Genstat®

Regression

www.vsni.co.uk

VSNi

A Guide to Regression, Nonlinear and Generalized Linear Models in Genstat®
(22nd Edition)

by Roger Payne.

Genstat is developed by VSN International Ltd, in collaboration with practising statisticians at Rothamsted and other organisations in Britain, Australia, New Zealand and The Netherlands.

Published by: VSN International, 2 Amberside, Wood Lane,
Hemel Hempstead, Hertfordshire HP2 4TP, UK
E-mail: info@genstat.co.uk
Website: http://www.genstat.co.uk/

First published 2008, for GenStat *for Windows* 11th Edition
This edition published 2022, for Genstat *for Windows* 22nd Edition

Contents

Introduction

Regression is one of the most popular methods in statistics, and one that is still producing new and exciting techniques. Genstat has a very powerful set of facilities for regression and generalized linear models that are nevertheless very straightforward and easy to use.

This book shows how Genstat's menus guide you from simple even to very complicated analyses, and also introduces the regression commands that you can use to program any non-standard analyses that you need. We start by explaining ordinary linear regression (with one or several variables), and then extend the ideas to nonlinear models and on to generalized linear models – so that you can analyse counts and proportions as well as the more usual numeric variables. Finally we introduce some of the most recent developments in generalized linear models, including Youngjo Lee and John Nelder's hierarchical generalized linear models, to bring you fully up-to-date with the range of possibilities. The book was written to provide the notes for VSN's 2-day course on Regression, Nonlinear and Generalized Linear Models, but it can be used equally well as a self-learning tool.

The chapters cover the following topics.

1 Linear regression: ranging from simple linear regression (with one variable) to multiple linear regression (several variables) and the modelling of parallel-line relationships (regression models with groups); plotting of residuals to assess the assumptions, and of the fitted model and data to assess the fit; methods for finding the best models when there are many explanatory variables.
2 Nonlinear models: Genstat's range of standard curves, and the facilities for defining your own nonlinear models.
3 Generalized models: how to analyse non-Normal data such as counts and proportions; recent advances – how to use generalized linear mixed models and hierarchical generalized linear models to handle additional sources of random variation.

Acknowledgement: Peter Lane's collaboration on the original Genstat regression courses – and on the regression source code itself – is gratefully acknowledged.

1 Linear regression

In this chapter you will learn

- how to fit a regression model with a single explanatory variable
- what the output means
- how to plot the fitted model
- what assumptions are made for the analysis, and how to check them
- what commands are used to fit, display and assess linear regressions ★
- how to perform a permutation test to assess a regression ★
- how to save results in Genstat data structures for future use ★
- how to make predictions from a regression analysis
- how to fit a multiple linear regression (with several explanatory variables)
- how to explore alternative models when there are several explanatory variables
- how to use all subsets regression to assess and summarize all available models ★
- how to fit parallel and non-parallel regression lines when you have an explanatory factor as well as an explanatory variate

Note: the topics marked ★ are optional.

1.1 Simple linear regression

Linear regression is a method of describing a relationship between one variable and one or more others:

- the *response variable* (also called the *y-variable* or *dependent variable*) is the variable to be described;
- the *explanatory variables* (also called the *x-variables* or *independent variables*) are the variables used to describe the response variable.

With a "simple linear regression" you have only one explanatory variable, say x. So you want to describe the response variate y by the *model*

$$y = b \times x + c$$

where the *parameters* of the model are

b the *regression coefficient*, and

c the *constant*.

In simple linear regression, the constant c is often called the *intercept* as it is the value of y when x is zero. We will explain later how you can fit models without a constant. Usually, however, the constant is included. The regression coefficient b is often called the *slope* of the regression line.

The model above represents the theoretical value that we are assuming for y, but in practical situations this is unlikely to be what we observe. There may be random variation, or the model may even be just an approximation to the true situation. Suppose we have made n observations of x and y, which we will label with the suffix i. We can define a statistical model to describe our observations as

$$y_i = b \times x_i + c + \varepsilon_i \qquad i = 1 \ldots n$$

where now

ε_i is the *residual* for observation i, representing the difference between the value y_i actually observed for observation i, and the theoretical value predicted by the model.

The theoretical value predicted by the model is known as the *fitted value*

$$f_i = b \times x_i + c \qquad i = 1 \ldots n$$

In ordinary linear regression, the residuals ε_i are assumed to come from independent Normal distributions, all with the same variance. In Section 1.3 we show how you can check this assumption, and in Chapter 3 we show how you can fit models to data from other distributions.

We estimate the parameter values by *least squares*, that is by taking the values that minimize the sum of the squared values of the residuals

$$\sum_i \varepsilon i^2 = \sum_i (y_i - b \times x_i - c)^2$$

If the residuals really do have Normal distributions, these estimates are the *maximum likelihood* estimates (that is, the parameter values that would give the highest probability for the data values that we have observed). The assumption of a Normal distribution is also required for the statistical tests described later in this section. However, we will not go into any more detail of the mathematics statistical theory here. More information can be found in standard statistical text books, such as *Applied Regression Analysis* by Draper & Smith (1981, Wiley, New York).

The data sets that are used in the examples and practicals in this Guide can be all be accessed from within Genstat. Click on File on the menu bar, and select the Open Example Data Sets option, as shown in Figure 1.1.

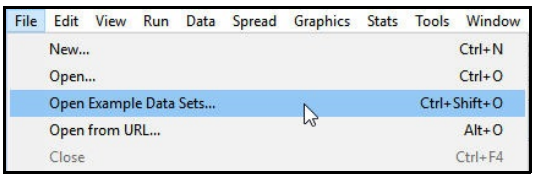


Figure 1.1

This opens the Example Data Sets menu, shown in Figure 1.2. It is easier to find the relevant file if you set the Filter by topic drop-down list to A Guide to Regression, Nonlinear and Generalized Linear Models. Here we shall open the Spreadsheet file `Pressure.gsh` (Figure 1.3) which contains recordings of blood-pressure from a sample of 38 women whose ages range from 20 to 80.

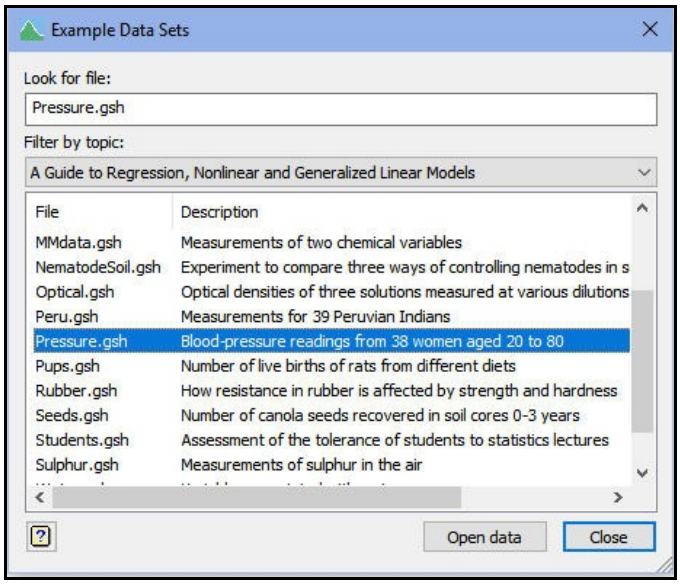
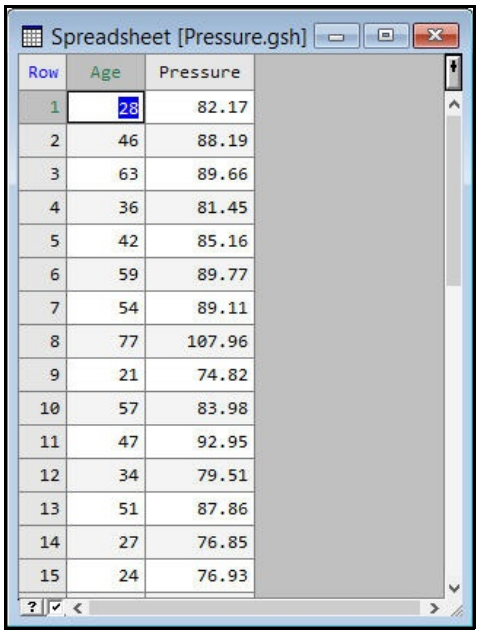


Figure 1.2



Row	Age	Pressure
1	28	82.17
2	46	88.19
3	63	89.66
4	36	81.45
5	42	85.16
6	59	89.77
7	54	89.11
8	77	107.96
9	21	74.82
10	57	83.98
11	47	92.95
12	34	79.51
13	51	87.86
14	27	76.85
15	24	76.93

Figure 1.3

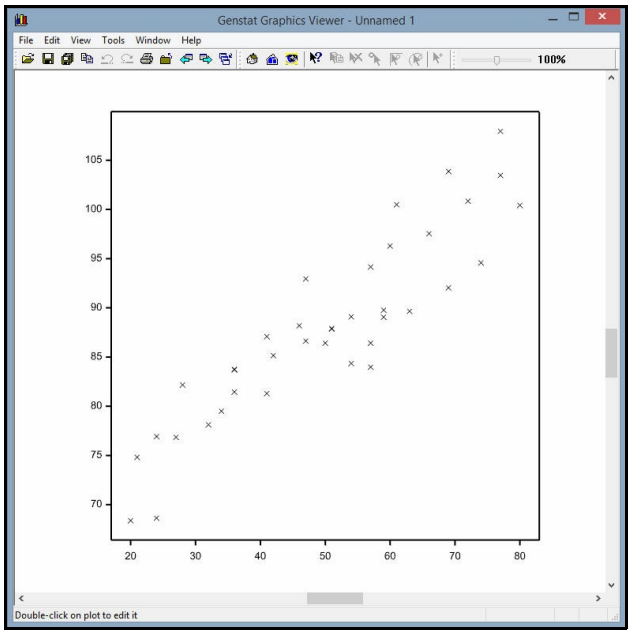


Figure 1.4

We can plot a graph of pressure against age (Figure 1.4) by using the Graphics menu and selecting 2D Scatter Plot. This shows a fairly linear relationship between blood-pressure and age, so it would seem sensible to fit a linear regression.

Figure 1.5 shows the regression line, and the residuals as vertical lines joining the fitted value on the regression line to the data point.

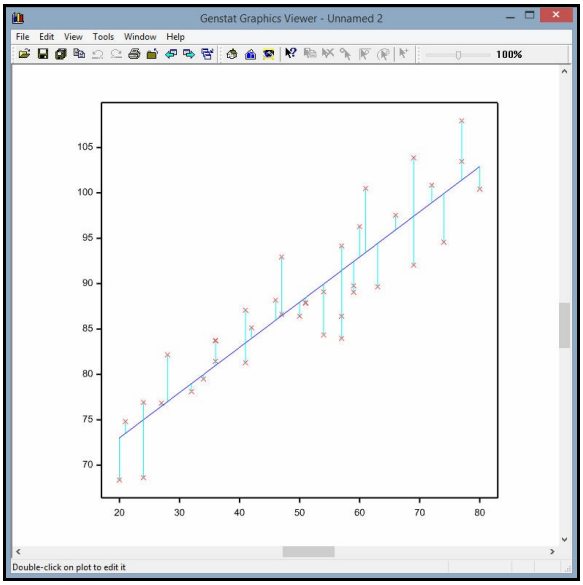


Figure 1.5

To fit the regression in Genstat, you select the Regression Analysis option of the Stats menu on the menu bar, and then clicking on the Linear sub-option as shown in Figure 1.6.

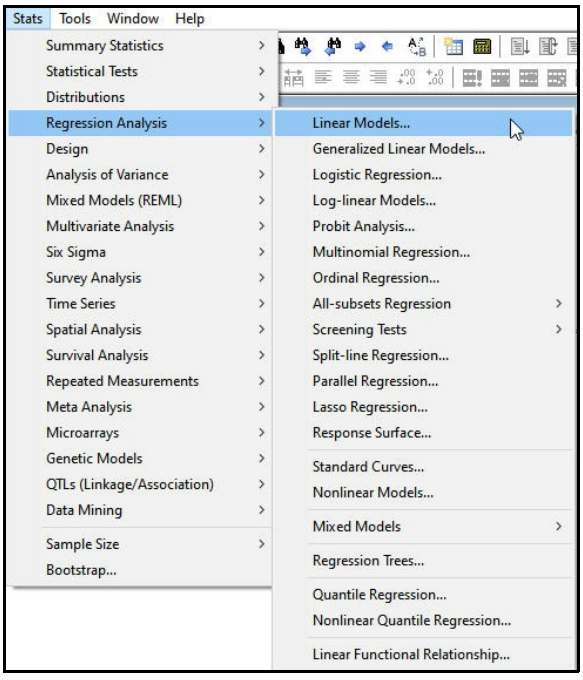


Figure 1.6

This opens the Linear Regression menu, shown in Figure 1.7. If you select the Simple linear regression option in the drop-down list at the top of the menu, the menu customizes itself so that you just need to fill in boxes to specify the Response variate and Explanatory variate. Clicking on Run produces the output below.

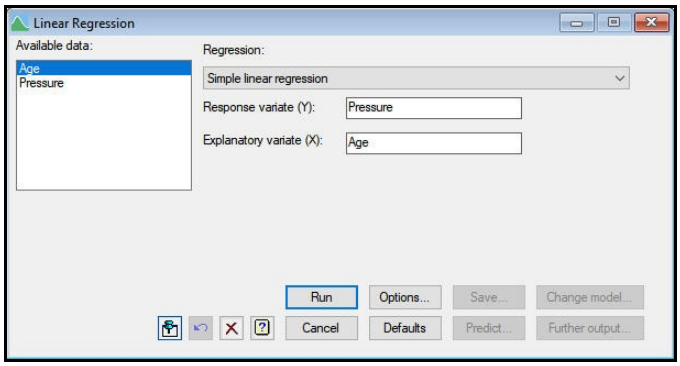


Figure 1.7

Regression analysis

Response variate: Pressure
Fitted terms: Constant, Age

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	2647.7	2647.69	169.73	<.001
Residual	36	561.6	15.60		
Total	37	3209.3	86.74		

Percentage variance accounted for 82.0
Standard error of observations is estimated to be 3.95.

Estimates of parameters

Parameter	estimate	s.e.	t(36)	t pr.
Constant	63.04	2.02	31.27	<.001
Age	0.4983	0.0382	13.03	<.001

The output to display is controlled by the Linear Regression Options menu (Figure 1.8), which is opened by clicking on the Options button in the Linear Regression menu. The default output begins with a description of the model, listing the response variable and the fitted terms: these are the constant and the explanatory variable. The constant is included by default; if you want to omit it, you should uncheck the Estimate constant term box. This would constrain the fitted line to pass through the origin (that is, the response must be zero when the explanatory is zero), but remember that the analysis would still be based on the assumptions that the variability about the line is constant for the whole range of the data, and that the relationship is linear right down to the origin. So this may not be sensible, particularly if you have observations close to the origin.

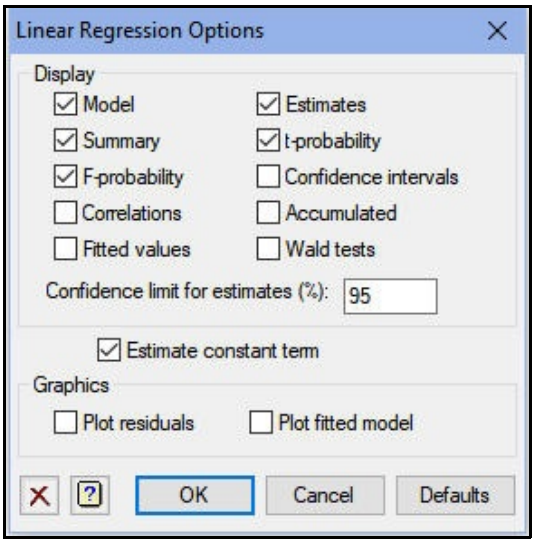


Figure 1.8

The next section of output contains an analysis of variance to help you assess the model. In the "`s.s.`" column, the "`Residual`" line contains the sum of squares of the residuals, and this is regarded as random variation. The "`Total`" line contains the residual sum of squares for a model that contains just the constant. In this model, the constant will be estimated as the mean of the values of the response variable (i.e. the *grand mean*). So this line contains

$$\sum_i (y - \mu)^2$$

where μ is the grand mean

$$\mu = \sum_i y / n$$

and, more accurately, is the total sum of squares "corrected" for the grand mean.

In a linear regression, we are interested to see whether there really is evidence of a linear relationship with the explanatory variable. So we want to compare the model containing just the constant with the model containing the constant *and* (a regression coefficient for) the explanatory variable. The difference between the residual sums of squares of these two models is printed in the `Regression` line:

$$2647.7 = 3209.3 - 561.6$$

This is known as the sum of squares "due to the regression", and represents the amount of variation that can be "explained" (i.e. removed from the residual) by including a regression coefficient for the explanatory variable in the model.

The "`d.f.`" (degrees of freedom) column records the number of independent parameters that are contributing to each sum of squares. In the "`Total`" line this is 37 (the number of observations minus one, as we have fitted a constant term. In the "`Residual`" line this is 36 (the number of observations minus two, as we have fitted a constant term and the regression coefficient for the explanatory variable). In the `Regression` line this is one, as this line represents the effect of adding one more parameter to the model.

The "`m.s.`" (mean square) column contains the sums of squares divided by the degrees of freedom, which converts them to variances. The "`v.r.`" (variance ratio) column shows the regression mean square divided by the residual mean square. Under the null hypothesis that there is no linear relationship between the response and the explanatory variables, this will have an F distribution with the degrees of freedom in the `Regression` and `Residual` lines i.e. the printed value 169.73 would be from an F distribution on one and 36 degrees of freedom. The "`F pr.`" column prints the corresponding probability. The value here is less than 0.001 (<.001), so the relationship is significant at a 0.1% level of significance.

It is important to remember, however, that the use of the F distribution depends on the assumption that the residuals have independent Normal distributions, all with the same variance, and we will show how you can assess that in Section 1.3.

The *percentage variance accounted for* is a summary of how much of the variability of this set of observations has been explained by the fitted model. It is the difference between residual and total mean squares expressed as a percentage of the total mean square. When expressed as a proportion rather than a percentage, this statistic is called the *adjusted* R^2; it is not quite the same as R^2, the squared coefficient of correlation. The adjustment takes account of the number of parameters in the model compared to the number of observations.

The final section of the output shows the estimated values for the parameters in the model. The regression coefficient for `Age` is 0.4983, with a standard error of 0.0382. So the model predicts that blood pressure will rise by 0.4983 units with each additional year. The corresponding t-statistic is large, 13.03 with 36 degrees of freedom, again indicating that there is a significant linear relationship between pressure and age. In fact, when the regression model has only one degree of freedom, the t-statistic in the table of estimates is the square root of the F statistic in the analysis of variance. So this is actually making the same test. Again, the use of the t distribution is based on the assumptions of the regression.

You can obtain further output by clicking on Further Output in the Linear Regression menu. So, if you are trying several different regression models, as we show in Section 1.11, you may want to omit some of the default output by unchecking the relevant boxes of the Linear Regression Options menu (Figure 1.8) until you have decided which model is best.

The resulting Linear Regression Further Output menu is shown in Figure 1.9. For example, if we check the Fitted values box and click on Run, we obtain the output below.

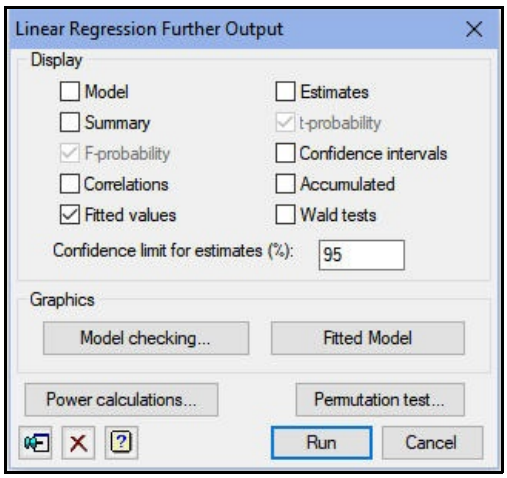


Figure 1.9

Regression analysis

Fitted values and residuals

Unit	Response	Fitted value	Standardized residual	Leverage
1	82.17	77.00	1.36	0.072
2	88.19	85.97	0.57	0.028
3	89.66	94.44	-1.24	0.042
4	81.45	80.98	0.12	0.045
5	85.16	83.97	0.31	0.032
6	89.77	92.44	-0.69	0.034
7	89.11	89.95	-0.22	0.028
8	107.96	101.41	1.74	0.095
9	74.82	73.51	0.35	0.105
10	83.98	91.45	-1.92	0.031
11	92.95	86.46	1.67	0.027
12	79.51	79.99	-0.12	0.050
13	87.86	88.46	-0.15	0.026
14	76.85	76.50	0.09	0.076
15	76.93	75.00	0.51	0.090
16	87.09	83.47	0.93	0.034
17	97.55	95.93	0.42	0.050
18	92.04	97.43	-1.41	0.060
19	100.85	98.92	0.51	0.072
20	96.30	92.94	0.87	0.036
21	86.42	87.96	-0.39	0.026
22	94.16	91.45	0.70	0.031
23	78.12	78.99	-0.23	0.057
24	89.06	92.44	-0.87	0.034
25	94.58	99.92	-1.41	0.080
26	103.48	101.41	0.55	0.095
27	81.30	83.47	-0.56	0.034
28	83.71	80.98	0.71	0.045
29	68.38	73.01	-1.24	0.111
30	86.64	86.46	0.05	0.027
31	87.91	88.46	-0.14	0.026

32	86.42	91.45	-1.29	0.031
33	103.87	97.43	1.68	0.060
34	83.76	80.98	0.72	0.045
35	84.35	89.95	-1.44	0.028
36	68.64	75.00	-1.69	0.090
37	100.50	93.44	1.82	0.038
38	100.42	102.91	-0.67	0.111
Mean	87.95	87.95	0.00	0.053

As explained earlier, the *fitted values* are those predicted by the model for each observation: $b \times x_i + c$. The residuals ε_i are differences between the observed values of the explanatory variable y_i and the fitted values. However, in the table these *simple residuals*, ε_i, have been divided by their standard errors. The resulting *standardized residuals* should be like observations from a Normal distribution with unit variance (again if the assumptions of the analysis are valid). The *leverage* values indicate how influential each observation is: a large value indicates that the fit of the model depends strongly on that observation; see the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 3.1.1 for more details.

You can display the fit graphically by clicking on the Fitted model button in the Linear Regression Further Output (Figure 1.9). This displays the picture shown in Figure 1.10, which shows the observed data with the fitted line and 95% confidence limits for the line.

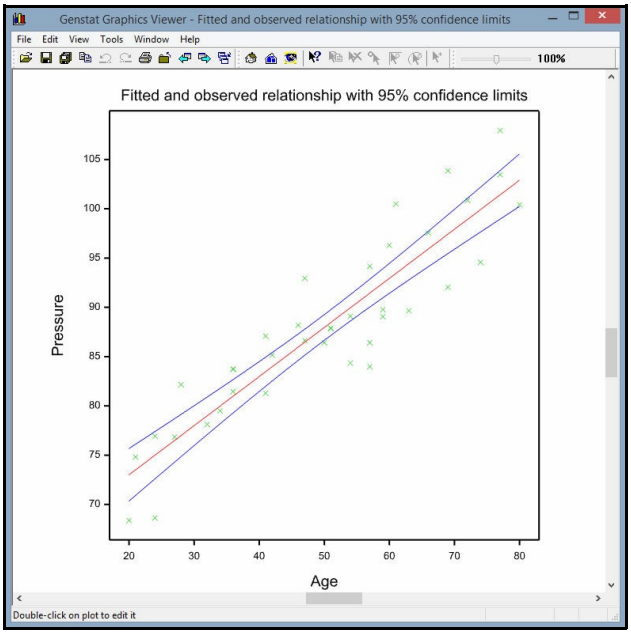
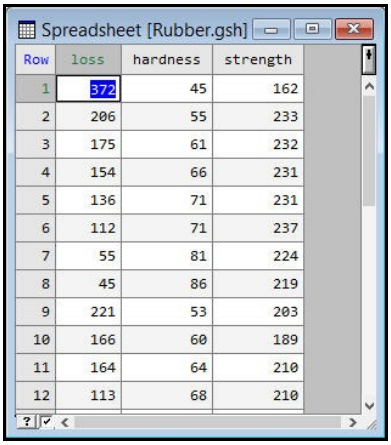


Figure 1.10

1.2 Practical

Spreadsheet file `Rubber.gsh`, contains data from an experiment to study how the resistance of rubber to abrasion is affected by its strength and hardness. The data are from Davies & Goldsmith (1972, *Statistical Methods in Research & Production*, Oliver & Boyd, Edinburgh), and are also used by McConway, Jones & Taylor (1999, *Statistical Modelling using GENSTAT*, Arnold, London, Chapter 4).

Use linear regression to see how loss depends on hardness.


Spreadsheet [Rubber.gsh]

Row	loss	hardness	strength
1	372	45	162
2	206	55	233
3	175	61	232
4	154	66	231
5	136	71	231
6	112	71	237
7	55	81	224
8	45	86	219
9	221	53	203
10	166	60	189
11	164	64	210
12	113	68	210

Figure 1.11

1.3 Checking the assumptions

The efficiency of the estimates in ordinary linear regression and the validity of the statistical tests depends on the assumption that the residuals ε_i come from independent Normal distributions, all with the same variance.

Genstat gives a warning message about any large residuals, as part of the summary of the analysis. The threshold is the value h that gives an upper-tail probability of $1/d$ in a standard Normal distribution, where d is the number of residual degrees of freedom. However, h is set to 2 (instead of any smaller value) when d is less than 20, and to 4 (instead of any larger value) when d is greater than 15773. So messages should appear for extreme outliers, but they should not be set off too often by random variation.

A warning message is also given if there are any particularly large leverage values. The threshold is $h \times k / N$, where k and N are the number of parameters and number of units used in the regression model, and h is as defined above. The sum of the leverages is always k, so this should draw attention to any observations with more than about twice the average influence. This does not mean that assumptions are broken, but rather that the analysis may be unduly affected by some observations.

