**For isolate 1**

$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right)=\theta_i \text{ for } i = 1, 2, 3 \text{ and } 4.$$

The back-transform is given in the output as the antilog of estimate. Thus, the odds for a score of 1 are 0.076420. Hence the estimate of the probability for a score of 1 ($\gamma_1$) is 0.076420/(1+0.076420) = 0.070995. For this cut-point, $\gamma_1$ and $\pi_1$ are the same.

The odds for a score of 1 or 2 are 0.800074. Hence the estimate of the probability for a score of 1 or 2 ($\gamma_2$) is 0.800074/(1+0.800074) = 0.444467. By subtraction, the estimate for $\pi_2$ is 0.444467-0.070995 = 0.373472. And so on.

**For isolate 2**

$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right)=\theta_i+0.752669 \text{ for } i = 1, 2, 3 \text{ and } 4.$$

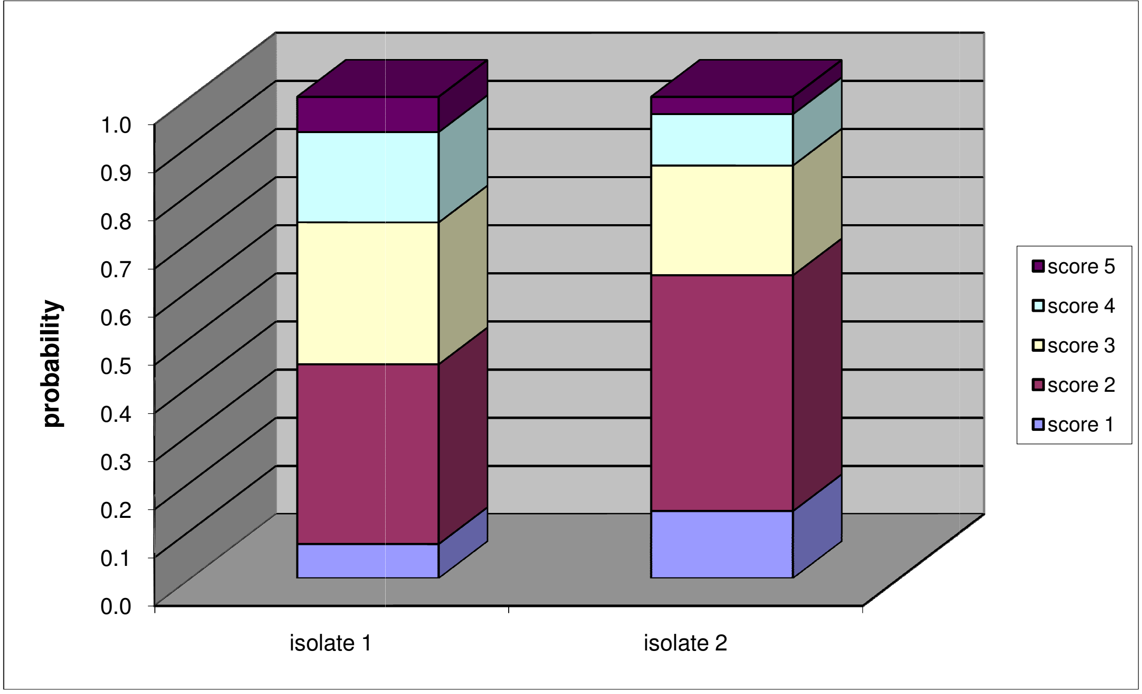
This means that the odds already worked out for the reference isolate simply need to be multiplied by $e^{0.752669}$, the antilog of $b_1$ being 2.122658. Thus:

The odds for a score of 1 are 0.076420×2.122658 = 0.162214. Hence the estimate of the probability for a score of 1 ($\gamma_1$) is 0.162214/(1+0.162214) = 0.139537. For this cut-point, $\gamma_1$ and $\pi_1$ are the same. And so on.

This is very easy to do in Excel. Here the cells used are marked (starting from V3), and formulae for the two calculation columns shown alongside.

|  | V | W | X | Y | X | Y |
|---|---|---|---|---|---|---|
| **13** |  |  | antilog | antilog | antilog | antilog |
| **14** |  |  | isolate 1 | isolate 2 | isolate 1 | isolate 2 |
| **15** |  |  | **Odds** | 2.122658 | **Odds** | =EXP(W20) |
| **16** | Cut-point 0/1 | -2.571507 | 0.076420 | 0.162214 | =EXP(W16) | =X16*Y$15 |
| **17** | Cut-point 1/2 | -0.223051 | 0.800074 | 1.698283 | =EXP(W17) | =X17*Y$15 |
| **18** | Cut-point 2/3 | 1.041947 | 2.834732 | 6.017167 | =EXP(W18) | =X18*Y$15 |
| **19** | Cut-point 3/4 | 2.541250 | 12.695533 | 26.948277 | =EXP(W19) | =X19*Y$15 |
| **20** | isolate 2 | 0.752669 |  |  |  |  |
| **21** |  |  |  |  |  |  |
| **22** |  | **score** | **gammas** |  | **gammas** |  |
| **23** |  | 1 | 0.070995 | 0.139573 | =X16/(1+X16) | =Y16/(1+Y16) |
| **24** |  | 2 | 0.444467 | 0.629394 | =X17/(1+X17) | =Y17/(1+Y17) |
| **25** |  | 3 | 0.739226 | 0.857492 | =X18/(1+X18) | =Y18/(1+Y18) |
| **26** |  | 4 | 0.926983 | 0.964220 | =X19/(1+X19) | =Y19/(1+Y19) |
| **27** |  | 5 | 1 | 1 | 1 | 1 |
| **28** |  | **score** | **probabilities** |  | **probabilities** |  |
| **29** |  | 1 | 0.070995 | 0.139573 | =X23 | =Y23 |
| **30** |  | 2 | 0.373472 | 0.489821 | =X24-X23 | =Y24-Y23 |
| **31** |  | 3 | 0.294758 | 0.228098 | =X25-X24 | =Y25-Y24 |
| **32** |  | 4 | 0.187758 | 0.106727 | =X26-X25 | =Y26-Y25 |
| **33** |  | 5 | 0.073017 | 0.035780 | =X27-X26 | =Y27-Y26 |

Excel has very good plotting techniques to illustrate the difference in the estimated probability distributions in cells X29:Y33:


probability
isolate 1
isolate 2
score 5
score 4
score 3
score 2
score 1