If there are at least 20 observations, Genstat makes two checks to see if the variance is constant. The fitted values are ordered into three roughly equal-sized groups. Levene tests are then carried out to compare the variance of the standardized residuals in the bottom group with those in the top group, and to compare the variance of the middle group with the variance of the bottom and top groups combined. Each test will generate a message if the test statistic is significant at the 2.5% level, which would indicate that the assumption of constant variance may not be valid.

Finally, Genstat sorts the standardized residuals according to the fitted values, and does "runs" test. A message is given if the sign of successive residuals does not change sufficiently often (again using a 2.5% significance level). This would indicate that there is still some systematic pattern in the residuals.

See the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 3.1.2 for more details.

You can also check the assumptions of the analysis visually, using the Model Checking menu (Figure 1.12) You open the menu by clicking on the Model checking button in the Regression Further Output menu (Figure 1.9). The menu allows you to choose between five types of graph for either the residuals, the leverage values or the *Cook's statistics* (a combination of the residual and leverage information).

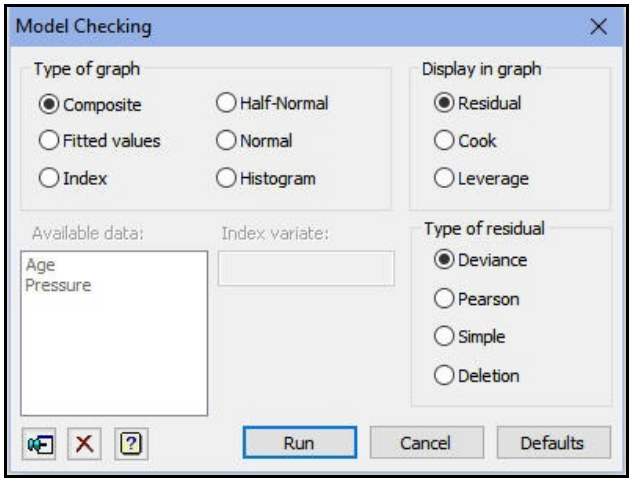


Figure 1.12

Figure 1.12 shows the default, which is a composite of four of these graphs: a histogram of the residuals, so that you can check that the distribution is symmetrical and reasonably Normal; a plot of residuals against fitted values, so that you can check whether the residuals are roughly symmetrically distributed with constant variance; a *Normal plot* which plots the ordered residuals against Normal distribution statistics – if they lie roughly on a straight line, the residuals are roughly Normally distributed; and a *half-Normal plot* which does the same for the absolute values of the residuals, and can be more useful for small sets of data.

The plots in Figure 1.13 indicate that the variance seems unrelated to the size of the observation, but that the distribution seems to be more constrained than the Normal: the largest residuals are a little smaller than would be expected from a Normal distribution. Experience shows the analysis is robust to small departures from Normality. However, we should be cautious in interpreting the F-statistics and t-statistics (which rely on the assumption of Normality), if the histogram looks very non-Normal.

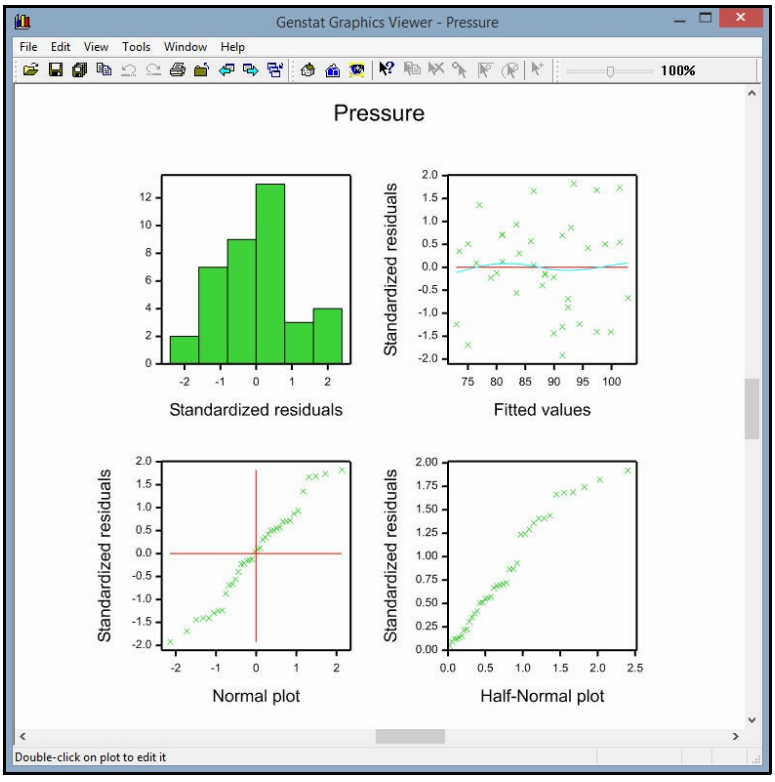


Figure 1.13

With the leverage and Cook statistics, there is no assumption of Normality. So the interest is in how these vary with the fitted values.

Figure 1.14 plots the leverages against the fitted values, showing (as you might expect) that the observations at the lowest and highest ages have most influence in determining the parameter estimates.

Cook's statistics (Figure 1.15) combine residual and leverage information. So they assess whether an observation is both influential and an outlier i.e. whether it is having an unduly large and perhaps detrimental effect on the parameter estimates. You might then need to investigate, for example, to see if some sort of mistake has been made.

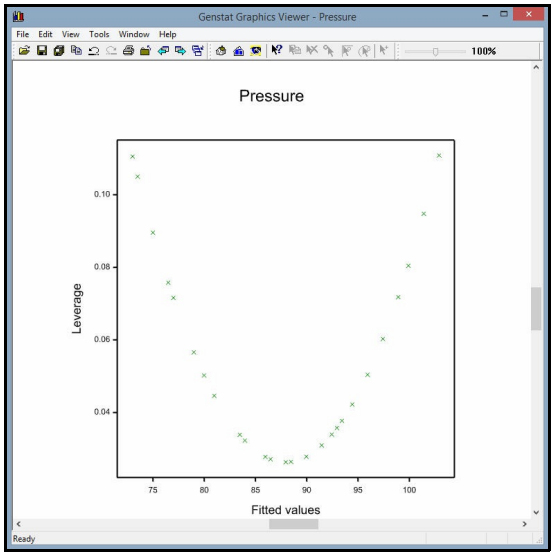


Figure 1.14

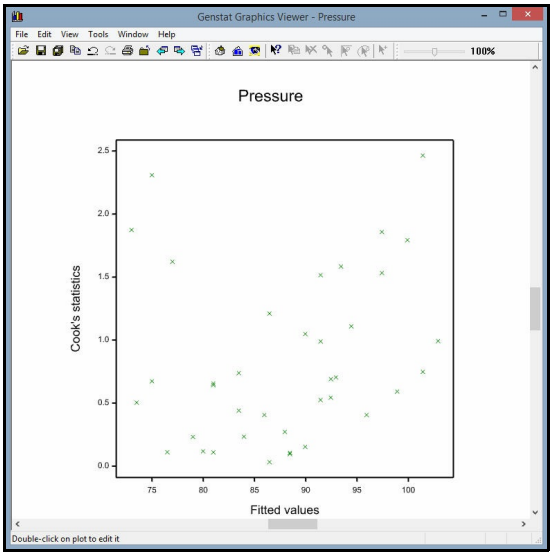


Figure 1.15

1.4 Practical

How well are the assumptions satisfied for the analysis in Practical 1.2?

1.5 Commands for linear regression analysis

The regression menus cover most situations, but it may still be worth learning the commands for their extra control and flexibility. For example Chapter 10 of the *Introduction to Genstat for Windows* explains how you can can write "loops" of commands to perform the same analysis on several data sets.

The menus communicate with the Genstat analysis engine by writing scripts of commands, and these are recorded in the Input Log. You can save these commands and run them later to recreate the analysis. You can also cut and paste the commands into a text window, so that you can edit and rerun them to modify the analysis. Or you can simply examine the commands to see how they work.

The analyses that we have done so far have generated the following set of commands.

```
"Simple Linear Regression"
MODEL Pressure
TERMS Age
FIT [PRINT=model,summary,estimates; CONSTANT=estimate;\
```

```
    FPROB=yes; TPROB=yes] Age
RDISPLAY [PRINT=fittedvalues; TPROB=yes]
RGRAPH [GRAPHICS=high; CIPLOT=yes]
RCHECK [RMETHOD=deviance; GRAPHICS=high] residual; composite
RCHECK [RMETHOD=deviance; GRAPHICS=high] leverage; fitted
RCHECK [RMETHOD=deviance; GRAPHICS=high] cook; fitted
```

The MODEL directive must be used before any regression analysis, to specify the response variate, as in the first line of the program above.

```
MODEL Pressure
```

MODEL can also define the distribution and link function of a generalized linear model (Chapter 3) using its DISTRIBUTION and LINK options, but those are not needed here.

The TERMS command is unnecessary in this example. However, it is useful when you have several explanatory variates or factors and want to examine a sequence of models, adding or dropping terms. It defines the most complicated model that you may want to fit, so that Genstat can construct the overall set of usable units (omitting those that have missing values for any of the variates or factors).

The FIT directive fits the regression.

```
FIT [PRINT=model,summary,estimates; CONSTANT=estimate;\
    FPROB=yes; TPROB=yes] Age
```

The PRINT option controls the output that is produced, so you could ask for all sections of output by setting:

```
PRINT=model,summary,estimates,correlations,fitted,\
      accumulated,confidence
```

Alternatively, after fitting a model you can use the RDISPLAY directive to display further sections of output without refitting the model; it has a PRINT option just like FIT.

The RGRAPH procedure allows you to draw a picture of the fitted model. For example,

```
RGRAPH
```

draws a graph of a simple linear regression. After multiple regression, you can specify the explanatory variate or a grouping factor or both, as in

```
RGRAPH Logsulphur; GROUPS=Rain
```

(see Section 1.15).

The RCHECK procedure provides model checking. It has two parameters: the first specifies what to display in the graph (residuals, Cook or leverages) and the second specifies the type of graph (composite, histogram, fittedvalues, index, normal or halfnormal). For example,

```
RCHECK residual; composite
```

draws the composite picture (Figure 1.13), while the plot of leverages against fitted-values graph (Figure 1.14) can be drawn by

```
RCHECK leverage; fitted
```

The RMETHOD option of RCHECK controls how the residuals are calculated. In an ordinary linear regression, deviance residuals are the ordinary (simple) residuals, divided by their standard errors i.e. they are *standardized* residuals. The GRAPHICS option controls whether the graph is displayed as a high-resolution plot in the graphics viewer (GRAPHICS=high, the default), or whether it is displayed as character plot in the Output window (GRAPHICS=lineprinter).

Full details of the regression commands are in Chapter 3 of the *Guide to the Genstat Command Language, Part 2 Statistics* or in the *Genstat Reference Manual*. The information in the *Reference Manual* is also in the on-line help, and can be accessed easily by putting the cursor within the name of the command (e.g. in the Input Log), and pressing the F1 key.

1.6 Permutation tests

If the assumptions do not seem to be satisfied, an alternative way to assess the significance of the regression might be to to use a permutation test. Clicking on the Permutation test button in the Linear Regression Further Output menu (Figure 1.9) produces the menu in Figure 1.16. This asks Genstat to make 4999 random permutations of the values of the response variate (see the Number of permutations box), and refit the regression. The Seed box specifies the seed to use for the random-number generator that is used to construct the permutations. The value 0 initializes the seed automatically (and prints the value in the output) if this is the first use of the generator in this run of Genstat; otherwise the seed is chosen to continue the existing sequence.

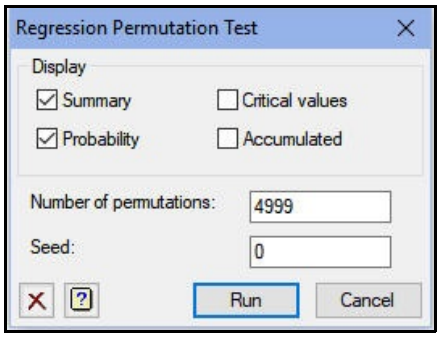


Figure 1.16

The probability for the regression is now determined from its distribution over the randomly permuted data sets. The output below shows a probability <.001, which means that the observed data set is one of the 5 with the largest variance ratios out of the 5000 sets that have been examined (1 observed data set + 4999 randomly permuted data sets).

Message: Default seed for random number generator used with value 909577

Probability for model <.001 (determined from 4999 random permutations)

If you ask for more permutations than the number that are possible for your data, Genstat will instead do an *exact test*, which uses each permutation once. There are $n!$ (n factorial) permutations for a data set with n observations. So, we would obtain an exact test with 5 observations by setting the number of permutations to 120 or more.

The test is performed using the `RPERMTEST` procedure.

1.7 Practical

Do a permutation test for the simple linear regression analysis in Practical 1.2.

1.8 Saving information from the analysis

As well as displaying the results of an analysis, the regression menus allow you to save the results in standard data structures. This is a common feature of most of the analysis menus in Genstat. After a regression analysis you can click on the Save button of the Linear Regression menu (Figure 1.7), which generates the Linear Regression Save Options menu. The residuals, fitted values, parameter estimates and standard errors can all be saved in variates: if you check one of these boxes, you will be prompted for the name of the variate to store the results, as shown in Figure 1.17. The variance-covariance matrix of the parameter estimates can also be saved in a symmetric matrix, another of Genstat's standard data structures. The information is saved using the RKEEP directive.

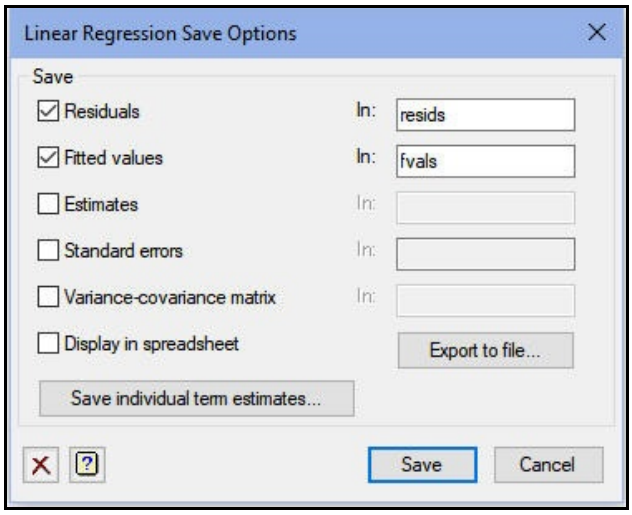


Figure 1.17

If you check the Display in Spreadsheet box, the results are put into a Genstat spreadsheet, which can then be saved in a file on your computer for use in a later run of Genstat, or in another program such as Excel. Alternatively you can save results automatically to a spreadsheet file by clicking on the Export to file button. This opens the Save Regression Results in Spreadsheet File menu. Figure 1.18, shows the menu with the default output components selected in the check boxes, and the Save in file box filled in to save them in the Excel file PressureResults.xlsx.

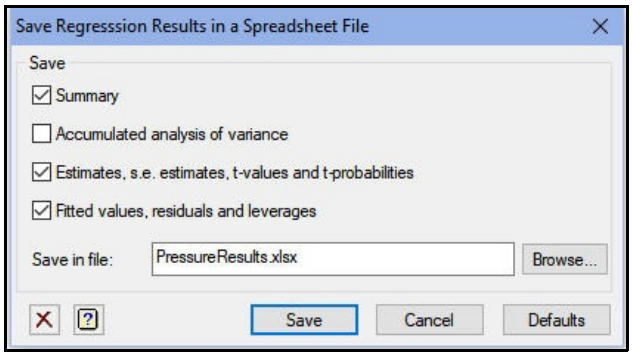


Figure 1.18

Each output component is saved on a separate page in the spreadsheet file. Figure 1.19 shows the page containing the summary of the analysis. Other pages save the estimates (with their standard errors etc.), and the fitted values (with residuals etc.).

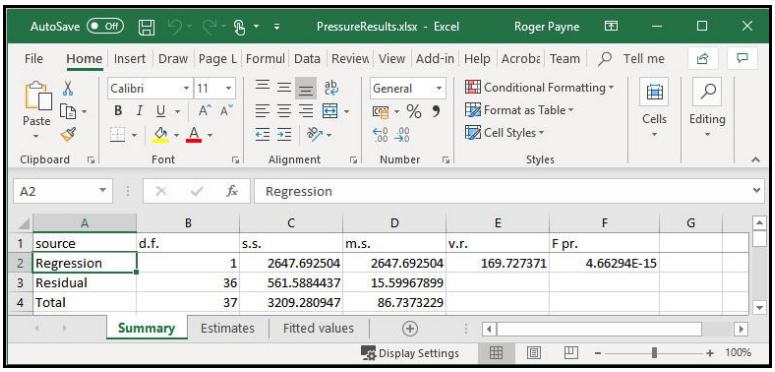


Figure 1.19

The file is saved using the RSPREADSHEET procedure.

1.9 Predictions from linear regression

The fitted values provide predictions of the response variable at the values of the explanatory variable that actually occurred in the data. If you want predictions at other values, you can use the prediction menu, obtained by clicking on the Predict button in the Linear Regression menu. This generates the Predictions - Simple Linear Regression menu shown in Figure 1.20. Initially the Predict values at box has mean filled in, so that a prediction would be formed for pressure at the mean value of the ages. However, we have changed this to ask for predictions at ages 25, 50, 75 and 100. The Display box has boxes that can be checked to provide predictions, standard errors, standard errors of differences between predictions, least significant differences of predictions, confidence limits and a description of how the predictions are formed. Here we print predictions, standard errors and the description.

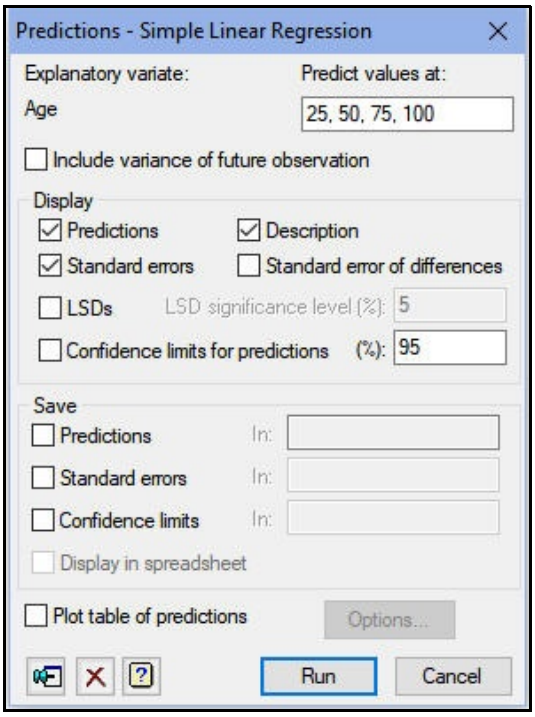


Figure 1.20

Predictions from regression model

These predictions are estimated mean values.

The standard errors are appropriate for interpretation of the predictions as summaries of the data rather than as forecasts of new observations.

Response variate: Pressure

Age	Prediction	s.e.
25	75.50	1.150
50	87.96	0.641
75	100.42	1.152
100	112.87	2.018

The output explains that the standard errors are appropriate as predictions for fitted values for these ages in this data set, not as predictions for new observations. We can augment the standard errors by the additional variability arising from a new set of observations at ages 25 - 100 by checking the box Include variance of future observation. (For further details see Section 3.3.4 of Part 2 of the *Guide to the Genstat Command*

Language.)

The predictions are made using the PREDICT directive.

1.10 Practical

Form predictions from the simple linear regression analysis in Practical 1.2 for hardness values 50, 60, 70, 80 and 90.

1.11 Multiple linear regression

In multiple linear regression you have several explanatory variables. This creates the extra problem, that you need to decide which ones are needed in the model. So you need to be able to explore models, comparing alternative variables or sets of variables, as well as to display and check the model that you finally select.

We illustrate this approach with a short set of data from a production plant, on page 352 of *Applied Regression Analysis* by Draper & Smith (1981, Wiley, New York). Information was collected over 17 months on variables possibly associated with water usage: the average temperature, the amount of production, the number of operating days and the number of employees. The data are loaded from the spreadsheet file Water.gsh.

Linear models with more than one explanatory variable are called *multiple linear regression models*. If you choose this title from the drop-down list in the Linear Regression menu, you can then specify several explanatory variables as well as the single response variable, as shown in Figure 1.21.

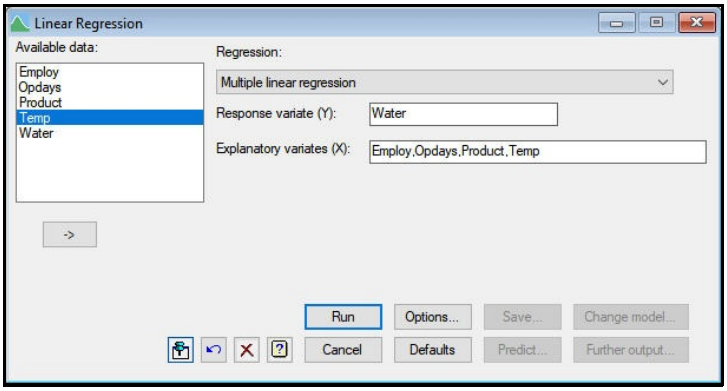


Figure 1.21

However, rather than just fitting the full model in one step, we shall illustrate how you can fit a sequence of regression models. This is best done using the General linear regression option from the drop-down list (Figure 1.22). This allows you to modify the model as many times as you like, using the Change model button in the Linear Regression menu.

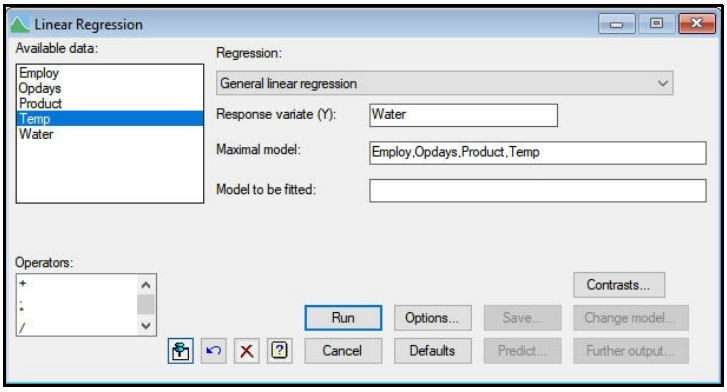


Figure 1.22

It is useful in a sequential study to start by specifying a *maximal model*, which includes all the explanatory terms that may be used in the sequence of models to be fitted. Genstat is then

able to customize the Change Model menu so that the Available data box is replaced by a Terms box containing all the terms that may be fitted. Also, if any explanatory variables have missing values, a common set of units (for which all variables have values) is identified at the start, so that all models can be properly compared. To start with, we leave the Model to be fitted box blank and fit only the constant, as shown in Figure 1.22.

It is important to note a small difference between the model boxes in General linear regression compared to the other types. Here, you can construct model formulae using the operators given in the Operators box: therefore, if you want just a list of explanatory variates, as here, you must type in commas to separate the identifiers. With Multiple linear regression these are added automatically.

Here is the output from this first analysis.

Regression analysis

Response variate: Water
Fitted terms: Constant

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	0	0.000	*		
Residual	16	3.193	0.1995		
Total	16	3.193	0.1995		

Percentage variance accounted for 0.0
Standard error of observations is estimated to be 0.447.

Message: the following units have large standardized residuals.

Unit	Response	Residual
16	4.488	2.73

Estimates of parameters

Parameter	estimate	s.e.	t(16)	t pr.
Constant	3.304	0.108	30.49	<.001

We can build the model using the Change Model menu (Figure 1.21), obtained by returning to the Linear Regression menu and clicking on Change model. This has a Terms window, in which you select the explanatory variables that you want to change. As you click on each one it is highlighted to show that it has been selected. As usual, you can hold down the Ctrl key when you click a line, so that this will not affect the highlighting of the other lines. Or you can click on Select all if you want all of them.

Once you have selected the variables of interest, you can click the Add button to add them to the model. Alternatively, you can click the Drop button to remove them from the model, or click the Switch button to remove those that are in the model and add those that are not. The Try button allows you to assess the effect of switching each of the selected variables, before making any change. There is also a section of the menu for stepwise

regression which is discussed in Section 1.13.

In Figure 1.21, we have selected all the variables, and checked just the Display changes box in the Explore section of the menu. Clicking Try now generates a succinct summary of the effect of each potential change. The first column describes the change. Subsequent columns give the degrees of freedom, sum of squares and mean square of the change. Here we are simply adding single potential x-variates, so the degrees of freedom are all one. Also, the residual of the initial model is printed to indicate the general level of variation. You might want to add terms with large mean squares (or remove terms with small mean squares, if there were any in the model already).

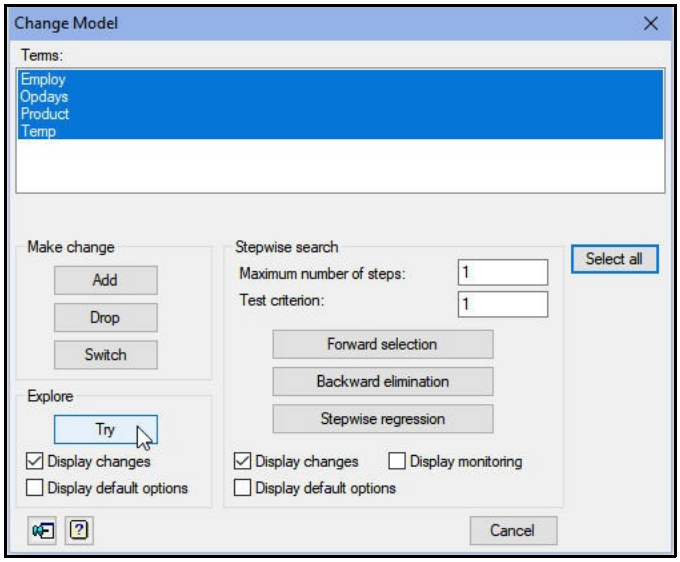


Figure 1.23

Changes investigated by TRY

Change	d.f.	s.s.	m.s.
+ Employ	1	0.545	0.545
+ Opdays	1	0.025	0.025
+ Product	1	1.270	1.270
+ Temp	1	0.261	0.261
Residual of initial model	16	3.193	0.200

Try is useful particularly if you have many explanatory variables and do not wish to fit them all. Here we shall be adding them all to the model, and so we will not use Try again. However, we will take its advice as to which variable to add to the model first. The output shows that `Product` has the largest mean square, so we use the Change model menu to add this (by selecting the Product line, and then clicking Add). The output is given below.

Regression analysis

Response variate: Water
Fitted terms: Constant, Product

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	1.270	1.2702	9.91	0.007
Residual	15	1.922	0.1282		
Total	16	3.193	0.1995		
Change	-1	-1.270	1.2702	9.91	0.007

Percentage variance accounted for 35.8
Standard error of observations is estimated to be 0.358.

Message: the following units have large standardized residuals.

Unit	Response	Residual
16	4.488	2.31

Message: the following units have high leverage.

Unit	Response	Leverage
2	2.828	0.27
3	2.891	0.25

Estimates of parameters

Parameter	estimate	s.e.	t(15)	t pr.
Constant	2.273	0.339	6.71	<.001
Product	0.0799	0.0254	3.15	0.007

The messages in the summary warn about one large residual, and two months with high leverage. So we would have to be careful in interpreting the results if we suspected that these two months were special in some way. Otherwise, the output from this analysis is similar to that in Section 1.1, and it shows that the model here accounts for only 35.8% of the variance in water use.

We can attempt to account for more of the variance by including the effect of another explanatory variable. We shall try the effect of temperature, so the model will become:

$$water = a + b \times production + c \times temperature$$

This can be fitted easily by returning to the Linear Regression menu and clicking on Change model again (Figure 1.21). You can then select `Temp` from the Terms box and click on Add as before to fit the modified model. The output is shown below.

Regression analysis

Response variate: Water
Fitted terms: Constant, Product, Temp

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	1.560	0.7798	6.68	0.009
Residual	14	1.633	0.1167		
Total	16	3.193	0.1995		
Change	-1	-0.289	0.2894	2.48	0.138

Percentage variance accounted for 41.5
Standard error of observations is estimated to be 0.342.

Message: the following units have large standardized residuals.

Unit	Response	Residual
16	4.488	2.04

Estimates of parameters

Parameter	estimate	s.e.	t(14)	t pr.
Constant	1.615	0.528	3.06	0.008
Product	0.0808	0.0242	3.34	0.005
Temp	0.00996	0.00632	1.57	0.138

The `Change` line and the t-statistic for `Temp` tell the same story here: the extra explanatory variable accounts for a further 5.7% of the variance, but does not seem to have a significant effect in conjunction with the amount of production.

We now include the effect of the number of operating days in each month, by adding it via the Change Model menu. To decrease the amount of output, we have clicked on Options first, and cancelled the display of the parameter estimates in the resulting General Linear Regression Options menu, so that we just get the model summary, as shown in Figure 1.24. (Notice that this menu would also allow you to specify a variate of weights if you wanted to do a *weighted* linear regression.) In the output, shown below, the percentage variance accounted for has increased to 50%. So this variable has a marked effect on water usage.

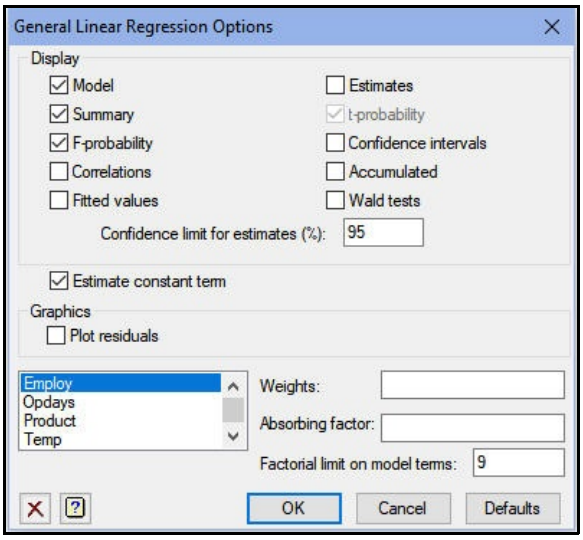


Figure 1.24

Regression analysis

Response variate: Water
Fitted terms: Constant, Product, Temp, Opdays

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	3	1.893	0.63093	6.31	0.007
Residual	13	1.300	0.09999		
Total	16	3.193	0.19954		
Change	-1	-0.333	0.33328	3.33	0.091

Percentage variance accounted for 49.9
Standard error of observations is estimated to be 0.316.

Finally, we add the fourth explanatory variable, the number of employees, returning to the default output.

Regression analysis

Response variate: Water
Fitted terms: Constant, Product, Temp, Opdays, Employ

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	4	2.4488	0.61221	9.88	<.001
Residual	12	0.7438	0.06198		
Total	16	3.1926	0.19954		
Change	-1	-0.5560	0.55603	8.97	0.011

Percentage variance accounted for 68.9
Standard error of observations is estimated to be 0.249.

Message: the following units have high leverage.

Unit	Response	Leverage
1	3.067	0.59

Estimates of parameters

Parameter	estimate	s.e.	t(12)	t pr.
Constant	6.36	1.31	4.84	<.001
Product	0.2117	0.0455	4.65	<.001
Temp	0.01387	0.00516	2.69	0.020
Opdays	-0.1267	0.0480	-2.64	0.022
Employ	-0.02182	0.00728	-3.00	0.011

This variable, too, has a large effect, raising the percentage variance accounted for to 69%.

Notice that the t-statistics now provide evidence of a significant effect of each variable

when all the others are taken account for. The estimate for the Temp parameter is larger than in the model with just production and temperature, 0.01387 compared to 0.00996, and its standard error is smaller, 0.00516 compared to 0.00632. The first effect is caused by the fact that there is correlation, or *confounding*, between the effects of the explanatory variables: so any effect is estimated differently in the presence of a different set of other explanatory variables. The difference in standard errors is caused both by this and by the fact that more variance has been accounted for in the last model.

The effect of this confounding can also be highlighted by looking at an accumulated analysis of variance. This shows the sequential effects of including the variables, in the order in which they were listed, rather than their effects in the presence of all the other variables. This summary is available from the Linear Regression Further Output menu, shown in Figure 1.25, and is displayed below.

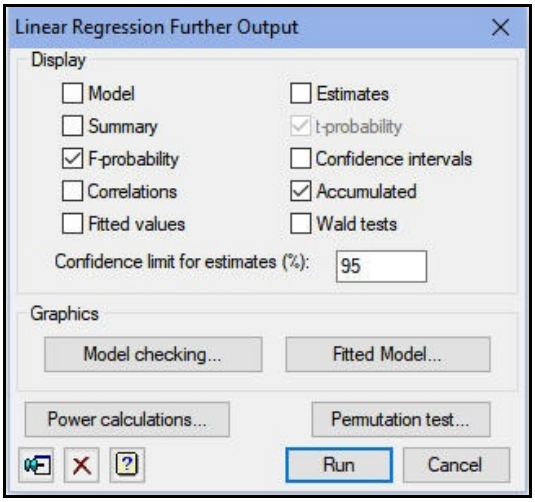


Figure 1.25

Regression analysis

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ Product	1	1.27017	1.27017	20.49	<.001
+ Temp	1	0.28935	0.28935	4.67	0.052
+ Opdays	1	0.33328	0.33328	5.38	0.039
+ Employ	1	0.55603	0.55603	8.97	0.011
Residual	12	0.74380	0.06198		
Total	16	3.19263	0.19954		

The F-probability for Temp here could be used to test the effect of temperature eliminating the effect of Product but ignoring Opdays and Employ; the t-probability with the estimate of Temp above, tests the effect eliminating the effects of all the other explanatory variables.

In this section, we have fitted the model sequentially, starting with just the constant, and then using the Change Model menu to decide which terms to add into the model. (In this example, the terms are x-variates, but you will see later, in Section 1.15, that regression models can also include factors and interactions with factors.) Provided you do not have too many terms, an alternative strategy would be to include them all, and then see sequentially whether any one can be left out.

If you do have only variates in the model, you can use the t-statistics of their regression

coefficients to assess whether they are needed. However, if you have factors, these may contribute several parameters to the model, making the assessment more difficult. Wald statistics (available from either the Options or Further Output menus) can then be used instead, to assess whether any term can be dropped from the model. The output below shows Wald statistics for the final model fitted to the water data. In an ordinary linear regression, Genstat also prints an F statistic (calculated as the Wald statistic divided by its degrees of freedom), and uses this to obtain the probability for each term. Provided there is no aliasing between the parameters of the terms, these F statistics and probabilities will be identical to those that would be printed in the Change lines of the Summary of Analysis if the terms were dropped from the model explicitly by using the Change Model menu. The advantage of the Wald statistics is that the model does not have to be refitted (excluding each term) to calculate the information. They thus provide a more efficient method of assessing whether all the terms are needed in the model.

Wald tests for dropping terms

Term	Wald statistic	d.f.	F statistic	F pr.
Product	21.61	1	21.61	<0.001
Temp	7.22	1	7.22	0.020
Opdays	6.96	1	6.96	0.022
Employ	8.97	1	8.97	0.011

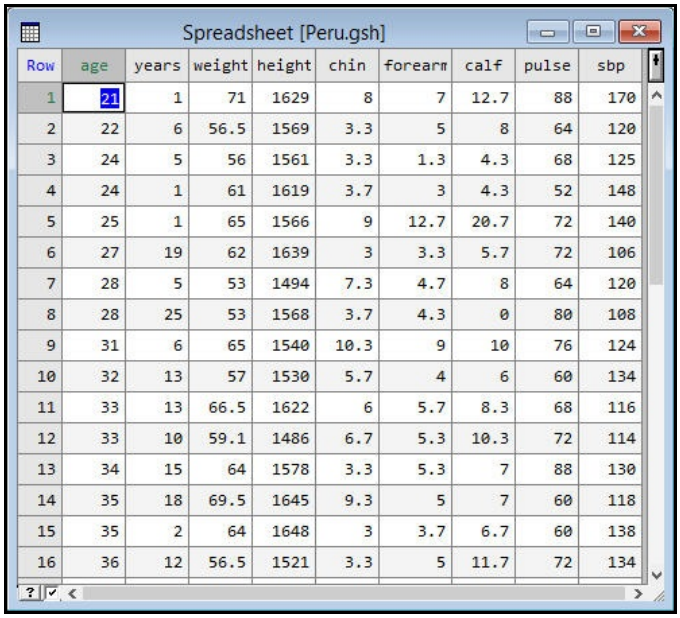
Residual d.f. 12

To perform a stepwise regression using commands, you first define the response variate, using the MODEL directive, in the usual way. You should also use the TERMS command to define the most complicated model that you may want to fit, so that Genstat can initialize the analysis e.g. by constructing the overall set of usable units (omitting those that have missing values for any of the variates or factors). If you do not do this and the explanatory variables do not all have the same units, the accumulated summary of the analysis may need to reinitialize itself part-way through the sequence of models. The first model is fitted using the FIT directive as usual. This can then be modified using the directives ADD, DROP, STEP, SWITCH and TRY. See the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 3.2. Wald statistics are calculated by the RWALD procedure.

1.12 Practical

Spreadsheet file Peru.gsh, contains a data set recording blood pressure and physical characteristics of some Peruvian indians (see McConway, Jones & Taylor 1999, *Statistical Modelling using GENSTAT*, Arnold, London, Section 6.2). The aim is to see whether blood pressure, sbp, can be explained effectively by regression models involving the physical variables. Use the Change Model menu to build a model containing up to two variables.

Can that model be improved by adding further variables?



Row	age	years	weight	height	chin	forearm	calf	pulse	sbp
1	21	1	71	1629	8	7	12.7	88	170
2	22	6	56.5	1569	3.3	5	8	64	120
3	24	5	56	1561	3.3	1.3	4.3	68	125
4	24	1	61	1619	3.7	3	4.3	52	148
5	25	1	65	1566	9	12.7	20.7	72	140
6	27	19	62	1639	3	3.3	5.7	72	106
7	28	5	53	1494	7.3	4.7	8	64	120
8	28	25	53	1568	3.7	4.3	0	80	108
9	31	6	65	1540	10.3	9	10	76	124
10	32	13	57	1530	5.7	4	6	60	134
11	33	13	66.5	1622	6	5.7	8.3	68	116
12	33	10	59.1	1486	6.7	5.3	10.3	72	114
13	34	15	64	1578	3.3	5.3	7	88	130
14	35	18	69.5	1645	9.3	5	7	60	118
15	35	2	64	1648	3	3.7	6.7	60	138
16	36	12	56.5	1521	3.3	5	11.7	72	134

Figure 1.26

1.13 Stepwise and all subsets regression

The sequential fitting methods described in Section 1.11 can be very labour intensive if there are many variables. The Change Model menu (Figure 1.27) also provides stepwise facilities that allow you to build up the model automatically.

To illustrate these with the water usage data, we first fit a model with just the constant (using the menu in Figure 1.22 in Section 1.11), and then click the Change button to produce the Change Model menu as before.

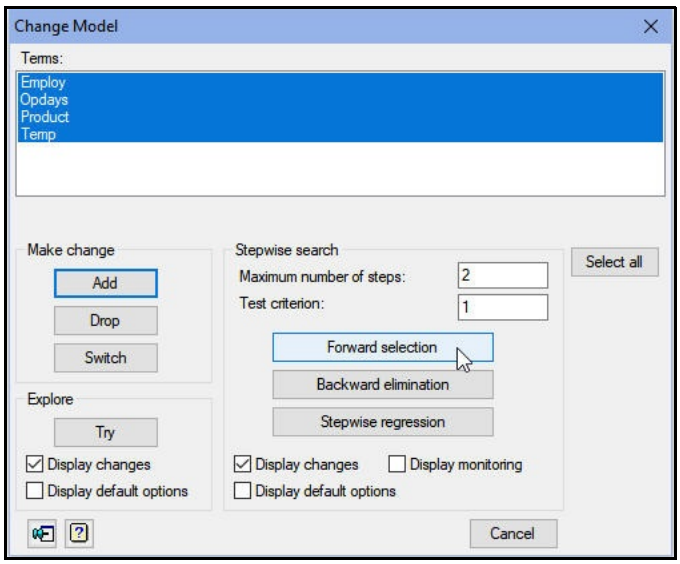


Figure 1.27

The process takes the form of a number of steps (specified in the Maximum number of steps box) in which variables are added or dropped from the model. The possible changes to consider are selected in the Terms box; in Figure 1.27 we have decided to consider all the variables. Each possible change is assessed using a variance ratio calculated as the mean square of the change line divided by the residual mean square of the original model.

If you click the Forward selection button, at each step Genstat adds the variable with the largest variance ratio, provided that variance ratio exceeds the value specified in the Test criterion box. The default value for the criterion is one, but many users prefer the value

four; see for example page 153 of McConway, Jones & Taylor (1999, *Statistical Modelling using GENSTAT*, Arnold, London).

If we click on Forward selection in Figure 1.27, two steps are taken, adding first `Product` and then `Employ`, as shown below.

Step 1: Residual mean squares

0.1282	Adding	Product
0.1765	Adding	Employ
0.1955	Adding	Temp
0.1995	No change	
0.2112	Adding	Opdays

Chosen action: adding Product.

Step 2: Residual mean squares

0.09710	Adding	Employ
0.11665	Adding	Temp
0.12816	No change	
0.13174	Adding	Opdays
0.19954	Dropping	Product

Chosen action: adding Employ.

As only the Display changes box is checked in the menu, Genstat simply produces a brief summary of the changes. The residual mean square of the original model at each step is given in the "No change" line. Notice that, for information, Genstat also shows the effect of dropping terms.

Thus, if you set the maximum number of steps equal to the number of variables, Genstat will perform a complete forward stepwise fit automatically, stopping only when no further variable seems to be useful.

The Backward elimination button examines the effect of dropping variables from the model. Suppose we now select Employ and Product in the Change Model menu (Figure 1.28), and click on Backward elimination. At each step, Genstat now drops the term with the smallest variance ratio, provided that variance ratio is less than the test criterion. As the output below shows, both variance ratios are greater than the criterion, so the process stops after a single step.

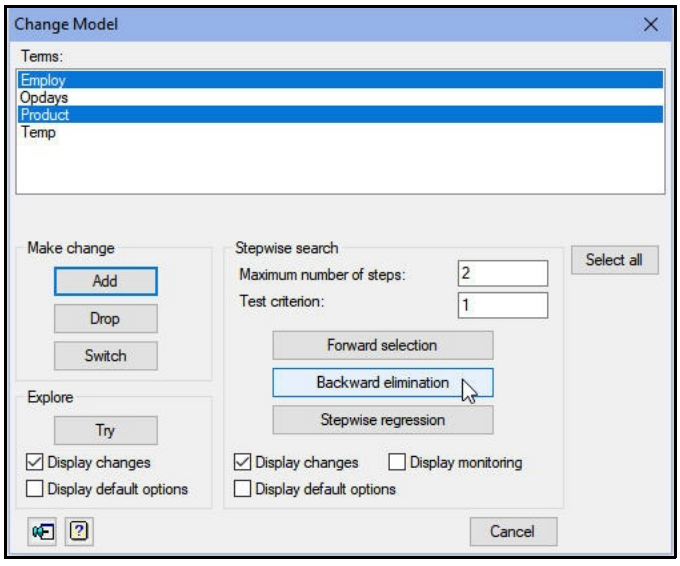


Figure 1.28

Step 1: Residual mean squares

```
0.09710   No change
0.12816   Dropping  Employ
0.17649   Dropping  Product

Chosen action: no change.
```

The menu can thus be used for full automatic backwards stepwise regression by first fitting the full model with the General Linear Regression menu (Figure 1.22). Then select all the variables in the Change Model menu, set a maximum number of steps equal to the number of variables and click on Backward Elimination.

Finally, if you click the Stepwise Regression button, Genstat will first look to see if any variable can be dropped. Then, if that is not possible, it looks to see if any can be added.

Automatic stepwise procedures result in only one model, and alternative models with an equivalent or even better fit can easily be overlooked. In observational studies with many correlated variables, there can be many alternative models, and selection of just one well-fitting model may be unsatisfactory and perhaps misleading. Another method is to fit all possible regression models, and to evaluate these according to some criterion. In this way several best regression models can be selected. However the fitting of all possible regression models can be very time-consuming. It should also be used with caution, because models can be selected that appear to have a lot of explanatory power, but contain only noise variables (those representing random variation). This can occur particularly when the number of parameters is large in comparison to the number of units. The models should therefore not be selected on the basis of a statistical analysis alone, but by considering the physical plausibility of models and by taking account of any previous modelling experience.

All subsets regression can be performed using the All Subsets Regression menu. This is obtained by selecting Regression Analysis from the Stats menu, clicking on All Subsets Regression and then Linear Models (as we shall be investigating a linear regression model again), as shown in Figure 1.29.

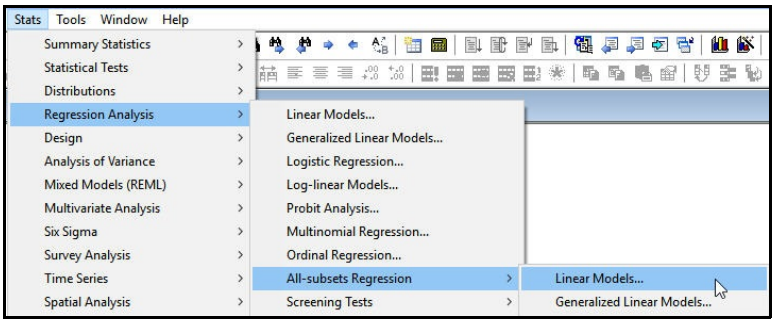


Figure 1.29

Figure 1.30 shows the menu set up to examine all possible regression models for the water usage data. Water is entered as the response variate, the explanatory variates are listed (separated by commas) in the Model formula or list of explanatory data box, and the All possible box is checked.

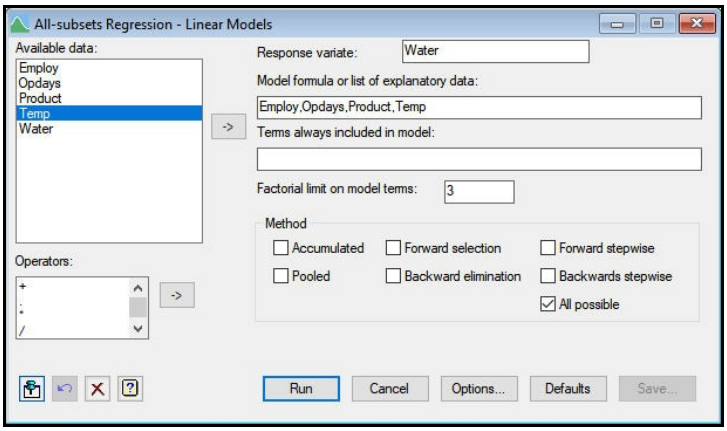


Figure 1.30

The output provides a brief summary of all the regressions. By default, the models with each number of explanatory variables are ordered according to their percentage variances accounted for (the column header "Adjusted"), and a statistic known as Mallows C_p is provided for further information. C_p is rather more conservative than the percentage variance accounted for (see Section 3.2.6 of the *Guide to the Genstat Command Language, Part 2 Statistics*) but here they lead to the same conclusions. Other statistics can be selected using the All Subsets Regression Options menu (obtained by clicking the Options button as usual). This also allows you to set a limit on the total number of terms in the subsets. (It may be impracticable to fit them all if there are many variables.)

Model selection

Response variate: Water
Number of units: 17
Forced terms: Constant
Forced df: 1
Free terms: Employ + Opdays + Product + Temp

All possible subset selection

Message: probabilities are based on F-statistics, i.e. on variance ratios.

Best subsets with 1 term

Adjusted	Cp	Df	Employ	Opdays	Product	Temp
35.77	18.02	2	-	-	.007	-
11.55	29.71	2	.099	-	-	-
2.04	34.30	2	-	-	-	.266
<0.00	38.10	2	-	.735	-	-

Best subsets with 2 terms

Adjusted	Cp	Df	Employ	Opdays	Product	Temp
51.34	10.93	3	.030	-	.003	-
41.54	15.35	3	-	-	.005	.138
33.98	18.76	3	-	.454	.007	-
16.99	26.41	3	.075	-	-	.181
6.42	31.18	3	.107	.679	-	-
1.51	33.39	3	-	.354	-	.168

Best subsets with 3 terms

Adjusted	Cp	Df	Employ	Opdays	Product	Temp
54.70	9.96	4	.042	-	.004	.177
54.06	10.22	4	.019	.199	.002	-
49.89	11.97	4	-	.091	.002	.036
19.70	24.61	4	.062	.247	-	.092

Best subsets with 4 terms

Adjusted	Cp	Df	Employ	Opdays	Product	Temp
68.94	5.00	5	.011	.022	.001	.020

The output shows that the best model with a single explanatory variable is the one with production (confirming the conclusion from our use of Try in Section 1.11), the best with two variables has production and number of employees, and so on.

The menu also provides some rather more flexible and powerful stepwise regression facilities which we will not demonstrate. For details see the on-line help or Section 3.2.6 of the *Guide to the Genstat Command Language, Part 2, Statistics*, which describes the RSEARCH procedure that the menu uses.

1.14 Practical

Use all subsets regression to see whether you can find alternative or improved models to the model that you fitted to the data on blood pressure of Peruvian indians in Practical 1.12.

1.15 Regression with grouped data

This section looks at the types of model that you can fit when you have factors as well as variates in the set of explanatory variables. Suppose you have one explanatory factor and one explanatory variate. You may then want to see how the regression line for the explanatory variate is the same within all the groups defined by the factor. Or perhaps the slope is the same for all the groups but the intercepts differ. Or perhaps the lines have different slopes and different intercepts.

We illustrate these ideas using some data collected in 1990 to investigate changing levels of air pollution. The response variable that we want to model is the logarithm of amount of sulphur in the air each day. We choose the logarithm because it seems natural to expect effects on sulphur to be proportionate to the amount of sulphur present. Also, previous experience of the data set (see the *Introduction to Genstat for Windows*, Chapter 2) shows that the measurements are skewed to the right. Our explanatory variables are a variate `Windsp` recording the strength of the wind, and a factor `Rain` indicating whether or not it rained. The data are available in the spreadsheet file Sulphur.gsh (Figure 1.31) and can be read using the Example Data Sets menu as shown in Section 1.1.

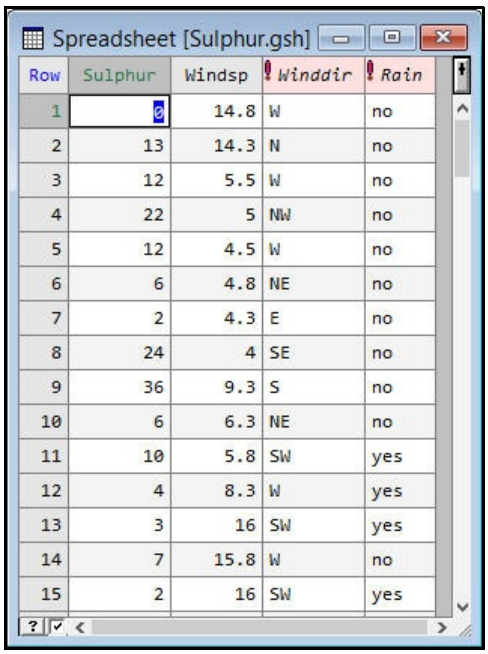


Figure 1.31

To transform the sulphur values, we select the Calculations option of the Data menu on the menu bar to open the Calculate menu. Figure 1.32 shows the menu with the necessary fields filled in to do the calculation and save the results in a new variate, `LogSulphur`. You will see that we get a missing value (and warning) for unit 1, which contains zero. (For further details about the Calculate menu see Chapter 2 of the *Introduction to Genstat for Windows*.)

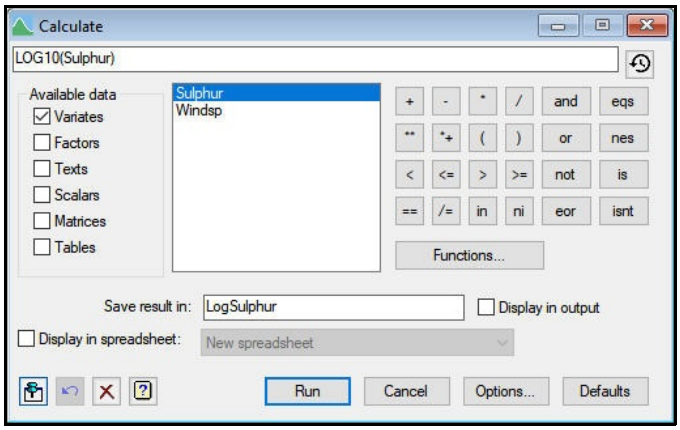


Figure 1.32

First we fit a simple linear regression on the wind speed using the Linear Regression menu, as shown in Figure 1.7.

Regression analysis

Response variate: LogSulphur
Fitted terms: Constant, Windsp

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	1.50	1.4952	10.35	0.002
Residual	110	15.89	0.1445		
Total	111	17.39	0.1567		

Percentage variance accounted for 7.8
Standard error of observations is estimated to be 0.380.

Message: the following units have large standardized residuals.

Unit	Response	Residual
98	1.633	2.68

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.076
72	0.699	0.052
95	1.146	0.055
100	1.398	0.051

Estimates of parameters

Parameter	estimate	s.e.	t(110)	t pr.
Constant	1.1066	0.0892	12.41	<.001
Windsp	-0.02557	0.00795	-3.22	0.002

The decrease in sulphur measurements with wind speed is estimated to be about 5.7% per km/h (the antilog of −0.02557 is 94.3%), and is statistically significant.

We would also like to estimate the difference between wet and dry days, and see if the relationship between sulphur and wind speed is different in the two categories. We can investigate this by selecting Simple linear regression with groups from the drop-down list in the Linear Regression menu. This customizes the menu to include an extra box where you can specify a factor to define the groups; the filled-in box is shown in Figure 1.31, with the factor `Rain` entered as the grouping factor.

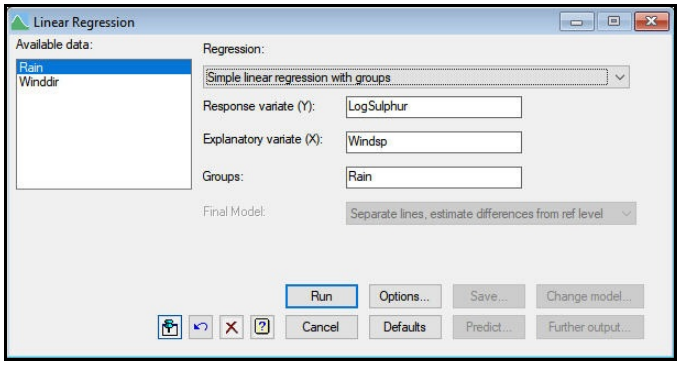


Figure 1.33

The menu performs three successive analyses. The first is exactly the same as that produced already with the Simple linear regression option, so we did not need to do that analysis separately. The second analysis fits a model with a separate intercept for wet and dry days, as shown below.

Regression analysis

Response variate: LogSulphur
Fitted terms: Constant + Windsp + Rain

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	1.89	0.9442	6.64	0.002
Residual	109	15.50	0.1422		
Total	111	17.39	0.1567		
Change	-1	-0.39	0.3933	2.77	0.099

Percentage variance accounted for 9.2
Standard error of observations is estimated to be 0.377.

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.102
72	0.699	0.073

Estimates of parameters

Parameter	estimate	s.e.	t(109)	t pr.
Constant	1.1235	0.0891	12.62	<.001
Windsp	-0.02193	0.00818	-2.68	0.008
Rain yes	-0.1240	0.0745	-1.66	0.099

Parameters for factors are differences compared with the reference level:

Factor	Reference level
Rain	no

The effect of rainfall is quantified here in terms of the difference between dry and wet days: that is, by comparing level `yes` of the factor `Rain` to its *reference level* `no`. By default the reference level is the first level of the factor, but you can change that by selecting the Attributes/Format sub-option of the Column option of the Spread menu on the menu bar. This opens the Column Attributes/Format menu, which has a section where you can choose the reference level for a factor column. Alternatively, you can use the `REFERENCELEVEL` option of the `FACTOR` directive.

So the model is

$$Logsulphur = a + b \times Windsp$$

for dry days, and

$$Logsulphur = a + d + b \times Windsp$$

for wet days. The model thus consists of two parallel regression lines. The estimates show that rainfall decreases the sulphur on average by 25% (antilog(–0.1240) = 75%), but this effect is not statistically significant because of the large unexplained variation in the sulphur measurements. This version of the model is very convenient if you want to make comparisons with the reference level (which may, for example, represent a standard set of conditions or treatment). However, we show later in this section how you can obtain the alternative version with a parameter in the model for each intercept.

We can investigate whether the linear effect of wind speed is different in the two categories of rainfall by looking at the third and final analysis.

Regression analysis

Response variate: LogSulphur
Fitted terms: Constant + Windsp + Rain + Windsp.Rain

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	3	1.92	0.6402	4.47	0.005
Residual	108	15.47	0.1432		
Total	111	17.39	0.1567		
Change	-1	-0.03	0.0323	0.23	0.636

Percentage variance accounted for 8.6
Standard error of observations is estimated to be 0.378.

Message: the following units have large standardized residuals.

Unit	Response	Residual
98	1.633	2.61

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.160
72	0.699	0.112
95	1.146	0.111
104	1.580	0.093

Estimates of parameters

Parameter	estimate	s.e.	t(108)	t pr.
Constant	1.153	0.109	10.57	<.001
Windsp	-0.0252	0.0107	-2.36	0.020
Rain yes	-0.208	0.193	-1.08	0.283
Windsp.Rain yes	0.0079	0.0167	0.47	0.636

Parameters for factors are differences compared with the reference level:

Factor	Reference level
Rain	no

This model includes the *interaction* between the explanatory factor and variate. In Genstat, interactions are represented using the dot operator, so that `Windsp.Rain` represents the interaction between wind speed and rain (i.e. a model term to fit a different regression coefficient from wind speed for each level of rain). The output now shows the slope of the regression for dry days, titled `Windsp`, and the difference in slopes between wet and dry, titled `Windsp.Rain yes`. So again we can see immediately that the difference between the slopes is small and not significant. The graph of the fitted model is shown in Figure 1.34.

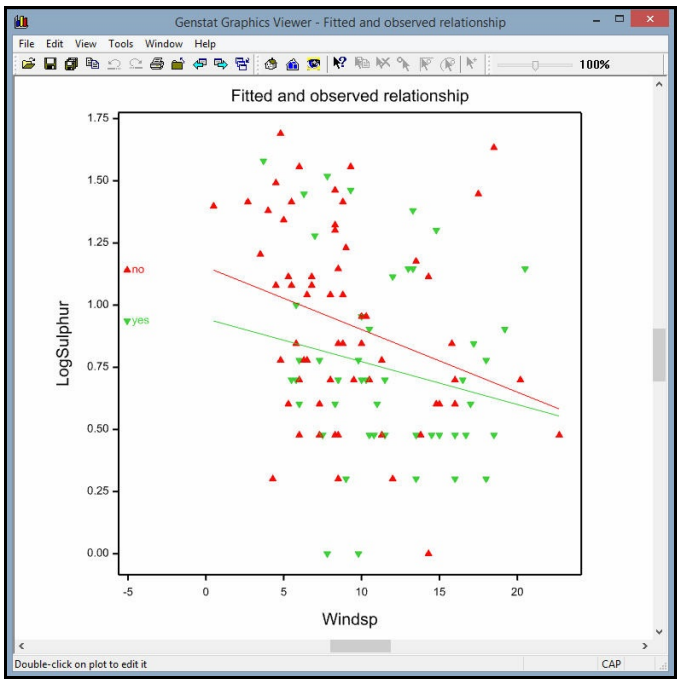


Figure 1.34

An *analysis of parallelism* can be carried out using the Accumulated option of the Linear Regression Further Output menu, as shown in Figure 1.9. This allows you to make a formal assessment of how complicated a model you need. You can then select the appropriate model from the Final model box (see Figure 1.33) and click on Run to fit it.

Regression analysis

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ Windsp	1	1.4952	1.4952	10.44	0.002
+ Rain	1	0.3933	0.3933	2.75	0.100
+ Windsp.Rain	1	0.0323	0.0323	0.23	0.636
Residual	108	15.4677	0.1432		
Total	111	17.3884	0.1567		

Here a Common line (in fact, a simple linear regression) would be enough, but to illustrate the fitted parallel lines we have selected Parallel lines, estimate lines and clicked on Run. This fits parallel lines but now with a parameter for each intercept, rather than parameters for differences from the reference level (which would be given by the alternative setting Parallel lines, estimate differences from ref. level). The other settings are: Common line; Parallel lines, estimate lines; and Parallel lines, estimate differences from ref. level. The fitted parallel lines are shown in Figure 1.35.

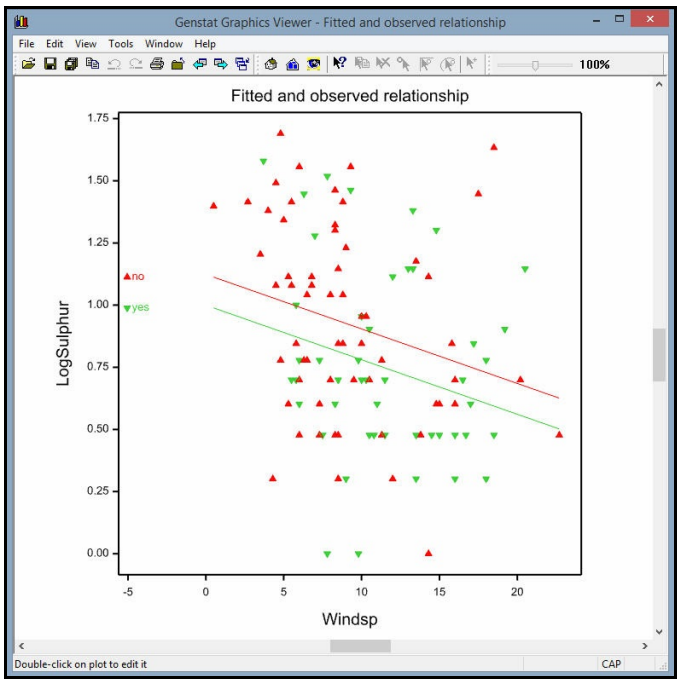


Figure 1.35

Regression analysis

Response variate: LogSulphur
Fitted terms: Windsp + Rain

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	1.89	0.9442	6.64	0.002
Residual	109	15.50	0.1422		
Total	111	17.39	0.1567		
Change	-1	-0.39	0.3933	2.77	0.099

Percentage variance accounted for 9.2
Standard error of observations is estimated to be 0.377.

Message: the following units have high leverage.

Unit	Response	Leverage
30	0.477	0.102
72	0.699	0.073

Estimates of parameters

Parameter	estimate	s.e.	t(109)	t pr.
Windsp	-0.02193	0.00818	-2.68	0.008
Rain no	1.1235	0.0891	12.62	<.001
Rain yes	1.000	0.109	9.14	<.001

1.16 Predictions from regression with groups

If we now click on the Predict button in the Linear Regression menu (Figure 1.33), we can obtain predictions from this parallel-line model. The predictions menu (Figure 1.36) is now customized to include the grouping factor (Rain).

In Figure 1.36, the drop-down list box Predict at levels is set to all, to indicate that we want to form predictions for all the levels of Rain. The alternative setting, standardize, forms averages over the levels of Rain, and the Standardization method box then allows you to indicate whether you want ordinary averages (Equal), or whether you want the levels weighted according to their replication in the data set (Marginal), or whether you want to specify your own weights (Specify) which might correspond to the numbers of wet and dry days that you would anticipate in some future period.

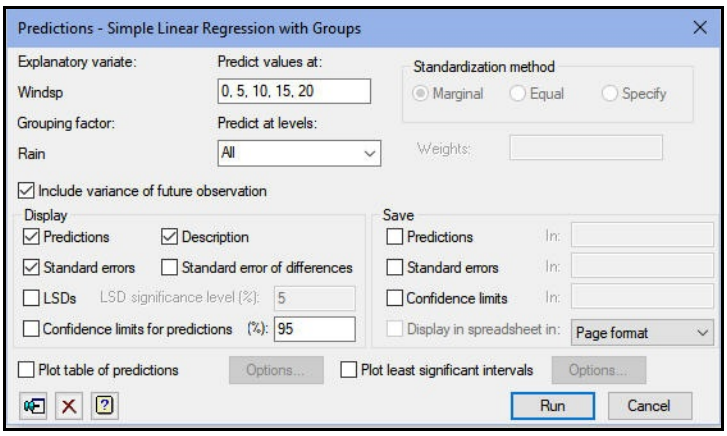


Figure 1.36

The other box specifies the values of the explanatory variate (Windsp) for which we want predictions, here 0, 5, 10, 15 and 20. We have also checked the box to include variance of future observation (unlike Figure 1.20 in Section 1.9), so the standard errors in the output below are relevant for the values as predictions of the amounts of sulphur on future occasions.

Predictions from regression model

These predictions are estimated mean values.

The predictions have been formed only for those combinations of factor levels for which means can be estimated without involving aliased parameters.

The standard errors are appropriate for interpretation of the predictions as forecasts of new observations rather than as summaries of the data.

Response variate: LogSulphur

Rain	no		yes	
	Prediction	s.e.	Prediction	s.e.
Windsp				
0	1.1235	0.3875	0.9996	0.3926
5	1.0138	0.3816	0.8899	0.3848
10	0.9042	0.3801	0.7802	0.3812
15	0.7945	0.3830	0.6705	0.3819
20	0.6848	0.3902	0.5609	0.3870

1.17 Practical

Spreadsheet file `Calcium.gsh`, contains data from an investigation to study associations between the environment and mortality. It records the annual mortality rate per 100000 for males, averaged over the years 1958-1964, and the calcium concentration (parts per million) in the drinking water supply in 61 large towns in England and Wales (see McConway, Jones & Taylor (1999, *Statistical Modelling using GENSTAT*, Arnold, London, Chapter 4).

Use linear regression with groups to investigate whether the relationship between mortality and calcium differs between regions.

Spreadsheet [Calcium.gsh]

Row	town	mortality	calcium	region
1	Bath	1247	105	South
2	Birkenhead	1668	17	North
3	Birmingham	1466	5	South
4	Blackburn	1800	14	North
5	Blackpool	1609	18	North
6	Bolton	1558	10	North
7	Bootle	1807	15	North
8	Bournemouth	1299	78	South
9	Bradford	1637	10	North
10	Brighton	1359	84	South
11	Bristol	1392	73	South
12	Burnley	1755	12	North
13	Coventry	1307	78	South
14	Croydon	1254	96	South
15	Darlington	1491	20	North
16	Derby	1555	39	North
17	Doncaster	1428	39	North
18	East Ham	1318	122	South
19	Exeter	1260	21	South
20	Gateshead	1723	44	North

Figure 1.37

2 Nonlinear regression

In this chapter you will learn

- how to fit polynomials ★
- how to fit smoothing splines ★
- how to fit a standard curve, using a negative exponential curve as an example
- what other standard curves are available
- how to fit parallel and non-parallel standard curves ★
- how to define and fit your own nonlinear models ★

Note: the topics marked ★ are optional.

2.1 Polynomials

In this section we show how to fit polynomial models in Genstat, using data from an experiment to study the relationship between yields of sugar cane and amounts of a nitrogen fertilizer. The data, in spreadsheet file `Cane.gsh` (Figure 2.1), consist of yields of sugar from four replicates of each of five amounts of the fertilizer.

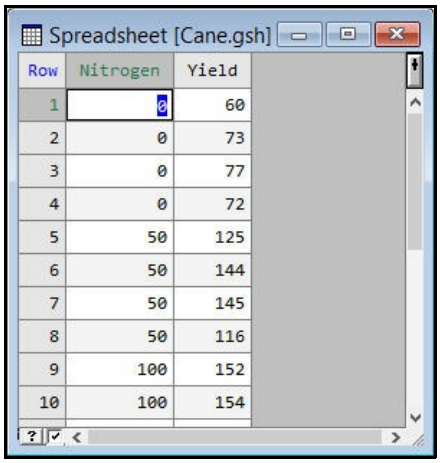


Figure 2.1

To illustrate polynomial regression we shall fit the *quadratic polynomial*

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

In this equation, y is the yield of sugar cane, and x is the corresponding amount of fertilizer. Notice that the model is still linear in the parameters a, b and c, even though there is no longer a linear relationship between y and x. So we can use the Linear Regression menu, as in Chapter 1.

In the Linear Regression menu (Figure 2.2), we select Polynomial regression in the drop-down list at the top of the menu, and choose quadratic as the model. We can then specify `Yield` as the Response variate, `Nitrogen` as the Explanatory variate, and click on Run to fit the model.

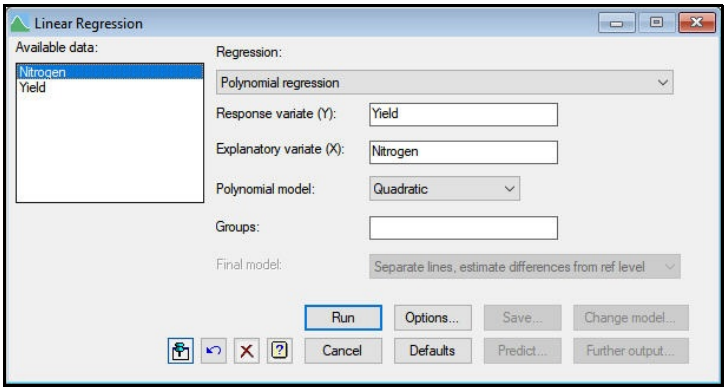


Figure 2.2

Regression analysis

Response variate: Yield
Fitted terms: Constant + Nitrogen
Submodels: POL(Nitrogen; 2)

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	34798.	17398.9	156.90	<.001
Residual	17	1885.	110.9		
Total	19	36683.	1930.7		

Percentage variance accounted for 94.3
Standard error of observations is estimated to be 10.5.

Message: the following units have large standardized residuals.

Unit	Response	Residual
6	144.0	2.10
7	145.0	2.20

Estimates of parameters

Parameter	estimate	s.e.	t(17)	t pr.
Constant	74.19	4.96	14.97	<.001
Nitrogen Lin	1.112	0.117	9.47	<.001
Nitrogen Quad	-0.002721	0.000563	-4.83	<.001

There is a message in the output about two large residuals: Genstat automatically checks to see if any residuals are large compared to a standard Normal distribution (see Section 3.1.2 of Part 2 of the *Guide to the Genstat Command Language* for the exact criterion). However, these two are only just outside the range (–1.96, 1.96) which contains 95% of observations from a Normally distributed variable.

The parameter estimates indicate that the fitted curve has the equation:

$$Yield = 74.19 + 1.112 \times Nitrogen - 0.002721 \times Nitrogen^2$$

The Polynomial regression option uses the Genstat `POL` function. (This is shown in the model description at the start of the output above, which indicates that a *submodel* `POL(Nitrogen;2)` has been fitted.) The `POL` function is also available in the Operators box if you select General linear regression as the regression type. So you can include polynomials in more complicated regressions like those in Section 1.11. The `POL` function (and this menu) will allow models with up to the fourth power. If you want to use higher powers, you would need to fit *orthogonal polynomials* using the `REG` function (see Section 3.4.2 of Part 2 of the *Guide to the Genstat Command Language*).

You can display the fitted model by clicking on the Fitted model button of the Regression Further Output menu as before. The resulting picture, in Figure 2.3, shows the segment of the quadratic curve that has been fitted.

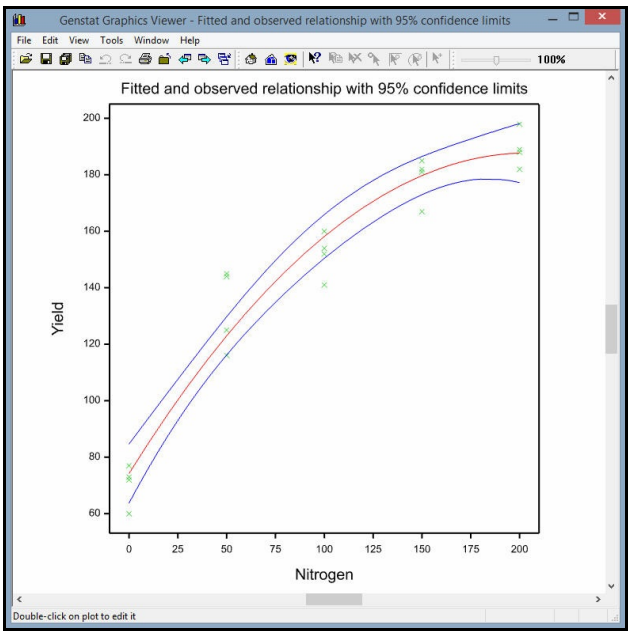


Figure 2.3

The polynomial model that we have fitted above provides a good way of checking for curvature in the relationship between yield and nitrogen. However, it may be unrealistic from a scientific point of view. The shape of the curve is constrained in two important ways: it is a quadratic that must be symmetrical about the maximum, and the curvature changes in a fixed way. As there is no scientific reason

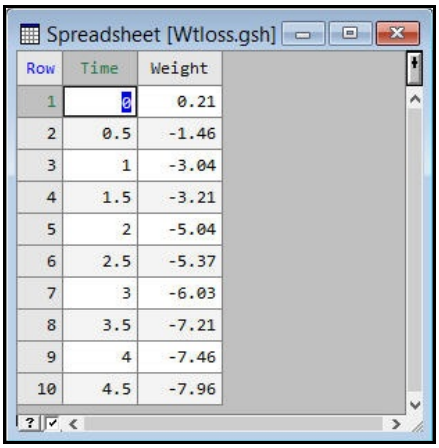
why this shape should fit well, we should be cautious in using it to summarize the results. We should certainly beware of trying to extrapolate outside the range of the data: for larger amounts of fertilizer nitrogen, the model will predict falling yields, and there is no evidence of any fall in yields here. In fact, the model predicts negative yields for nitrogen values greater than 467.1!

2.2 Practical

Spreadsheet file `Wtloss.gsh`, contains data giving the loss in weight of a product following manufacturing (data from Draper & Smith 1981, *Applied Regression Analysis*, Wiley, New York).

Fit a quadratic polynomial of weight on time, examine the residuals, and form predictions for times 0, 5, 10 and 15.

Remove the quadratic term, and plot the residuals against fitted values to see the effect of omitting this term.



Row	Time	Weight
1	0	0.21
2	0.5	-1.46
3	1	-3.04
4	1.5	-3.21
5	2	-5.04
6	2.5	-5.37
7	3	-6.03
8	3.5	-7.21
9	4	-7.46
10	4.5	-7.96

Figure 2.4

2.3 Smoothing splines

A smoothing spline is useful for indicating the shape of a relationship without imposing too much pre-defined structure. You can fit these using Linear Regression menu (Figure 2.5), by selecting Smoothing spline in the drop-down list at the top of the menu.

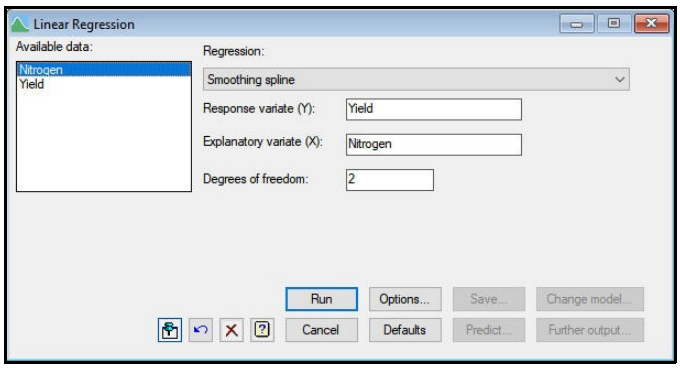


Figure 2.5

We have entered `Yield` as the Response variate, and `Nitrogen` as the Explanatory variate, as before. We have also specified 2 as the number of degrees of freedom for the spline. Essentially this defines how much the original data are to be smoothed. As we have only 5 different nitrogen values in the data, we can chose from 1 to 4 degrees of freedom: 1 corresponds to perfect smoothing (i.e. a straight line), while here 4 correspond to no smoothing (i.e. a curve passing through the mean yield at each of the five distinct values of `Nitrogen`). The fitted model is plotted in Figure 2.6 and the output is shown below.

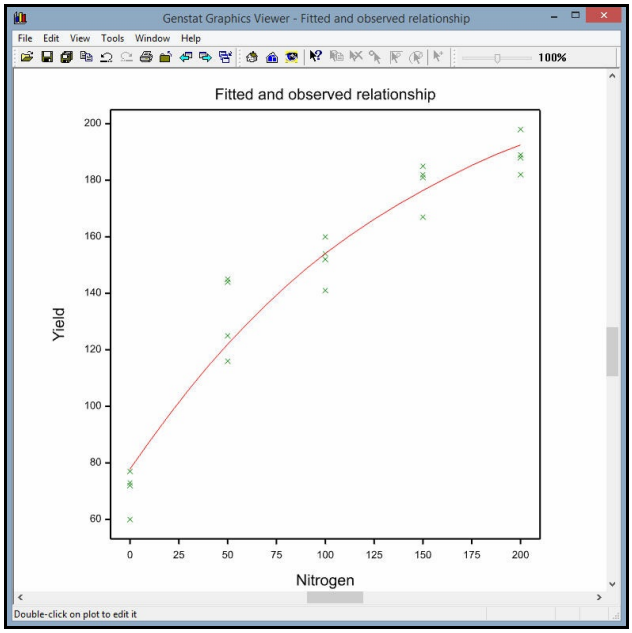


Figure 2.6

Regression analysis

Response variate: Yield
Fitted terms: Constant + Nitrogen
Submodels: SSPLINE(Nitrogen; 2)

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	34649.	17324.6	144.82	<.001
Residual	17	2034.	119.6		
Total	19	36683.	1930.7		

Percentage variance accounted for 93.8
Standard error of observations is estimated to be 10.9.

Message: the following units have large standardized residuals.

Unit	Response	Residual
6	144.0	2.14
7	145.0	2.23

Estimates of parameters

Parameter	estimate	s.e.	t(17)	t pr.
Constant	87.80	4.24	20.73	<.001
Nitrogen Lin	0.5675	0.0346	16.41	<.001

The output does not show the equation of the fitted curve: it is rather complicated,

involving cubic polynomials fitted between each distinct pair of values of Nitrogen. The linear component is, however, estimated and displayed as before. The point of this analysis is to draw the picture, shown in Figure 2.6. This shows a smooth curve quite similar to the previous polynomial curve, but still rising at the largest value of Nitrogen rather than reaching a maximum there.

The SSPLINE function used for the smoothing spline option, and the LOESS function used for locally weighted regression (another option in the drop-down list), are also available in the Operators box if you select General Linear Regression option. A model that contains a smoothing spline or a locally weighted regression is called an *additive model*; for further details, see Section 3.4.3 of Part 2 of the *Guide to the Genstat Command Language*.

2.4 Practical

Fit a smoothing spline to explain the loss in weight data in Practical 2.2. Try different numbers of degrees of freedom to find an appropriately smooth model.

2.5 Standard curves

Genstat provides a range of standard nonlinear curves, chosen to represent many standard situations. These are fitted for you automatically, by the Standard Curves menu. So behind the scenes, Genstat fits the curves by finding the parameter values that maximize the likelihood of the data. Genstat takes care of all the complications that arise in nonlinear model fitting, such as the choice of initial values for the search. It also uses stable forms of parameterization to make the search more reliable (see Ross, G.J.S. 1990, *Nonlinear Estimation*, Springer-Verlag, New York). So you can fit these curves as easily as an ordinary regression.

You open the menu by clicking on the Standard Curve sub-option of the Regression option of the Stats menu. The type of curve is chosen using the drop-down list box at the top. The menu then customizes itself for the selected curve, and displays a small example plot in the box in the left-hand bottom corner. In Figure 2.7 we have chosen an exponential curve. This has the equation

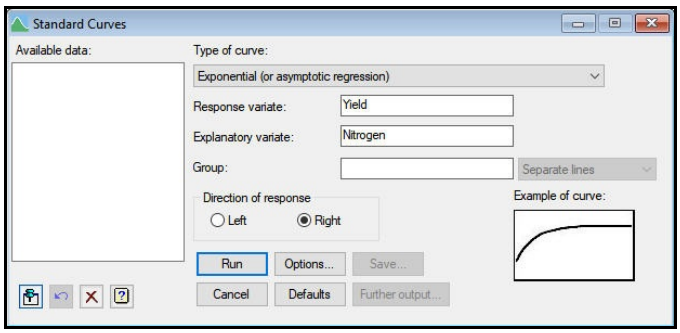


Figure 2.7

$$yield = \alpha + \beta \times \rho^{\,\text{nitrogen}}$$

which represents a curve rising or falling to a plateau or *asymptote* at the value defined by the parameter α. In addition to standard boxes where you to enter the response and explanatory variates, and a group factor (if required), the menu also has a box where you select the Direction of response. If you select left, curve rises or falls from an asymptote on the left of the graph (this corresponds to a value of ρ greater than 1), whereas right gives a curve that rises or falls to an asymptote on the left of the graph (this corresponds to a value of ρ greater 0 but less than 1).

With the sugar-cane data it is clear that we need an asymptote to the right. The results of fitting the curve are shown below (where A represents α, B represents β, and R represents ρ).

Nonlinear regression analysis

Response variate: Yield
Explanatory: Nitrogen
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	35046.	17523.18	182.02	<.001
Residual	17	1637.	96.27		
Total	19	36683.	1930.68		

Percentage variance accounted for 95.0
Standard error of observations is estimated to be 9.81.

Estimates of parameters

Parameter	estimate	s.e.
R	0.98920	0.00213
B	-131.1	10.6
A	203.0	10.8

Note that no t-probabilities are shown in this nonlinear analysis, because both the standard errors and the t-statistics are approximations, which depend on the amount of curvature of the model and on how well it fits the data.

The fitted model is plotted in Figure 2.8. It seems to fit the data well, and has reasonable behaviour at both extremes of the nitrogen fertilizer treatments.

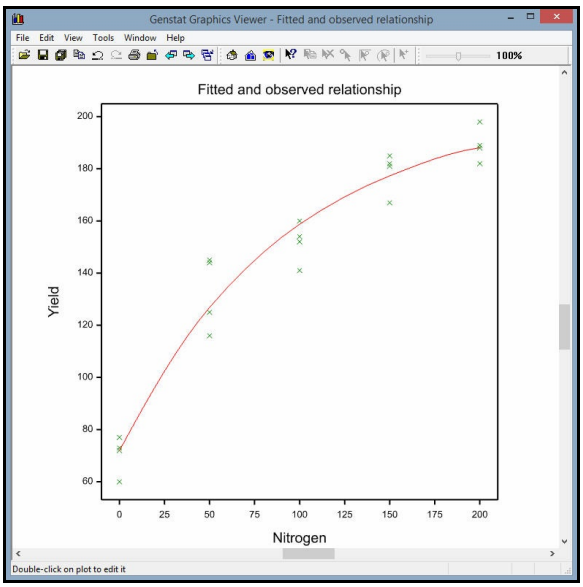


Figure 2.8

The help system has a page with the shapes of all the standard curves (Figure 2.9).

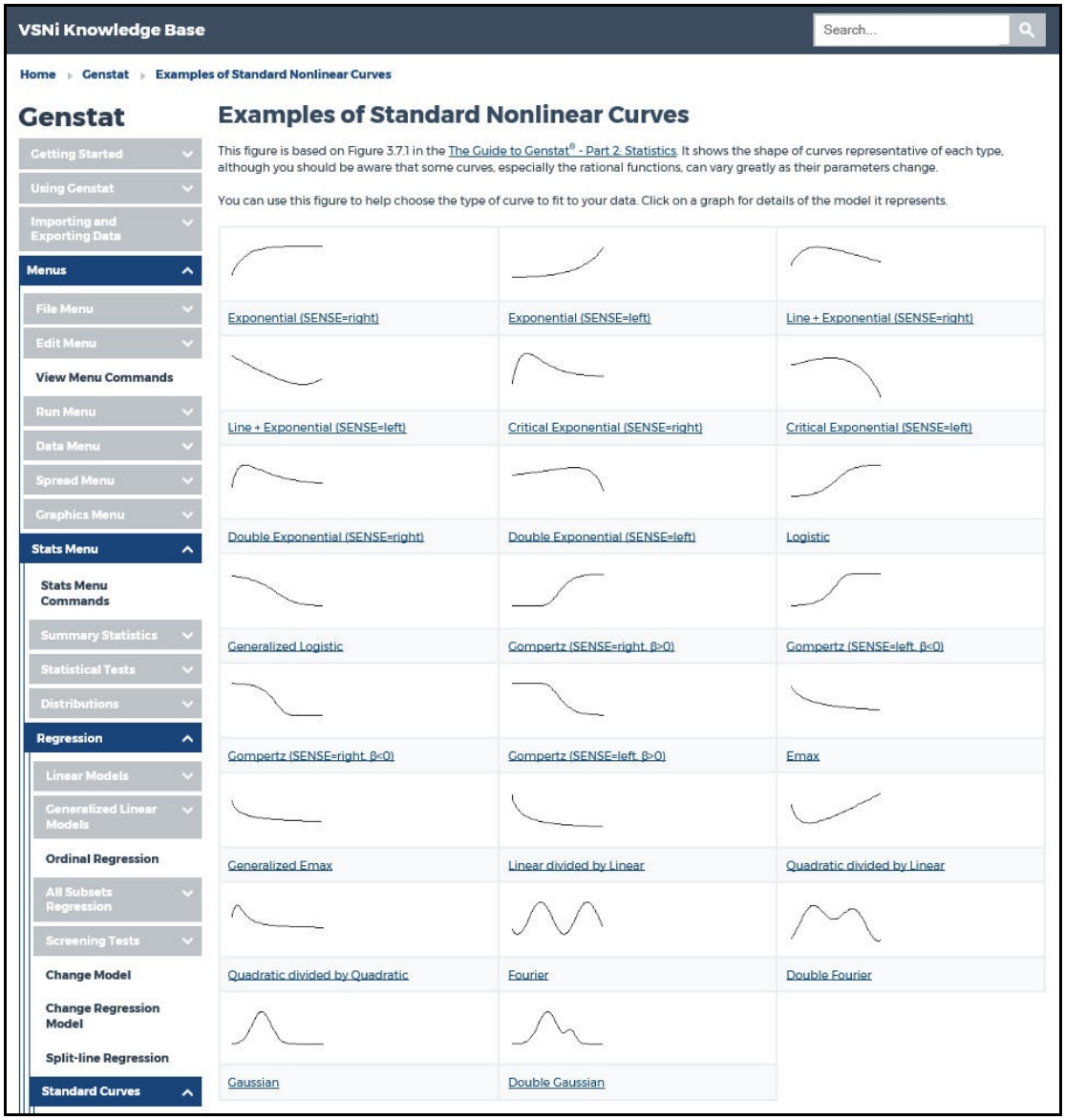


Figure 2.9

Their equations are as follows:

Exponential

exponential $y_i = \alpha + \beta \rho^{x_i} + \varepsilon_i$

dexponential $y_i = \alpha + \beta \rho^{x_i} + \gamma \sigma^{x_i} + \varepsilon_i$

cexponential $y_i = \alpha + (\beta + \gamma x_i) \rho^{x_i} + \varepsilon_i$

lexponential $y_i = \alpha + \beta \rho^{x_i} + \gamma x_i + \varepsilon_i$

Logistic

logistic $$y_i = \alpha + \frac{\gamma}{1 + \exp(-\beta(x_i-\mu))} + \varepsilon_i$$

glogistic $$y_i = \alpha + \frac{\gamma}{(1+\tau\ \exp(-\beta(x_i-\mu)))^{\tau^{-1}}} + \varepsilon_i$$

gompertz $$y_i = \alpha + \gamma\ \exp(-\exp(-\beta(x_i-\mu))) + \varepsilon_i$$

emax $$y_i = \alpha + \frac{\gamma}{1 + \exp(-\beta(\log(x_i)-\mu))} + \varepsilon_i$$

gemax $$y_i = \alpha + \frac{\gamma}{(1+\tau\ \exp(-\beta(\log(x_i)-\mu)))^{\tau^{-1}}} + \varepsilon_i$$

Rational functions

ldl $$y_i = \alpha + \frac{\beta}{1 + \delta\ x_i} + \varepsilon_i$$

qdl $$y_i = \alpha + \frac{\beta}{1 + \delta\ x_i} + \gamma\ x_i + \varepsilon_i$$

qdq $$y_i = \alpha + \frac{\beta + \gamma\ x_i}{1 + \delta\ x_i + \eta\ x_i^2} + \varepsilon_i$$

Fourier

fourier $$y_i = \alpha + \beta\ \sin\left(\frac{2\pi(x_i-\eta)}{\omega}\right) + \varepsilon_i$$

dfourier $$y_i = \alpha + \beta\ \sin\left(\frac{2\pi(x_i-\eta)}{\omega}\right) + \gamma\ \sin\left(\frac{4\pi(x_i-\varphi)}{\omega}\right) + \varepsilon_i$$

Gaussian

gaussian $$y_i = \alpha + \frac{\beta}{\sqrt{2\pi\sigma^2}}\ \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) + \varepsilon_i$$

dgaussian $$y_i = \alpha + \frac{\beta}{\sqrt{2\pi\sigma^2}}\ \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) + \frac{\gamma}{\sqrt{2\pi\sigma^2}}\ \exp\left(\frac{-(x_i-\nu)^2}{2\sigma^2}\right) + \varepsilon_i$$

The standard curves are fitted using the `FITCURVE` directive. This has a parameter to specify the model, and a `PRINT` option just like `FIT`. There is also a `CURVE` option to choose the type of curve; for example:

```
FITCURVE [PRINT=summary; CURVE=exponential] Nitrogen
```

For more information, see the *Guide to the Genstat Command Language, Part 2 Statistics*, Section 3.7.1.

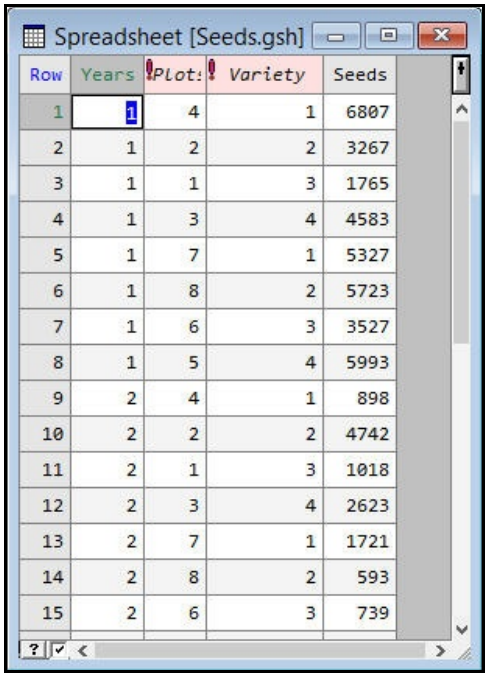
2.6 Practical

Fit an exponential curve to the weight-loss data from Practical 2.2.

2.7 Standard curves with groups

If you have a groups factor, you can investigate the consistency of a nonlinear relationship across the groups. The ideas are very similar to those used to define parallel and non-parallel regression lines in Section 1.15.

We shall illustrate them using the data in spreadsheet file `Seeds.gsh` (Figure 2.10). This records the number of canola seeds recovered in soil cores 0-3 years after growing 4 different varieties. The assumption is that the numbers of seeds will decline exponentially with time, but we would like to know if the rates and curvature of the curves differ according to the variety.



Spreadsheet [Seeds.gsh]

Row	Years	Plot	Variety	Seeds
1	1	4	1	6807
2	1	2	2	3267
3	1	1	3	1765
4	1	3	4	4583
5	1	7	1	5327
6	1	8	2	5723
7	1	6	3	3527
8	1	5	4	5993
9	2	4	1	898
10	2	2	2	4742
11	2	1	3	1018
12	2	3	4	2623
13	2	7	1	1721
14	2	8	2	593
15	2	6	3	739

Figure 2.10

We again use the Standard Curves menu (Figure 2.11), but now specify a group factor (`Variety`) as well as the response variate (`Seeds`) and the explanatory variate (`Years`).

The output is shown below.

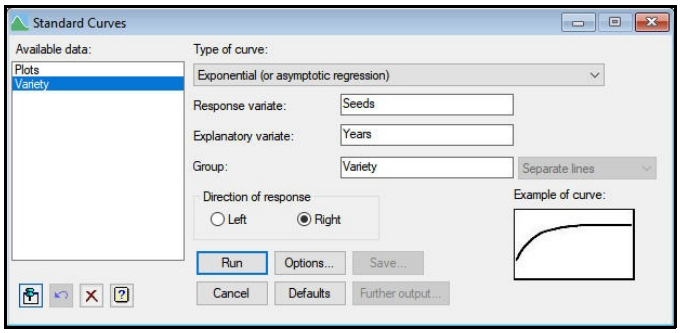


Figure 2.11

Nonlinear regression analysis

Response variate: Seeds
Explanatory: Years
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	86899878.	43449939.	34.31	<.001
Residual	29	36724908.	1266376.		
Total	31	123624785.	3987896.		

Percentage variance accounted for 68.2
Standard error of observations is estimated to be 1125.

Message: the following units have large standardized residuals.

Unit	Response	Residual
3	1765.	-2.72
10	4742.	2.67

Message: the error variance does not appear to be constant; large responses are more variable than small responses.

Estimates of parameters

Parameter	estimate	s.e.
R	0.393	0.155
B	11450.	3595.
A	120.	655.

Nonlinear regression analysis

Response variate: Seeds
Explanatory: Years
Grouping factor: Variety, constant parameters separate
Fitted Curve: A + B*(R**X)
Constraints: R < 1

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	5	94267932.	18853586.	16.70	<.001
Residual	26	29356854.	1129110.		
Total	31	123624785.	3987896.		
Change	-3	-7368054.	2456018.	2.18	0.115

Percentage variance accounted for 71.7
Standard error of observations is estimated to be 1063.

Message: the following units have large standardized residuals.

Unit	Response	Residual
10	4742.	2.53

Message: the error variance does not appear to be constant; large responses are more variable than small responses.

Estimates of parameters

Parameter	estimate	s.e.
R	0.393	0.103
B	11451.	
A Variety 1	260.1	
A Variety 2	284.7	
A Variety 3	-674.7	
A Variety 4	612.7	

Nonlinear regression analysis

Response variate:	Seeds
Explanatory:	Years
Grouping factor:	Variety, all linear parameters separate
Fitted Curve:	A + B*(R**X)
Constraints:	R < 1

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	8	102231932.	12778991.	13.74	<.001
Residual	23	21392854.	930124.		
Total	31	123624785.	3987896.		
Change	-3	-7964000.	2654667.	2.85	0.059

Percentage variance accounted for 76.7
Standard error of observations is estimated to be 964.

Message: the following units have large standardized residuals.

Unit	Response	Residual
10	4742.	2.85

Message: the error variance does not appear to be constant; large responses are more variable than small responses.

Estimates of parameters

Parameter	estimate	s.e.
R	0.3601	0.0881
B Variety 1	16969.	
A Variety 1	-275.3	
B Variety 2	11765.	
A Variety 2	469.2	
B Variety 3	6668.	
A Variety 3	214.8	
B Variety 4	13568.	
A Variety 4	547.8	

Nonlinear regression analysis

Response variate:	Seeds
Explanatory:	Years
Grouping factor:	Variety, all parameters separate
Fitted Curve:	A + B*(R**X)
Constraints:	R < 1

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	11	105574914.	9597719.	10.63	<.001
Residual	20	18049871.	902494.		
Total	31	123624785.	3987896.		
Change	-3	-3342983.	1114328.	1.23	0.323

Percentage variance accounted for 77.4
Standard error of observations is estimated to be 950.

Message: the following units have large standardized residuals.

Unit	Response	Residual
10	4742.	2.91
14	593.	-2.46

Message: the error variance does not appear to be constant; large responses are more variable than small responses.

Estimates of parameters

Parameter	estimate	s.e.
R Variety 1	0.141	0.161
B Variety 1	40105.	44011.
A Variety 1	432.	573.
R Variety 2	0.665	0.333
B Variety 2	9191.	2353.
A Variety 2	-1577.	4341.
R Variety 3	0.367	0.473
B Variety 3	6570.	6989.
A Variety 3	202.	1009.
R Variety 4	0.556	0.266
B Variety 4	10537.	2487.
A Variety 4	-562.	2260.

The analysis first fits a common line to all the years. Then it fits a different asymptote (A) for each variety. Then it generalizes the model further to have different rate parameters (B) for each variety. Then the final model includes different shape parameters (R), so that all the parameters differ between varieties. (The parameters A and B are the linear parameters in the model, and it seems more natural that they might vary between groups than the nonlinear parameter R. So the sequence varies those first.)

We can produce an *analysis of parallelism* using the Accumulated option of the Standard Curve Further Output menu, as shown in Figure2.12. So we can assess how complicated a model we need, and then perhaps set the Final Model box (alongside the Group box in Figure 2.11) and refit the model as in the ordinary linear regression with groups discussed in Section 1.15.

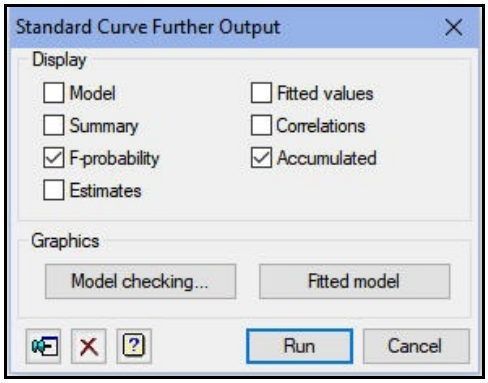


Figure 2.12

Nonlinear regression analysis

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ Years	2	86899878.	43449939.	48.14	<.001
+ Variety	3	7368054.	2456018.	2.72	0.072
+ Years.Variety	3	7964000.	2654667.	2.94	0.058
+ Separate nonlinear	3	3342983.	1114328.	1.23	0.323
Residual	20	18049871.	902494.		
Total	31	123624785.	3987896.		

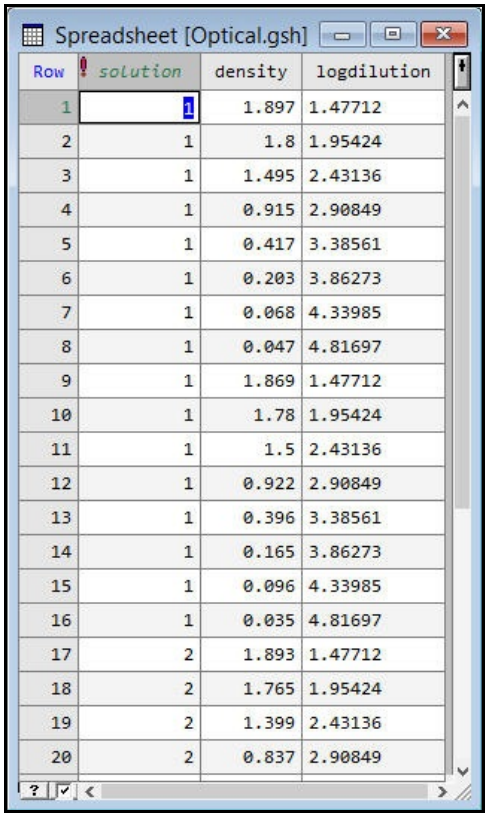
The output suggests that we need different asymptote and parameters (A), and possibly different rate parameters (B), but that there is no evidence that we need different nonlinear shape parameters (R).

2.8 Practical

Spreadsheet file `Optical.gsh`, contains optical densities of three solutions measured at various dilutions (transformed to log10).

Fit a logistic curve to the relationship between `density` and `logdilution`.

Include `solution` as the group factor to assess the consistency of the model over the different solutions.



Spreadsheet [Optical.gsh]

Row	solution	density	logdilution
1	1	1.897	1.47712
2	1	1.8	1.95424
3	1	1.495	2.43136
4	1	0.915	2.90849
5	1	0.417	3.38561
6	1	0.203	3.86273
7	1	0.068	4.33985
8	1	0.047	4.81697
9	1	1.869	1.47712
10	1	1.78	1.95424
11	1	1.5	2.43136
12	1	0.922	2.90849
13	1	0.396	3.38561
14	1	0.165	3.86273
15	1	0.096	4.33985
16	1	0.035	4.81697
17	2	1.893	1.47712
18	2	1.765	1.95424
19	2	1.399	2.43136
20	2	0.837	2.90849

Figure 2.13

2.9 Nonlinear models

If you want to fit a curve that the Standard Curves menu does not cover, Genstat has an alternative menu, shown in Figure 2.14, that allows you to define and fit your own nonlinear curves. This is obtained by clicking on the Nonlinear Models sub-option of the Regression option of the Stats menu. We illustrate it by refitting the exponential model to the sugar-cane data in Section 2.1.

First we enter `Yield` into the Response variate field in the usual way. Then we must define the model to be fitted. This can contain a mixture of linear and nonlinear terms. The nonlinear terms are defined by clicking on the New button in the Model expressions section. This opens the Generate Expression menu (Figure 2.15) which you use to specify expressions to define the nonlinear parts of the model. Here we have defined the expression

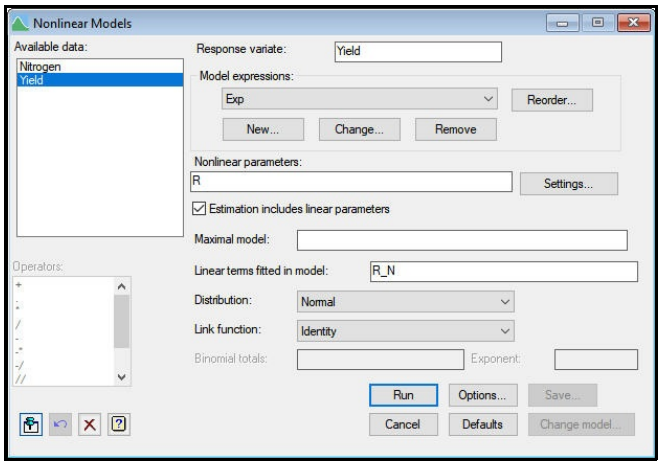


Figure 2.14

```
R_N = R ** Nitrogen
```

The expression has been given the *reference identifier* `Exp` and, when we click on OK, this identifier is entered (automatically) by Genstat into the Model expressions list box in the Nonlinear Models menu (Figure 2.14).

The variate `R_N` is a linear term, as the model can be written as

```
A + B * R_N
```

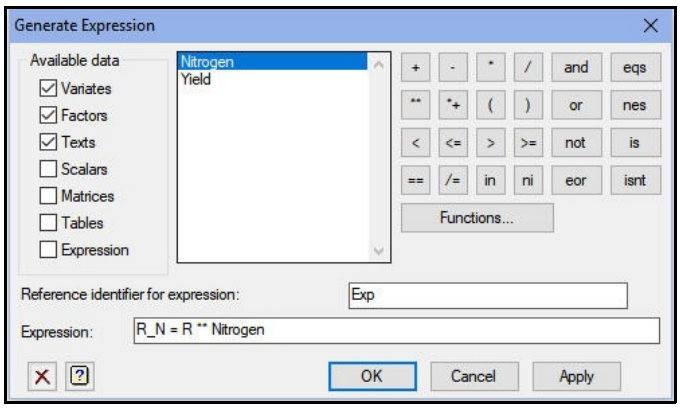


Figure 2.15

So we check the Estimation includes linear parameters box, and enter `R_N` into the Linear terms fitted in model box.

It is much more efficient to estimate the parameters `A` and `B` in this way. The alternative would be to define the whole model in the expression, for example by

```
FitYield = A + B * R**Nitrogen
```

The expression sets variate `FitYield` to the fitted values given by the model. If the Estimation includes linear parameters box is not checked, the MaximalmModel box is replaced by a box called Fitted values into which you should enter the name of the fitted values variate `FitYield`.

Notice that you can have more than one linear term. In fact you can define a maximal model and use the Change Model menu as in Section 1.11 to decide which ones are needed. The Distribution and LinkfFunction boxes allow you to define and fit generalized nonlinear models (see Section 3.5.8 of Part 2 of the *Guide to the Genstat Command Language*). The default settings of `Normal` and `Identity`, as in Figure 2.14 fit the usual type of nonlinear model in which the residuals are assumed to be Normally distributed.

The next step is to list the nonlinear parameters (in this case just `R`) in the Nonlinear parameters box of the Nonlinear Models menu (Figure 2.14). You will need to set initial values for these, and possibly also bounds and steplengths, by using the Nonlinear Parameter Settings menu (Figure 2.16), opened by clicking on the Settings button in the Nonlinear Models menu. Here we have set an initial value of 0.9, and defined an upper bound of 1.0, but have not defined any lower bound and have left Genstat to decide on the step length to be used.

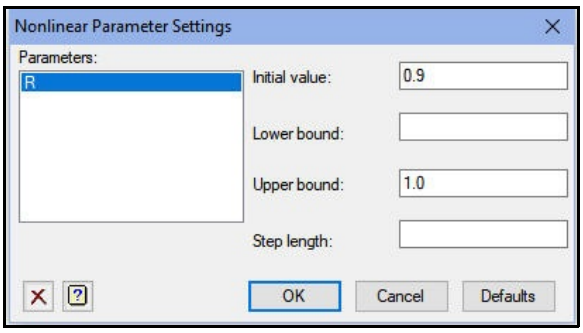


Figure 2.16

Finally, clicking on Run in Figure 2.14 produces the output below.

Nonlinear regression analysis

Response variate: Yield
Nonlinear parameters: R
Model calculations: Exp
Fitted terms: Constant, R_N

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	35046.	17523.13	182.01	<.001
Residual	17	1637.	96.28		
Total	19	36683.	1930.68		

Percentage variance accounted for 95.0
Standard error of observations is estimated to be 9.81.

Estimates of parameters

Parameter	estimate	s.e.
R	0.98926	0.00212
* Linear		
Constant	203.3	10.9
R_N	-131.4	10.7

The parameter `B` is now the regression coefficient of `R_N` and `A` is now the `Constant`, but otherwise the results are virtually identical to those given (rather more easily) by the Standard Curves menu. Further information about nonlinear curve fitting, and the directives `RCYCLE` and `FITNONLINEAR` that are used, is in Section 3.8 of Part 2 of the *Guide to the Genstat Command Language*.

2.10 Practical

Spreadsheet file `MMdata.gsh`, contains measurements of two chemical variables F and S.

Fit the Michaelis-Menten equation to the relationship between F and S. This is a hyperbola through the origin, usually parameterized as

$$S = p2 \times F / (p1 + F).$$

Hint 1: use starting values of $p1 = 1$ and $p2 = 15$.

Hint2 : you can fit $p2$ as a linear parameter.

Compare your model with the standard linear-by-linear curve.

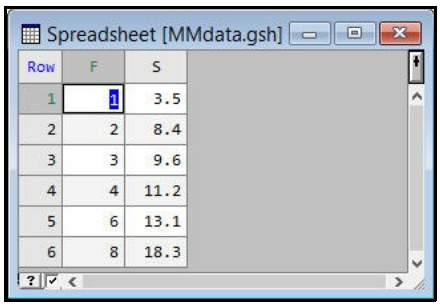


Figure 2.17

3 Generalized linear models

The regression menus that we have seen so far are intended for continuous data that can be assumed to follow a Normal distribution.

Generalized linear models extend the usual regression framework to cater for non-Normal distributions. For example:

- Poisson distributions – for counts, such as number of items sold in a shop, or numbers of accidents on a road, number of fungal spores on plants etc;
- binomial data recording *r* "successes" out of *n* trials, for example numbers of surviving patients out of those treated, or weeds killed out of those sprayed, or flies killed by an insecticide etc;
- gamma distributions for positively-skewed data.

They also incorporate a *link function* that defines the transformation required to make the model linear. For example:

- logarithm base e for Poisson data (counts);
- logit, probit or complementary log-log for binomial data;
- logarithm or reciprocal for the gamma distribution.

The most important point is that, once you have defined the distribution and link function, fitting a generalized linear model in Genstat is very similar to the way in which we fitted the ordinary regression models in Chapter 1. So you just need to know how your data are distributed, and the appropriate scale for the model.

So, in this chapter you will learn

- the terminology and equations that underlie generalized linear models (★)
- how to fit log-linear models to count data
- how to fit logistic regression models to binomial data
- how to fit probit models to binomial data
- how to use generalized linear mixed models to model non-Normal data when there are several sources of error variation ★
- how to use hierarchical generalized linear mixed models to model non-Normal data when there are several sources of error variation ★

Note: the topics marked ★ are optional.

3.1 Equations and terminology

In an ordinary regression the underlying model is

$$y = \mu + \varepsilon$$

where

μ is the mean to be predicted by the linear regression model, and

ε is the residual, assumed to come from a Normal distribution with mean zero and variance σ^2.

The mean is known as the *expected value* of y, and is estimated by the fitted value from the regression. For example, in the simple linear regression in Section 1.1, the fitted value was

$$f = b \times x + c$$

where b was the regression coefficient and c was the constant term (or intercept). Equivalently, we can say that y has Normal distribution with mean μ and variance σ^2.

In a generalized linear model the expected value of y is still μ, but the linear model now defines the *linear predictor*, usually represented by η, which is related to μ by the *link function* g():

$$\eta = g(\mu).$$

For example, in the log-linear model in the next section, we have a single variate `temperature` and a logarithmic link function . So we have

$$\log(\mu) = \eta = b \times \texttt{temperature} + c$$

The other extension is that y has a distribution with mean μ from a wider class of distributions known as the *exponential family*. This includes the binomial, Poisson, gamma, inverse-normal, multinomial, negative-binomial, geometric, exponential and Bernoulli distributions, as well as the usual Normal distribution.

The fitting of data from all of the distributions, apart from the Normal, is complicated by the fact that their variances change according to their means. For example the variance of a Poisson distribution is equal to its mean. The algorithm that is used to fit a generalized linear model allows for this by doing a weighted linear regression. The weights depend on the means, but the means are estimated by a regression that uses the weights. So an iterative process is used where the means and weights are recalculated alternately until the estimation converges. If you are interested, you can find full details in McCullagh & Nelder (1989, *Generalized Linear Models, second edition*). However, another important point is that you do not need to know how the algorithm works in order to use a generalized linear model – this is reliably programmed and safely concealed inside Genstat. It is worth remembering, though, that the fit is essentially achieved by a weighted linear regression. So we can still use the standard model checking plots described in Section 1.3.

3.2 Log-linear models

Often the data may consist of counts. For example, you may have recorded the number of various types of items that have been sold in a shop, or numbers of accidents occurring on different types of road, or the number of fungal spores on plants with different spray treatments. Such data are generally assumed to follow a Poisson distribution. At the same time, it is usually assumed also that treatment effects will be proportionate (that is, the effect of a treatment will be to multiply the expected count by some number, rather than

to increase it by some fixed amount). So, the model will be linear on a logarithmic scale rather than on the natural scale as used in ordinary linear regression. Models like this are known as *log-linear models* and form just one of the types of model covered by Genstat's facilities for generalized linear models.

The Generalized Linear Models menu is obtained by clicking on the Generalized Linear line in the Regression section of the Stats menu. For a log-linear model, you should then select Log-linear modelling in the Analysis drop-down list box, as shown in Figure 3.1. The menu now looks very similar to the General Linear Regression menu (Figure 1.22), and operates in a very similar way. So you can define a maximal model and then investigate which of its terms are required, as we did in Section 1.11.

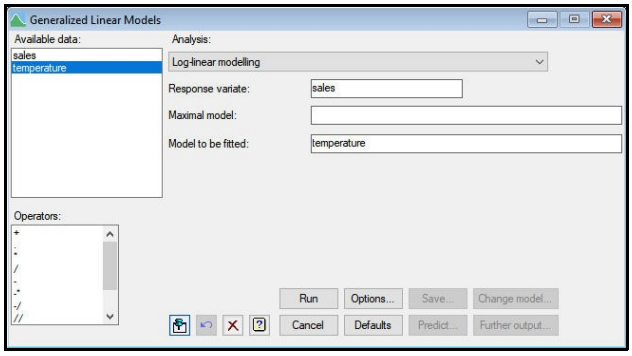
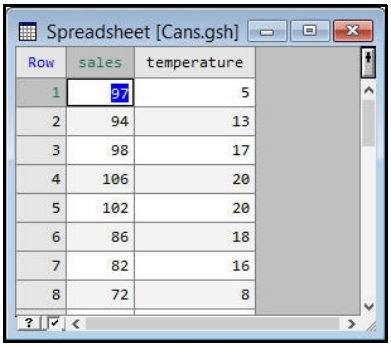


Figure 3.1

We shall use the menu to analyse a data set in Genstat spreadsheet file `Cans.gsh` (Figure 3.2). The response variate is the number of cans of drink (`sales`) sold by a vending machine during 30 weeks. The model to be fitted has just a single explanatory variate, `temperature`, which is the average temperature during the corresponding week. As we have a single explanatory variate, there is no need to specify the Maximal model. Clicking on Run produces the output below.



Row	sales	temperature
1	97	5
2	94	13
3	98	17
4	106	20
5	102	20
6	86	18
7	82	16
8	72	8

Figure 3.2

Regression analysis

Response variate: sales
Distribution: Poisson
Link function: Log
Fitted terms: Constant, temperature

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	1	52.61	52.614	52.61	<.001
Residual	28	32.05	1.145		
Total	29	84.66	2.919		

Dispersion parameter is fixed at 1.00.

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

Message: the following units have large standardized residuals.

Unit	Response	Residual
30	137.00	2.87

Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant	4.3410	0.0303	143.49	<.001	76.78
temperature	0.01602	0.00222	7.22	<.001	1.016

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

The initial description contains the extra information that the data have a Poisson distribution, and that the *link* function (the transformation required to give a scale on which the model is linear) is the logarithm to base e. These are the two aspects required to characterize a generalized linear model. In the Log-linear modelling menu they are set automatically, but you can also select General Model in the Analysis field to obtain a menu where you can set these explicitly, and thus fit any of Genstat's generalized linear models.

With generalized linear models, the summary of analysis contains *deviances* instead of sums of squares. Under the null hypothesis they have χ^2 distributions, and a quick rule-of-thumb is that their expected values are equal to their degrees of freedom.

However, some sets of data show *over-dispersion*. The residual deviance is then noticeably greater than its expectation and, instead of assessing the regression line by comparing its deviance with χ^2, you should use the deviance ratio (and assess this using an F distribution).

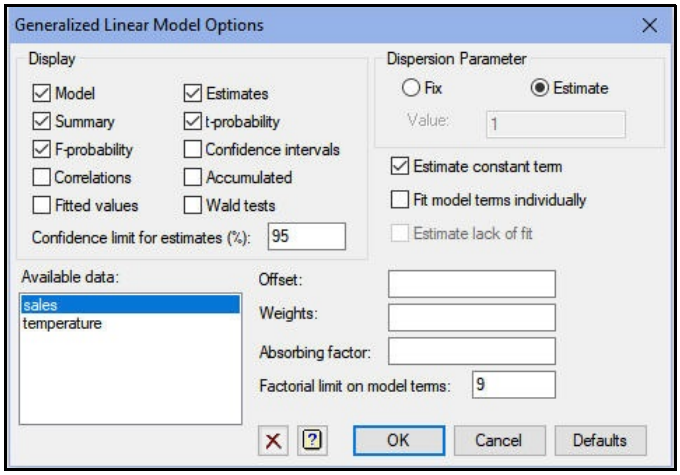


Figure 3.3

Genstat will do this for you if you ask it to estimate the *dispersion parameter*, by checking Estimate in the Dispersion parameter section of either the Generalized Linear Model Options menu (Figure 3.3) or the Generalized Linear Models Further Output menu (Figure 3.4). Genstat then also adjusts the standard errors of the parameter estimates to take account of the over dispersion.

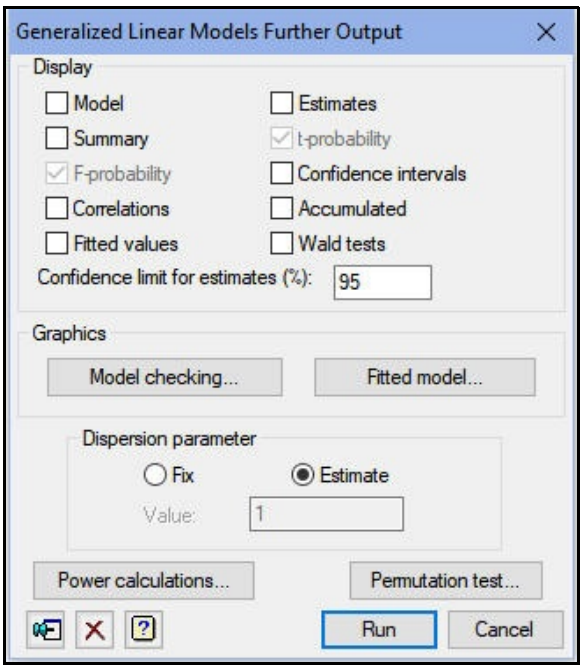


Figure 3.4

Note, however, that the residual deviance may be large not because of over dispersion, but simply because some important terms have been omitted from the model (and these may not even be available in the data set). You should then keep the dispersion parameter at the default value of 1, and continue to assess the deviances using χ^2 distributions. Further details are given in Section 3.5.1 of Part 2 of the *Guide to the Genstat Command Language*.

Here, though, the residual deviance is not substantially more than its expectation (as illustrated by the fact that its mean deviance is 1.145). So we can treat the regression deviance as χ^2 on one degree of freedom – and note that there seems to be a very strong effect of temperature on sales.

The Generalized Linear Model Options menu (Figure 3.3) contains several controls that do not occur in the Linear Regression Options menu (Figure 1.8). An offset variate is a variate that is included in the linear predictor with a constant regression parameter of 1. In log-linear models it can be used to adjust the model when the counts have been made over periods with different lengths. For example, suppose the can data had been collected over months instead of years. We would then need to take account of the fact that the months may contain between 28 and 31 days. If we include the logarithm of the number of days in the relevant month as an offset, the model becomes

$$\log(\text{sales}) = \log(\text{days}) + b \times \text{temperature} + c$$

This means that we have

$$\log(\text{sales/days}) = b \times \text{temperature} + c$$

So we have corrected for the unequal lengths of the months, and are using the linear model to describe the rates at which the sales are made. Notice that this is more valid than the alternative of adjusting the response variate itself; a response variate of sales/days would no longer follow a Poisson distribution.

The other important control is the check box where you can ask to fit the model terms individually. By default, all the terms are fitted in a single step. So the accumulated analysis of deviance will not allow you to assess their individual effects, as it would in an ordinary regression analysis. Here we have only one model term (`temperature`), so we can leave the box unchecked.

The fitted model can be displayed by clicking on the Fitted model button in the Generalized Linear Models Further Output menu (Figure 3.4) to obtain the Graph of Fitted Model menu (Figure 3.5).

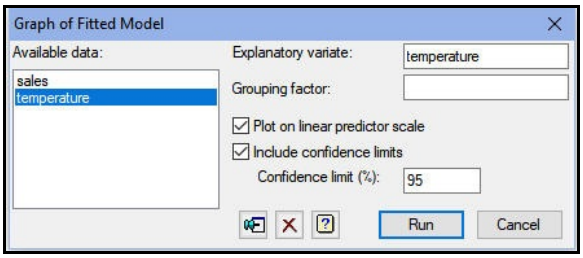


Figure 3.5

This menu appears whenever you ask to plot the fitted model from one of the regression menus where you yourself specify which model to fit. (So, for example, it would also have been used if we had chosen to plot the water data in Section 1.11). There may then be several variates or factors to use for the x-axis or to define groups. Here there is only the variate `temperature`, so we enter that as the explanatory variable, and click on Run to plot the graph.

When used with a generalized linear model, the menu has an extra box to allow you to plot the y-axis on the linear predictor (i.e. here the logarithmic) scale instead of the natural scale (here counts). In Figure 3.5 we have chosen to do that, and we can then include 95% confidence limits for the response; see Figure 3.6. The line should be straight, so this also allows us to assess any nonlinearity in the response. The alternative is to plot with the y-axis on the natural scale.

The scale of the y-axis in the graph (Figure 3.6) illustrates the logarithmic link transformation, and you can see the point with the large residual (on the top right of the plot).

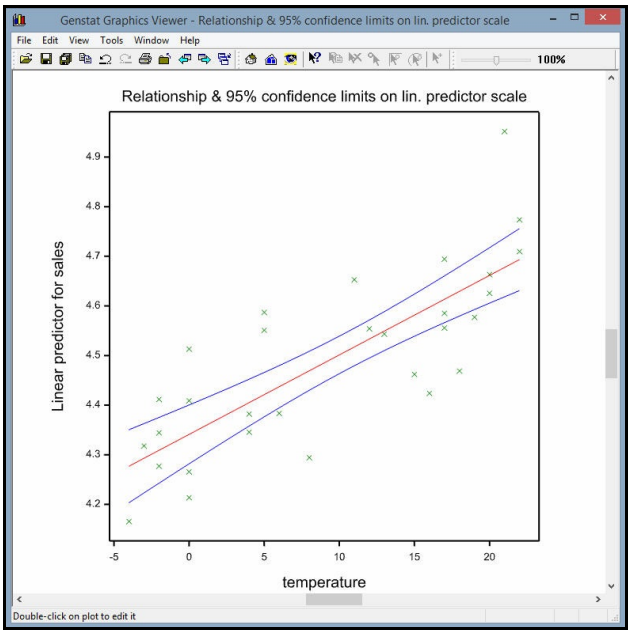


Figure 3.6

If we plot instead on the natural scale (Figure 3.7), you can see how the fitted values increase exponentially (the inverse transformation of the logarithm) with temperature.

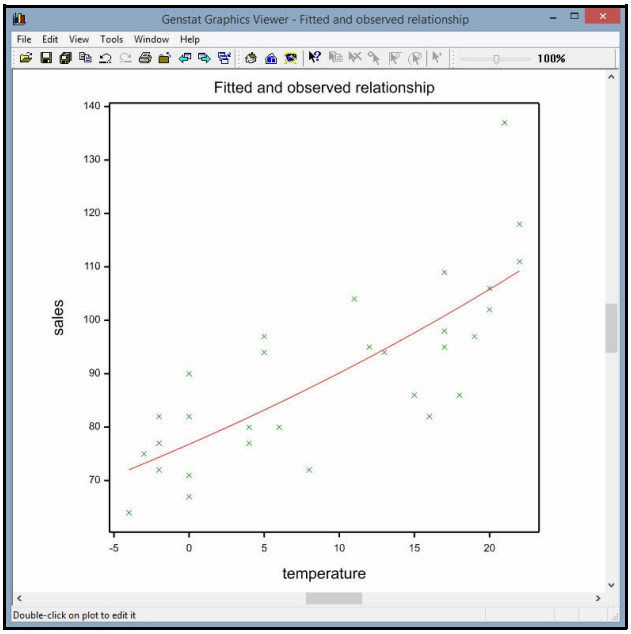


Figure 3.7

You can also produce the model-checking plots (Figure 3.8) in the same way as in earlier sections. Remember that the model is essentially fitted by a weighted regression (Section 3.1), and notice that the residuals are standardized by dividing each one by its variance. You would therefore expect that the residuals should look asymptotically like residuals from a Normal distribution. So, provided, we have a reasonably large data set, we should be able to assess the fit and model assumptions in the same way as in an ordinary linear regression.

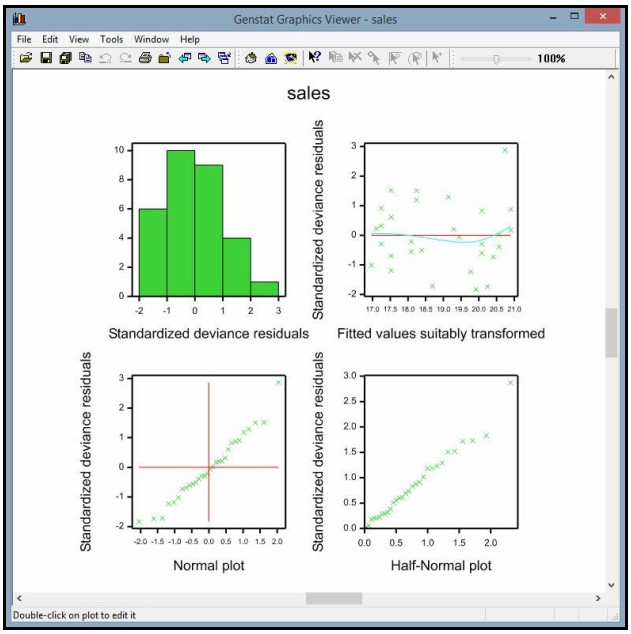


Figure 3.8

You will see, later in this chapter, that the Generalized Linear Models menu also has customized menus for binomial data, where each data value records a number of subjects responding out of a total number observed. Furthermore, as you will see in the next practical, the models can involve factors as well as variates.

The similarity of the menus for generalized linear models to those for ordinary linear regression is matched by the similarity of the commands that are used. The main point is that you must use the `MODEL` directive not only to define the response variate, but also to define the distribution and link function using its `DISTRIBUTION` and `LINK` options. For binomial data (Section 3.4), the response variate contains the number of "successes"

r, and you set the `NBINOMIAL` parameter to a variate containing the corresponding numbers observed *n*.

3.3 Practical

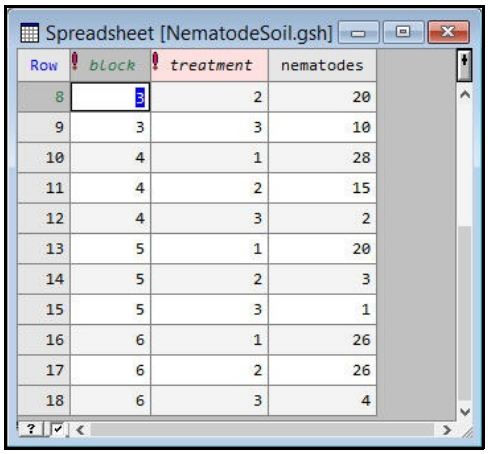
Spreadsheet file `NematodeSoil.gsh`, contains the results of an experiment to compare three ways of controlling nematodes in soil. There were six blocks, each containing a plot for each of the three methods of control (i.e. a randomized block design).

Analyse the counts, assuming they have Poisson distributions and that blocks and treatments have multiplicative effects

Hint 1: the model should be

```
block + treatment
```

Hint 2: remember to check the Fit model terms individually box in the Generalized Linear Model Options menu.



Row	block	treatment	nematodes
8	3	2	20
9	3	3	10
10	4	1	28
11	4	2	15
12	4	3	2
13	5	1	20
14	5	2	3
15	5	3	1
16	6	1	26
17	6	2	26
18	6	3	4

Figure 3.9

3.4 Logistic regression and probit analysis

Probit analysis and logistic regression model the relationship between a stimulus, like a drug, and a *quantal* response i.e. a response that may be either success or failure.

The probit model was originally derived by assuming there is a certain level of dose of the stimulus for each subject below which it will be unaffected, but above which it will respond. This level of dose, known as its *tolerance*, will vary from subject to subject within the population. In probit analysis, it is assumed that the tolerance to the dose (or often the logarithm of the dose) has a Normal distribution. So, if we plot the proportion of the population with each tolerance against log dose, we will obtain the familiar bell-shaped curve shown in Figure 3.10.

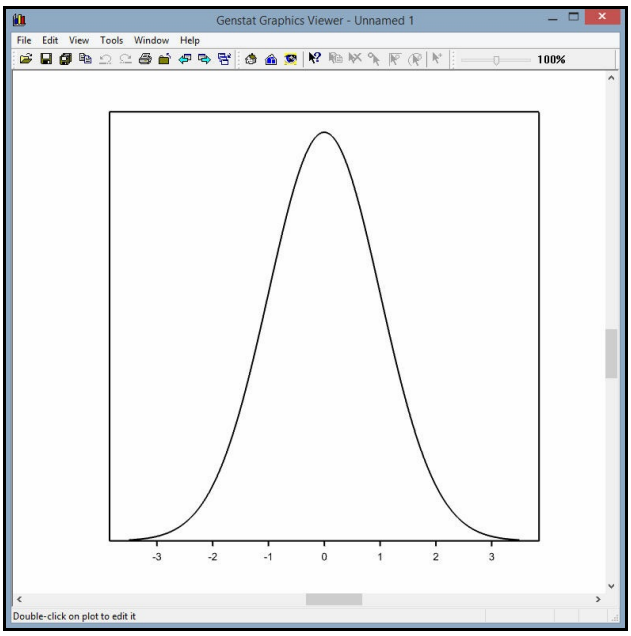


Figure 3.10

The probability of response to a dose x is the proportion of the population with a tolerance of less than x. We can read this off the cumulative Normal curve (Figure 3.11).

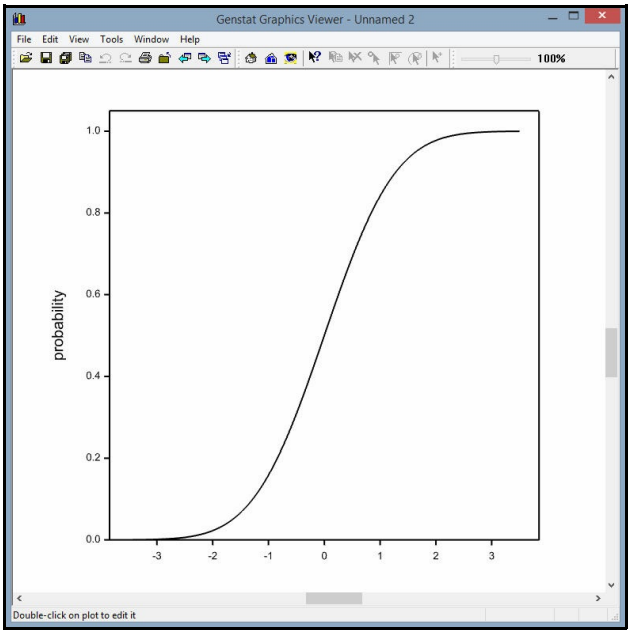


Figure 3.11

We can make the relationship linear by transforming the y-axis to Normal equivalent deviates or *probits* as shown in Figure 3.12. This "stretches" the axis at the top and bottom to make the response into a straight line.

The probit transformation is frequently used in biology, for example to model the effects of insecticides and other situations where the underlying assumption of a tolerance distribution seems natural. The *logit* transformation is also very popular:

$$\text{Logit}(p) = \log(p / q)$$

where p is the probability expressed as a percentage, and

$$q = 100 - p$$

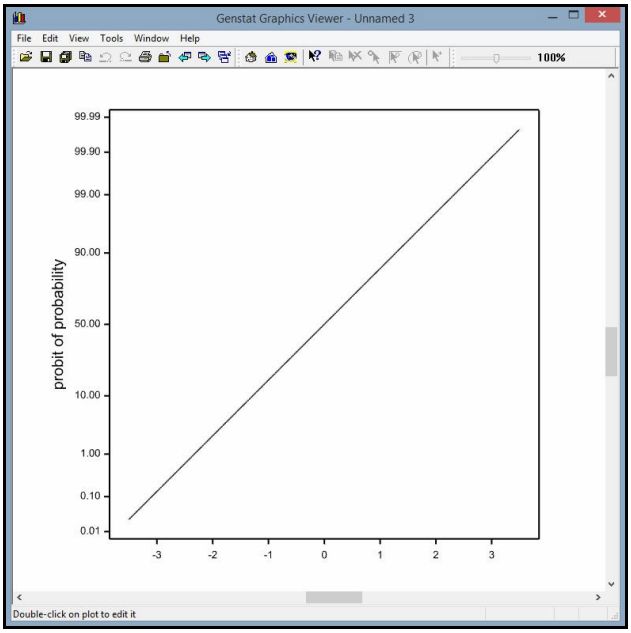


Figure 3.12

This looks similar to the probit, but its tolerance distribution has slightly fatter tails (see Figure 3.13) . It is the logarithm of the odds ratio, and involves proportionate changes in p when p is small, and proportionate changes in $100–p$ when p is near 100.

The third available transformation is the *complementary log-log* transformation. This can be derived from the "one-hit" model where "success" arises from infection by one or more particles that come from a Poisson distribution. It is defined as

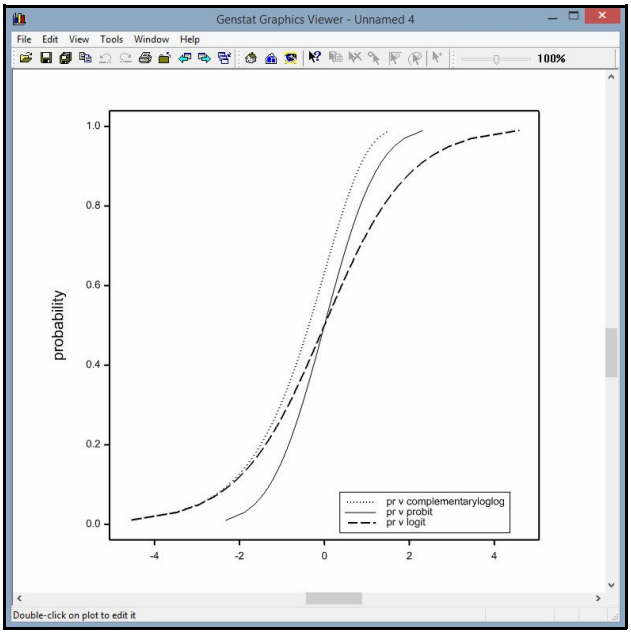


Figure 3.13

$$\text{complementary log-log}(p) = \log(-\log(q))$$

It is similar to the logit for small values of p, but then rises more steeply i.e. the tolerance distribution has a smaller upper tail (see Figure 3.13).

To investigate a logistic regression or probit relationship, you would do an experiment where subjects are given treatments of the drug (or other stimulus) and then observed to see if they give the desired reaction. Usually several subjects receive each treatment. So we have binomial data, with n subjects receiving a dose, and r responding. Usually n is greater than one; *binary* experiments with n=1 are sometimes used, but the model parameters tend to be less well estimated.

So we have a generalized linear model with a binomial distribution, and the choice of either a probit, a logit or a complementary-log-log link function.

Figure 3.14 shows spreadsheet file `Drug.gsh`, which contains an example from Finney (1971, *Probit Analysis, 3rd Edition*, page 103). This compares the effectiveness of three analgesic drugs to a standard drug, morphine. Fourteen groups of mice were tested for response to the drugs at a range of doses. The variate `N` records total number of mice in each group, and `R` records the number that responded. Instead of `Dose` we will fit `LogDose`, the logarithm (base 10) of the dose, which we can calculate in the usual way (see Figure 1.32).

Spreadsheet [Drug.gsh]

Row	Drug	Dose	N	R
1	Morphine	1.5	103	19
2	Morphine	3	120	53
3	Morphine	6	123	83
4	Amidone	1.5	60	14
5	Amidone	3	110	54
6	Amidone	6	100	81
7	Phenadoxone	0.75	90	31
8	Phenadoxone	1.5	80	54
9	Phenadoxone	3	90	80
10	Pethidine	5	60	13
11	Pethidine	7.5	85	27
12	Pethidine	10	60	32
13	Pethidine	15	90	55
14	Pethidine	20	60	44

Figure 3.14

Genstat has a custom setting of the Generalized Linear Models menu (Figure 3.15) for logistic regression which automatically sets the distribution to binomial. It has boxes where you enter the variates containing the total numbers of subjects, and the numbers of successes (i.e. the numbers responding). There is also a drop-down list box where you choose the link transformation.

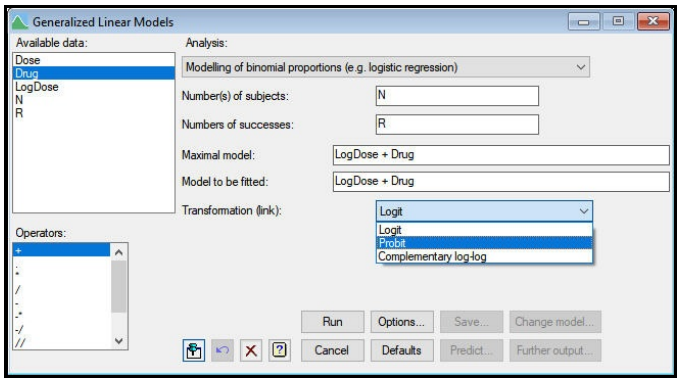


Figure 3.15

Otherwise, the menu has boxes for the maximal model and the model to be fitted, just like the General linear regression setting of the Linear Regression menu. So you can explore the available models, using the Change Model menu as we showed for an ordinary linear regression in Section 1.11.

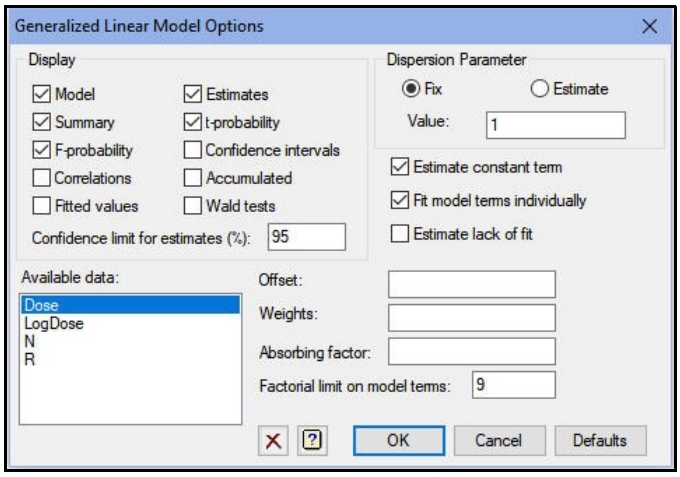


Figure 3.16

If we set the options menu (Figure 3.16) to fit terms individually and print the accumulated summary (and then click on Run in the Generalized Linear Models menu), we obtain the output below.

Regression analysis

Response variate: R
Binomial totals: N
Distribution: Binomial
Link function: Probit
Fitted terms: Constant + LogDose + Drug + LogDose.Drug

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	7	247.624	35.3748	35.37	<.001
Residual	6	2.334	0.3891		
Total	13	249.958	19.2275		
Change	-3	-1.534	0.5112	0.51	0.675

Dispersion parameter is fixed at 1.00.

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

Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.
Constant	-1.255	0.171	-7.34	<.001
LogDose	2.226	0.304	7.32	<.001
Drug Amidone	-0.006	0.272	-0.02	0.983
Drug Phenadoxone	1.205	0.197	6.11	<.001
Drug Pethidine	-1.194	0.402	-2.97	0.003
LogDose.Drug Amidone	0.475	0.485	0.98	0.328
LogDose.Drug Phenadoxone	0.480	0.475	1.01	0.313
LogDose.Drug Pethidine	0.134	0.464	0.29	0.772

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
Drug	Morphine

Accumulated analysis of deviance

Change	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
+ LogDose	1	39.4079	39.4079	39.41	<.001
+ Drug	3	206.6821	68.8940	68.89	<.001
+ LogDose.Drug	3	1.5336	0.5112	0.51	0.675
Residual	6	2.3344	0.3891		
Total	13	249.9579	19.2275		

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

The conclusion from the accumulated summary, is that there are (log)dose and drug effects, but no interaction. So the data can be described by parallel (log)dose lines with a different intercept for each drug. We can drop the interaction using the Change Model menu, as shown in Figure 3.17, to obtain parameter estimates for the parallel-line model.

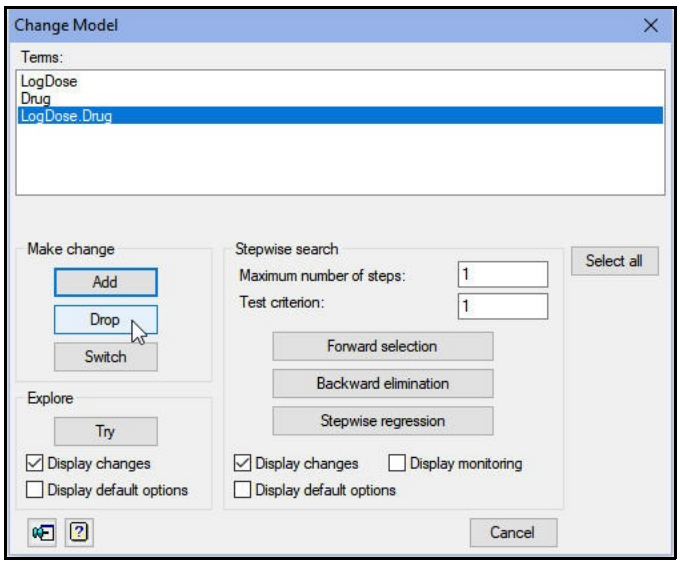


Figure 3.17

Regression analysis

Response variate:	R
Binomial totals:	N
Distribution:	Binomial
Link function:	Probit
Fitted terms:	Constant + LogDose + Drug

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	4	246.090	61.5225	61.52	<.001
Residual	9	3.868	0.4298		
Total	13	249.958	19.2275		
Change	3	1.534	0.5112	0.51	0.675

Dispersion parameter is fixed at 1.00.

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

Estimates of parameters

Parameter	estimate	s.e.	t(*)	t pr.
Constant	-1.379	0.114	-12.08	<.001
LogDose	2.468	0.173	14.30	<.001
Drug Amidone	0.238	0.108	2.20	0.028
Drug Phenadoxone	1.360	0.130	10.49	<.001
Drug Pethidine	-1.180	0.133	-8.87	<.001

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
Drug	Morphine

The Probit analysis setting of the Generalized Linear Models menu (Figure 3.18) provides further customization, but only for models with up to one variate and one factor. So it covers parallel and non-parallel lines like those described in Section 1.15. One of the extra controls allows you to make the log transformation automatically. Here we take logarithms base 10 and again store the results in `LogDose`.

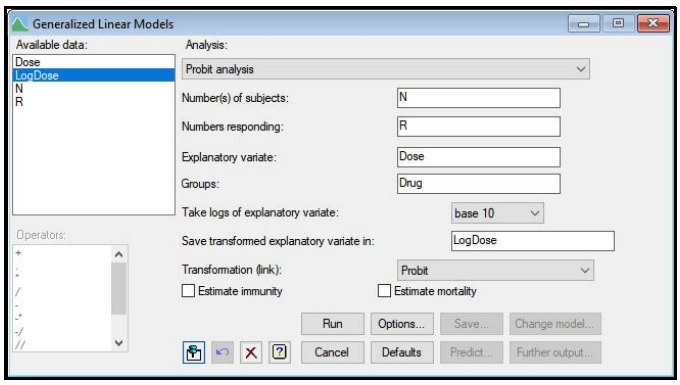


Figure 3.18

You can also estimate natural mortality and immunity. Natural mortality occurs when there are some subjects that will always respond even if there is no treatment. Conversely, natural immunity occurs when there are some subjects that will never respond however large the dose. These are illustrated in Figure 3.19.

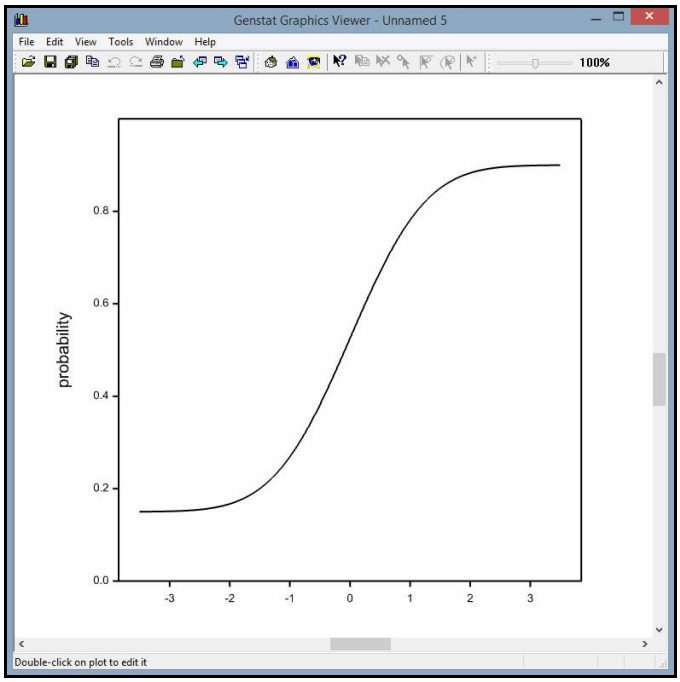


Figure 3.19

The Probit Analysis Options menu (Figure 3.20) also has some extensions. You can decide whether to estimate separate slopes, mortality or immunity parameters in the different groups. Here we have left the slope box unchecked, as we have already discovered that we do not need different slopes. The mortality and immunity boxes are irrelevant as we not estimating either of these in the

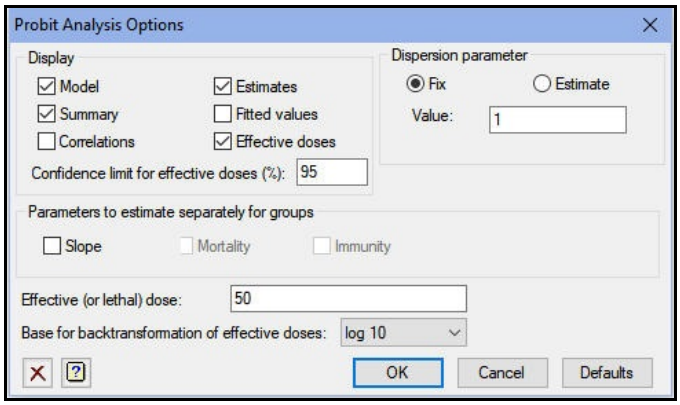


Figure 3.20

main menu.

The other addition is that we can print estimated lethal doses. Here we have asked Genstat to estimate LD50, that is the dose at which 50% of the population would respond. We have also asked to print the back transformed LD50s, by setting the drop-down list box to log 10 to indicate the base of the original transformation of the doses. (The alternative settings are None and log e.)

As the menu is customized for regression with groups, notice that the analysis estimates a separate intercept for each group (i.e. drug). The more general logistic regression menu estimated an intercept for the reference group Morphine and, for the other groups, differences between their intercepts and the Morphine intercept.

Regression analysis

Response variate:	R
Binomial totals:	N
Distribution:	Binomial
Link function:	Probit
Fitted terms:	Drug + LogDose

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio
Regression	4	246.090	61.5225	61.52
Residual	9	3.868	0.4298	
Total	13	249.958	19.2275	

Dispersion parameter is fixed at 1.00.

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

Estimates of parameters

Parameter	estimate	s.e.	t(*)
Drug Morphine	-1.379	0.114	-12.08
Drug Amidone	-1.141	0.120	-9.50
Drug Phenadoxone	-0.0197	0.0882	-0.22
Drug Pethidine	-2.559	0.189	-13.51
LogDose	2.468	0.173	14.30

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

Effective doses

Log10 scale

Group	LD	estimate	s.e.	lower 95%	upper 95%
Morphine	50.00	0.5587	0.02992	0.5015	0.6177
Amidone	50.00	0.4624	0.03362	0.3961	0.5267
Phenadoxone	50.00	0.0080	0.03628	-0.0649	0.0761
Pethidine	50.00	1.0367	0.02877	0.9812	1.0929

Natural scale

Group	LD	estimate	lower 95%	upper 95%
Morphine	50.00	3.620	3.173	4.147
Amidone	50.00	2.900	2.489	3.363
Phenadoxone	50.00	1.019	0.861	1.192
Pethidine	50.00	10.883	9.576	12.385

The Probit analysis setting of the Generalized Linear Models menu uses the PROBITANALYSIS procedure.

The final setting of the Generalized Linear Models menu (Figure 3.21) has boxes for you to specify the distribution and link function explicitly, so that you can fit any of the available generalized linear models.

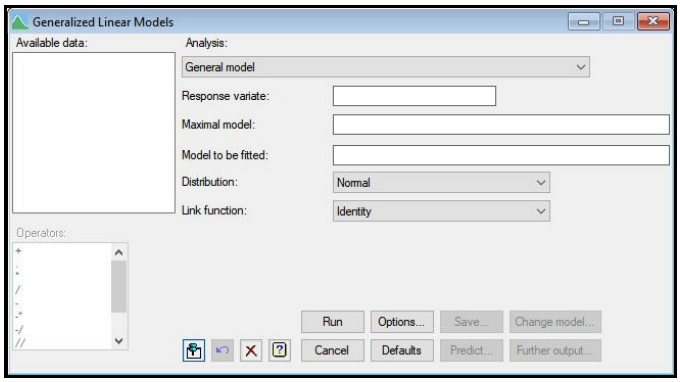


Figure 3.21

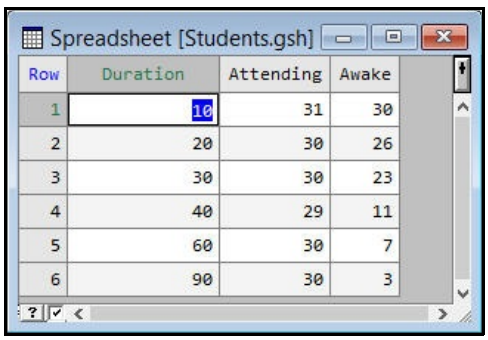
3.5 Practical

To assess the tolerance of students to statistics lectures, groups of students were subjected to talks of differing duration and were assessed just before the end to see whether they were awake or asleep.

The data are in spreadsheet file Students.gsh.

Fit a probit model to the data (taking logarithm of Duration)

Is there any evidence to suggest that some students will never fall asleep? (Hint: include natural immunity in the model.)



Row	Duration	Attending	Awake
1	10	31	30
2	20	30	26
3	30	30	23
4	40	29	11
5	60	30	7
6	90	30	3

Figure 3.22

3.6 Generalized linear mixed models

A major limitation of regression models and generalized linear models is that they cater for only one source of random, or *error*, variation. There are many situations, however, where there are several sources of error. For example, a medical trial might involve making observations on several subjects on a range of occasions. You would then need to allow for the random variation of the subjects, and perhaps also of the occasions, as well as the usual residual variation of each individual observation. Similarly, in agricultural experiments like the split-plot, where the units (plots of land) are divided into sub-units (sub-plots), the different subdivisions may all contribute to the error variation; see Section 5.1 of the *Guide to ANOVA and Design in Genstat*.

In an ordinary linear regression situation, you can handle several sources of variation by using either the analysis of variance or the REML mixed models menus or commands (see the *Guide to the Genstat Command Language, Part 2 Statistics*, Chapters 4 and 5, or the *Guide to ANOVA and Design in Genstat* and the *Guide to REML in Genstat*).

Methods for including additional sources of error variation in generalized linear models are more recent, and are still an active area of research. Genstat provides the reasonably well-established generalized linear mixed models method, described in this section, and also the more recent – and more flexible – hierarchical generalized linear models described at the end of this chapter.

Generalized linear mixed models extend the standard generalized linear models framework by allowing you to include additional random effects in the linear predictor. So the linear predictor vector becomes

$$\eta = \mathbf{X}\beta + \Sigma_j \mathbf{Z}_j v_j$$

The matrix **X** is the *design matrix* for the ordinary explanatory variables (known as the *fixed* effects), and β is their vector of regression coefficients. If the explanatory variables are variates, then **X** is a matrix whose first column contains the value one if the constant is being fitted, and whose later columns each contain the values from one of the explanatory variates. An explanatory factor would have an "indicator" column in **X** for each of its levels, with one in the units that took the level concerned, and zero elsewhere.

Similarly $\mathbf{Z}_j$ is the design matrix for the jth random term, and v_j is the corresponding vector of random effects. The random effects v_j are assumed to come from a Normal distribution with mean zero and variance σ_j^2.

As an example we consider some data from an experiment on rats (Weil 1970, *Food and Cosmetics Toxicology*). Pregnant rats were fed with either a control diet or one with an added chemical, and the numbers of live pups in the resulting litters were counted after four days and at the end of the 21-day lactation period. The data are available in the spreadsheet file Pups.gsh (Figure 3.23).

Spreadsheet [Pups.gsh]

Row	litter	time	diet	pups
1	1	4	control	13
2	1	21	control	13
3	2	4	control	12
4	2	21	control	12
5	3	4	control	9
6	3	21	control	9
7	4	4	control	9
8	4	21	control	9
9	5	4	control	8
10	5	21	control	8
11	6	4	control	8
12	6	21	control	8

Figure 3.23

The Generalized Linear Mixed Models menu (Figure 3.24) is opened by selecting the Generalized Linear Mixed Models sub-sub-option of the Mixed Models sub-option of the Regression Analysis option of the Stats menu on the menu bar.

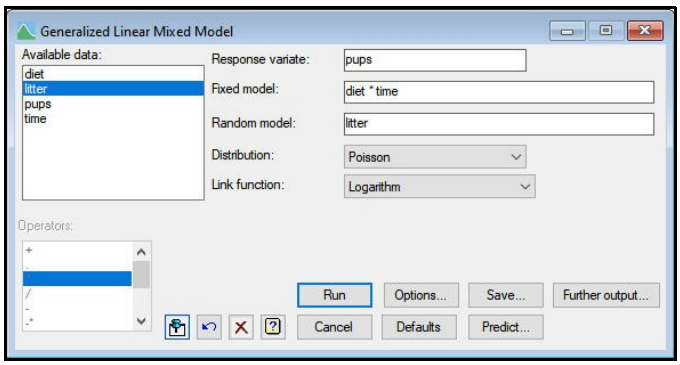


Figure 3.24

We need to fit a log-linear model (i.e. a generalized linear model with a Poisson distribution and a logarithmic link) but with an additional random effect to take account of the random variation of the litters. In the fixed model we want to look at the main effects of diet and time, and their interaction.

The Generalized Linear Mixed Models Options menu in Figure 3.25 selects the output, and controls the dispersion parameter, offset and so on, in a similar way to an ordinary generalized linear model. We have retained the default choices for output, and fixed the dispersion at 1 as usual for a Poisson distribution.

The output is shown below.

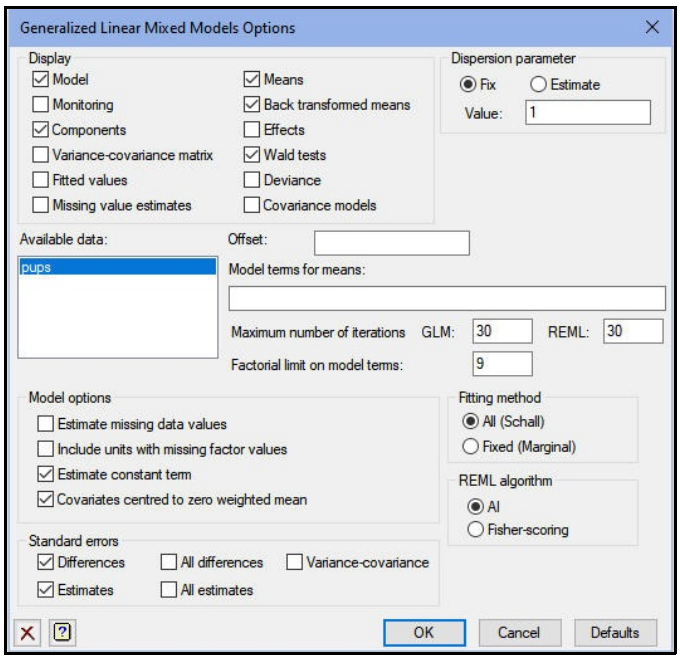


Figure 3.25

Generalized linear mixed model analysis

Method: c.f. Schall (1991) Biometrika
Response variate: pups
Distribution: poisson
Link function: logarithm
Random model: litter
Fixed model: Constant + diet*time
Dispersion parameter fixed at value 1.000

Estimated variance components

Random term	component	s.e.
litter	0.022	0.021

Residual variance model

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

Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
diet	2.22	1	2.22	29.3	0.147
time	4.28	1	4.28	58.0	0.043
diet.time	0.79	1	0.79	58.0	0.379

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
diet.time	0.79	1	0.79	58.0	0.379

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

Tables of means and back-transformed means

Table of means and back-transformed means for diet

	Means	Backtransform
diet		
control	2.234	9.339
treated	2.072	7.940

Minimum standard error	0.06859
Average standard error	0.07076
Maximum standard error	0.07294

Standard error of difference	0.1001

Table of means and back-transformed means for time

time	Means	Backtransform
4	2.244	9.434
21	2.062	7.860

Minimum standard error	0.06315
Average standard error	0.06576
Maximum standard error	0.06837
Standard error of difference	0.08543

Table of means and back-transformed means for diet.time

time	4		21	
	Means	Backtransform	Means	Backtransform
diet				
control	2.288	9.851	2.181	8.853
treated	2.201	9.035	1.943	6.979

Minimum standard error	0.08770
Average standard error	0.09293
Maximum standard error	0.10145
Minimum standard error of difference	0.1156
Average standard error of difference	0.1279
Maximum standard error of difference	0.1367

Notice that the output components are more like those of a REML analysis (see Chapter 5 of the *Guide to the Genstat Command Language, Part 2 Statistics* or the *Guide to REML in Genstat*) than those of a regression analysis. This reflects the fact that the GLMM procedure, which does the analysis, uses the REML directive to fit the fixed model.

Next we have requested Wald tests for the fixed effects. The Wald tests themselves depend on the asymptotic properties of the model. So, as in an ordinary REML analysis, they must be used with care as they tend to be optimistic; see Sub-section 5.3.6 of the *Guide to the Genstat Command Language, Part 2 Statistics* or Section 1.1 of the *Guide to REML in Genstat*. In an ordinary orthogonal analysis of variance, the Wald statistic divided by its degrees of freedom will have an F distribution, $F_{m,n}$, where m is the number of degrees of freedom of the fixed term, and n is the number of residual degrees of freedom for the fixed term. Unless the design is large or complicated, REML can estimate n, as above, and prints it in the column headed "d.d.f." (i.e. denominator degrees of freedom); m is shown the column headed "n.d.f." (i.e. numerator degrees of freedom). In other situations, the printed F statistics have approximate F distributions. These are more reliable than the Wald tests, but they should still be used with caution.

Caution is particularly necessary with binary data (i.e. binomial data where there is only a single subject in each group) or with very small data sets. The variance

components of the random terms then tend to be underestimated, and so the standard errors for the fixed effects may be too small and the Wald and F tests may be too large.

Another way to check whether a term is needed in the model is to see how deviance changes between analyses that include and exclude the term. Checking the Deviance box in the Generalized Linear Mixed Models Options menu prints the deviance from the generalized linear model. This can be used to assess the effect of omitting a term from the fixed model (while keeping the random model unchanged) or from the random model (while keeping the fixed model unchanged).

The Generalized Linear Mixed Models Further Output menu (Figure 3.26) provides two further types of deviance. When you check the Deviance box, the Method drop-down list is enabled. As well as the generalized linear model (GLM) deviance, it provides the REML deviance (Residual likelihood) which can be used to assess terms in the random model, and the deviance based on the full likelihood from the REML analysis (Full likelihood) which can be used to assess terms in either the fixed or the random model. The boxes also become available to print the Akaike and Schwarz Bayesian information coefficients, and the numbers of fixed and random degrees of freedom (and we will print these too).

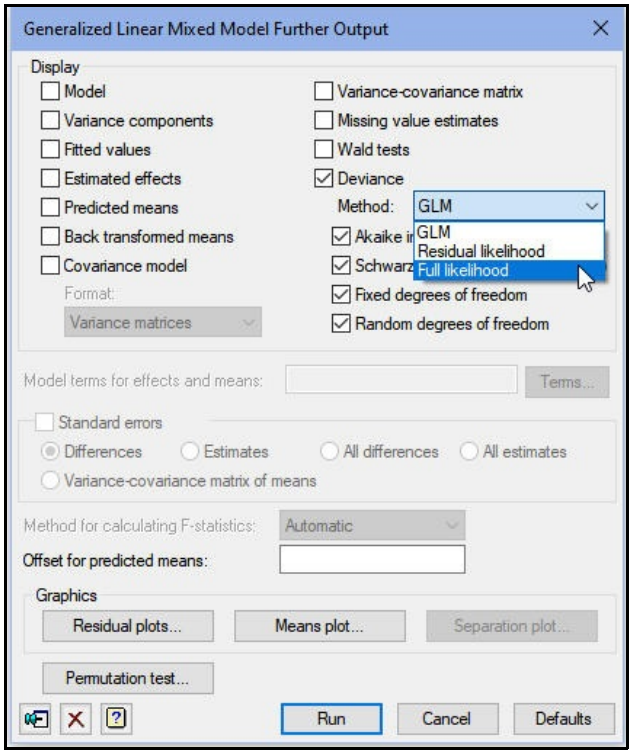


Figure 3.26

The output below shows the deviance and information coefficients based on the full likelihood, together with the degrees of freedom.

Generalized linear mixed model analysis

Deviance and information criteria from REML analysis

Deviance	-68.51
Akaike information coefficient	-77.58
Schwarz Bayes information coefficient	-66.78
d.f. of fixed model	4
d.f. of random model	1

(based on the full log-likelihood)

The deviance and the information coefficients omit constants that are unaffected by changes in the models. So they cannot be used to assess the general lack of fit. They should be used only to compare different models. So we change the fixed model to `diet+time` in the Generalized Linear Mixed Models menu, click on Run to redo the analysis, and use the Generalized Linear Mixed Models Further Output menu to print deviance and the information coefficients for the analysis excluding the `diet.time` interaction.

Generalized linear mixed model analysis

Method:	c.f. Schall (1991) Biometrika
Response variate:	pups
Distribution:	poisson
Link function:	logarithm
Random model:	litter
Fixed model:	Constant + diet + time

Dispersion parameter fixed at value 1.000

Estimated variance components

Random term	component	s.e.
litter	0.022	0.021

Residual variance model

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

Tests for fixed effects

Sequentially adding terms to fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
diet	2.42	1	2.42	29.2	0.130
time	4.30	1	4.30	59.0	0.043

Dropping individual terms from full fixed model

Fixed term	Wald statistic	n.d.f.	F statistic	d.d.f.	F pr
diet	2.42	1	2.42	29.2	0.130
time	4.30	1	4.30	59.0	0.043

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

Tables of means and back-transformed means

Table of means and back-transformed means for diet

diet	Means	Backtransform
control	2.232	9.316
treated	2.076	7.976

Minimum standard error	0.06862
Average standard error	0.07060
Maximum standard error	0.07258
Standard error of difference	0.09975

Table of means and back-transformed means for time

time	Means	Backtransform
4	2.242	9.414
21	2.066	7.892

Minimum standard error	0.06319
Average standard error	0.06561
Maximum standard error	0.06804
Standard error of difference	0.08507

Generalized linear mixed model analysis

Deviance and information criteria from REML analysis

Deviance	-71.93
Akaike information coefficient	-79.46
Schwarz Bayes information coefficient	-70.83
d.f. of fixed model	3
d.f. of random model	1

(based on the full log-likelihood)

The difference between the deviances is 3.42 (71.93−68.51) on 1 degree of freedom. This is non-significant, but closer to the 3.84 value for significance at 5%. Remember, though, that the algorithms for fitting generalized linear mixed models by REML involve several approximations. So the REML deviances should also be used with care.

The final way to assess terms in the fixed model is to do a permutation test, and this should not be subject to any biases. The menu is opened by clicking on the Permutation test button on the Generalized Linear Mixed Models Further Output menu (Figure 3.26). First, though, we need to redo the analysis with the interaction included in the fixed model, as in Figure 3.24.

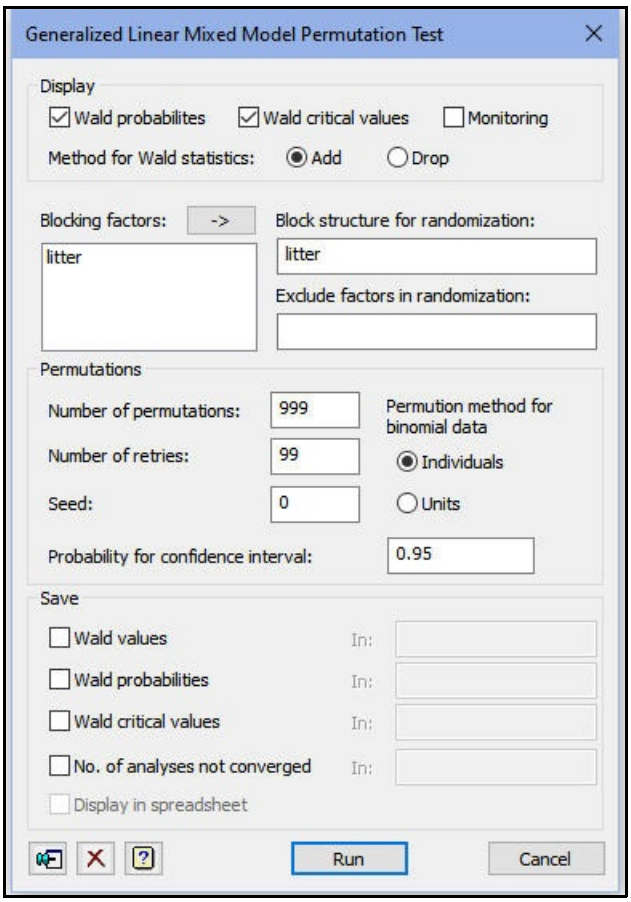


Figure 3.27

In the menu (Figure 3.27) we have chosen to display probabilities based on the Wald statistics calculated in the analyses of the permuted data sets, as well as critical values based on their distribution. These are to be Wald statistics from the sequential adding of the individual fixed terms into the model, so that we can obtain probabilities for the main effects as well as the interaction. The design has diets applied to the different litters. The times factor changes within the litters, and has equally replicated levels. So we can (and should) use a block structure of litters (and rats within the litters) for the random permutations. We have left the seed at its default of zero. As this is the first use of random numbers in this job, Genstat picks a seed at random (here 377901) using the computer's clock. The default on the menu is to do 99 permutations. However, as the analysis of this data set is not too time-consuming, we have increased this to 999 in order to provide a more reliable result. We have retained the default of 99 retries to replace any permuted data sets with unsuccessful analyses and, in fact, here there are none.

Message: Default seed for random number generator used with value 377901

Probabilities for Wald statistics

Source	
diet	0.138
time	0.001
diet.time	0.128

(determined from 999 random permutations)

Critical values for Wald statistics

Source	5%	1%	0.1%
diet	4.413	7.647	13.365
time	1.306	1.720	2.340
diet.time	1.167	1.820	2.482

The interaction has a probability of 0.128, reinforcing the conclusion that there is little evidence of an interaction between diet and time, nor is there any evidence for differences between the diets.

3.7 Practical

Spreadsheet file `Clinical.gsh` contains data from a multicentre randomized clinical trial (Beitler & Landis 1985, *Biometrics*). In each of eight centres, a group of patients was given a cream containing a control treatment and another group was given another cream containing an active drug to control an infection. The variate `Total` records the number of patients in each group, and the variate `Favorable` records the number that produced a favourable response.

Analyse the data as a generalized linear mixed model, treating the effects of the different clinics as a random effect.

Spreadsheet [Clinical.gsh]

Row	Clinic	Treatment	Favorable	Total
1	1	drug	11	36
2	1	control	10	37
3	2	drug	16	20
4	2	control	22	32
5	3	drug	14	19
6	3	control	7	19
7	4	drug	2	16
8	4	control	1	17
9	5	drug	6	17
10	5	control	0	12
11	6	drug	1	11
12	6	control	0	10
13	7	drug	1	5
14	7	control	1	9
15	8	drug	4	6
16	8	control	6	7

Figure 3.28

3.8 Hierarchical generalized linear models

Hierarchical generalized linear models (HGLMs) provide another way of modelling non-Normal data when there are several sources of error variation. Like generalized linear mixed models, they extend the familiar generalized linear models to include additional random terms in the linear predictor. However, they do not constrain these additional terms to follow a Normal distribution nor to have an identity link, as is the case in a generalized linear mixed model. They thus provide a much richer set of models, that may seem more intuitively appealing. The methodology provides improved estimation methods that reduce bias, by the use of the exact likelihood or extended Laplace approximations. In particular, the Laplace approximations seem to avoid the biases that

are often found when binary data are analysed by generalized linear mixed models.

So the linear predictor vector again becomes

$$\eta = \mathbf{X}\beta + \Sigma_j \mathbf{Z}_j v_j$$

and the response vector y still has a distribution from the exponential family. However, this is limited to binomial, gamma, Normal or Poisson. (These do cover all the most common generalized linear models though.) The additional random terms now have their own link functions

$$v_i = \mathrm{v}(u_i)$$

where the vectors of random effects u_i have beta, Normal, gamma or inverse gamma distributions. (These are distributions that are *conjugate* to the distributions available for y; for details see Lee, Nelder & Pawitan 2006, *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, CRC Press.)

The analysis involves fitting an extended generalized linear model, known as the *augmented mean model*, to describe the mean vector μ. This has units corresponding to the original data units, together with additional units for the effects of the random terms. The augmented mean model is fitted as a generalized linear model, but there may be different link functions and distributions operating on the additional units from those on the original units. The link function is the function v(), while the distribution is the one to which the distribution of u_i is conjugate; see Lee, Nelder & Pawitan (2006) Chapter 6 for details. The data values for the extra units contain the inverse-link transformations of the expected values of the random distributions. Further generalized linear models, with gamma distributions and usually with logarithmic links, model the dispersion for each random term (including the residual dispersion parameter). The models are connected, in that the y-variates for the dispersion models are deviance contributions from the augmented mean model divided by one minus their leverages, while the reciprocals of the fitted values from the dispersion models act as weights for the augmented mean model. So the models are fitted alternately until convergence, as shown in Table 7.3 of Lee, Nelder & Pawitan (2006).

The methodology has been implemented in Genstat, as a suite of procedures, and there are data files and programs to run many of the worked examples from Lee, Nelder & Pawitan (2006).

The example programs are accessed by selecting the Analysis Programs sub-option of the Examples option of the Help menu on the menu bar. The Filter by topic drop-down list box allows you to display only the Lee, Nelder & Pawitan examples, as shown in Figure 3.29.

The use of the procedures is explained in Section 3.5.11 of the *Guide to the Genstat Command Language, Part 2 Statistics.*

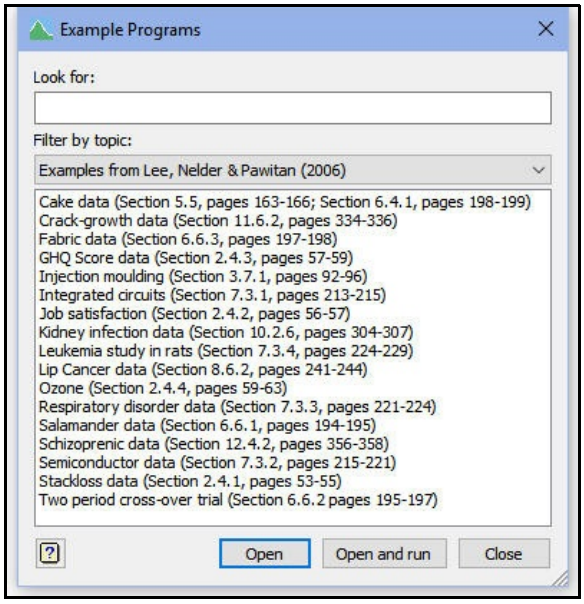


Figure 3.29

However, you do not need to know the details of the methodology to fit HGLMs, and the commands are needed only for the more advanced features.

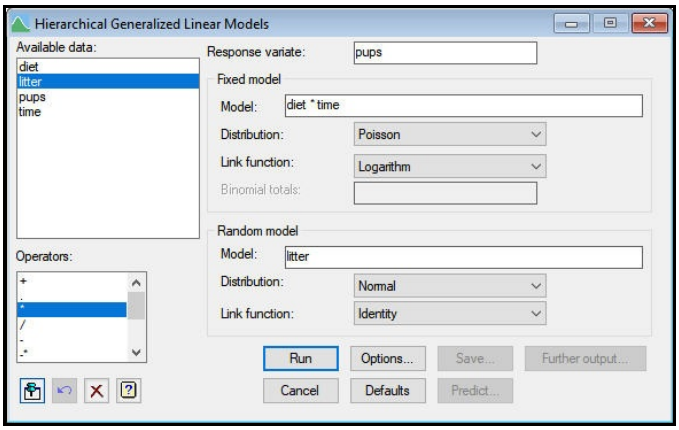


Figure 3.30

Instead you can open the Hierarchical Generalized Linear Models menu (Figure 3.30) by selecting the Hierarchical Generalized Linear Models sub-sub-option of the Mixed Models sub-option of the Regression Analysis option of the Stats menu on the menu bar. Here we reanalyse the data on rat litters from Section 3.6.

The output is controlled by the Hierarchical Generalized Linear Models Options menu (Figure 3.31). Here we have changed the default settings to print Wald tests, and to fix the dispersion parameter at 1 as the data have a Poisson distribution.

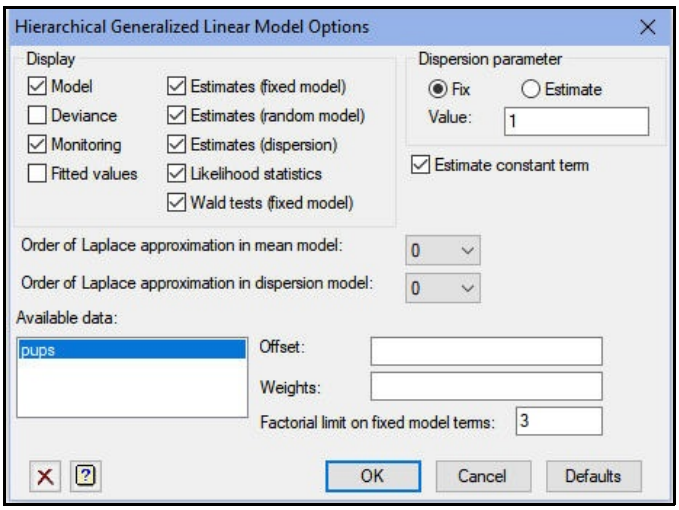


Figure 3.31

Monitoring

```
                cycle no., disp. components & max. absolute change
                  2          -2.933          0.6133
                  3          -3.240          0.3073
                  4          -3.425          0.1847
                  5          -3.544          0.1190
                  6          -3.625         0.08096
Aitken extrapolation OK
                  7          -3.806          0.1809
                  8          -3.812        0.006254
                  9          -3.816        0.004628
                 10          -3.820        0.003430
Aitken extrapolation OK
                 11          -3.830        0.009839
                 12          -3.830      0.00001578
```

Hierarchical generalized linear model

Response variate: pups

Mean model

Fixed terms: diet*time
Distribution: poisson
Link: logarithm
Random terms: litter
Distribution: normal
Link: identity
Dispersion: fixed

Dispersion model

Distribution: gamma
Link: logarithm

Estimates from the mean model

Estimates of parameters

Parameter	estimate	s.e.	t(*)
constant	2.2875	0.0877	26.08
diet treated	-0.086	0.126	-0.68
time 21	-0.107	0.116	-0.92
diet treated .time 21	-0.151	0.171	-0.89
litter 1	0.111	0.124	0.89
litter 2	0.081	0.125	0.65
litter 3	-0.011	0.126	-0.09
litter 4	-0.011	0.126	-0.09
litter 5	-0.042	0.126	-0.33
litter 6	-0.042	0.126	-0.33
litter 7	0.096	0.124	0.77
litter 8	0.066	0.125	0.53
litter 9	0.005	0.126	0.04
litter 10	0.005	0.126	0.04
litter 11	-0.026	0.126	-0.21
litter 12	0.081	0.125	0.65
litter 13	-0.153	0.128	-1.19
litter 14	-0.105	0.127	-0.82
litter 15	-0.026	0.126	-0.21
litter 16	-0.026	0.126	-0.21
litter 17	0.127	0.126	1.00
litter 18	0.095	0.127	0.75
litter 19	0.064	0.127	0.50
litter 20	0.032	0.128	0.25
litter 21	0.080	0.127	0.63
litter 22	0.048	0.128	0.37
litter 23	0.048	0.128	0.37
litter 24	0.016	0.128	0.12
litter 25	0.016	0.128	0.12
litter 26	-0.115	0.130	-0.88

litter 27	0.000	0.128	0.00
litter 28	-0.082	0.129	-0.63
litter 29	-0.016	0.129	-0.13
litter 30	-0.115	0.130	-0.88
litter 31	-0.049	0.129	-0.38
litter 32	-0.148	0.130	-1.13

Parameters for factors are differences compared with the reference level:

Factor	Reference level
diet	control
time	4
litter	1

Estimates from the dispersion model

Estimates of parameters

Parameter	estimate	s.e.	t(*)	antilog of estimate
lambda litter	-3.830	0.494	-7.76	0.02172

Likelihood statistics

-2 × h(y|v)	290.286
-2 × h	234.755
-2 × P_v(h)	308.741
-2 × $P_{\beta,v}$(h)	320.460
-2 × EQD(y|v)	290.317
-2 × EQD	234.786
-2 × P_v(EQD)	308.772
-2 × $P_{\beta,v}$(EQD)	320.491

Fixed parameters in mean model	4
Random parameters in mean model	32
Fixed dispersion parameters	1
Random dispersion parameters	0

Wald tests for dropping HGLM fixed terms

Term	Wald statistic	d.f.	approx. pr.
diet.time	0.7858	1	0.375

The output is more like the output from a regression, as you would expect as the algorithm involves fitting a generalized linear model to describe the mean μ and another to estimate the dispersion parameters. The dispersion estimates are of the logarithms (base e) of the variance components. So the exponentials of the HGLM estimates should correspond to those in the generalized linear mixed models analysis. Here we have an estimate of −3.830 for litters. Its exponential is 0.022, which is the same as the GLMM estimate.

The likelihood statistics allow you to assess the various components of the model. Changes in the fixed model can be assessed using changes in `-2*P_v(h)`; changes in the dispersion models are assessed using `-2*P_beta,v(h)`; and `-2*h(y|v)` could be used if you wanted to form the deviance information coefficient (DIC). The EQD statistics are approximations to the first four, (h-likelihood) statistics, calculated using quasi-likelihood instead of exact likelihood. There are two procedures `HGFTEST` and `HGRTEST` that can use these statistics to perform tests to see if terms can be dropped from the fixed and random models. More information, and an example, is given in Section 3.5.11 of the *Guide to the Genstat Command Language, Part 2 Statistics*.

The Wald tests provide another, quicker way of seeing whether terms can be dropped from fixed model. They are less accurate than the likelihood tests performed by `HGFTEST`. Here, though, the conclusion is clear – that we do not need the diet-by-time interaction. So we redo the analysis, with a fixed model containing just the main effects (see Figure 3.32). To save space we also modify the options menu to omit the monitoring information and the estimates from the random model (i.e. the litter effects).

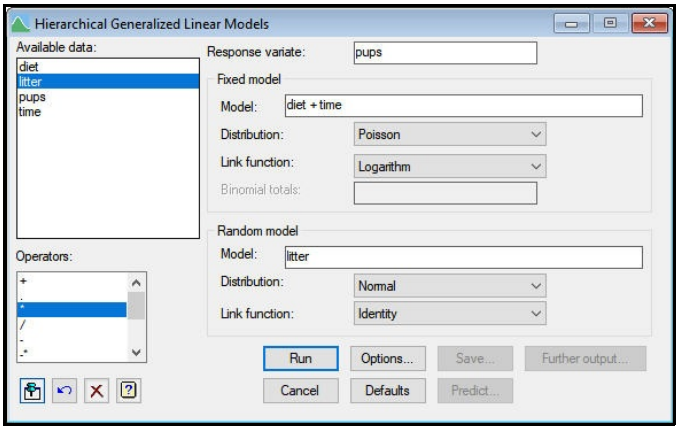


Figure 3.32

Hierarchical generalized linear model

Response variate: pups

Mean model

Fixed terms: diet + time
Distribution: poisson
Link: logarithm
Random terms: litter
Distribution: normal
Link: identity
Dispersion: fixed

Dispersion model

Distribution: gamma
Link: logarithm

Estimates from the mean model

	estimate	s.e.	t(*)
constant	2.31990	0.07874	29.46
diet treated	-0.15531	0.09975	-1.56
time 21	-0.17640	0.08507	-2.07

Estimates from the dispersion model

Estimates of parameters

Parameter	estimate	s.e.	t(*)	antilog of estimate
lambda litter	-3.830	0.494	-7.76	0.02171

Likelihood statistics

-2 × h(y|v)	291.073
-2 × h	235.541
-2 × P_v(h)	309.528
-2 × $P_{\beta,v}$(h)	319.552
-2 × EQD(y|v)	291.104
-2 × EQD	235.572
-2 × P_v(EQD)	309.558
-2 × $P_{\beta,v}$(EQD)	319.583

Fixed parameters in mean model	3
Random parameters in mean model	32
Fixed dispersion parameters	1
Random dispersion parameters	0

Wald tests for dropping HGLM fixed terms

Term	Wald statistic	d.f.	approx. pr.
diet	2.424	1	0.119
time	4.300	1	0.038

3.9 Practical

Reanalyse the data in the spreadsheet file `Clinical.gsh` using the hierarchical generalized linear models menus.

4
Other facilities

This Guide illustrates menus from the main regression analyses. Other menus are listed below with references to sections in the *Guide to the Genstat Command Language, Part 2 Statistics* or to procedures in the *Genstat Reference Manual, Part 3*, describing the associated commands and methodology:

Ordinal Regression	Section 3.5.5,
Split-line Regression	Procedure R2LINES,
Lasso Regression	Procedure RLASSO,
Response Surface	Procedures RQUADRATIC and VSURFACE,
Regression Tree	Section 3.9,
Quantile Regression	Section 3.10,
Nonlinear Quantile Regression	Procedure RQNONLINEAR, and
Linear Functional Relationship	Procedure RLFUNCTIONAL.

Chapter 3 of the *Guide to the Genstat Command Language, Part 2 Statistics* gives more details of the statistical methodology, and describes the most important commands.

Index