

## 5 Univariate techniques

In Zuur et al. (2007), various univariate statistical methods were discussed, namely linear regression, generalised linear modelling (GLM), additive modelling, generalised additive modelling (GAM), regression and classification trees, mixed effects modelling, generalised least squares (GL) and additive mixed effects modelling. More examples on (additive) mixed effects modelling (and extensions) can be found in Zuur et al. (2009). Their Appendix A contains a detailed example of linear regression.

In this chapter, we show how to apply these techniques in Brodgar. In Section 5.1, we show how to apply a linear regression analysis in Brodgar, and Section 5.2 tells you have to reproduce some of the analyses presented in Zuur et al. (2007; 2009). Applying any of the other univariate techniques in Brodgar follows the same clicking process, and we therefore *strongly* advise that you read these two sections, before reading the remaining text.

In Sections 5.3 and 5.4, we show how to apply GLM in Brodgar, and GAM is discussed in Section 5.5 and 5.6. Running regression and classification techniques in Brodgar are explained in Section 5.7. Mixed effects models are discussed in Section 5.8, and GLS (for detailing with heterogeneity) in Sections 5.9 and 5.10. Finally, adding temporal or spatial correlation structures to additive models and GLS is explained in Section 5.11.

All methods discussed in this chapter make use of R.

### 5.1 Linear regression in Brodgar

Linear regression is explained in Sections 5.1 and 5.2 in Zuur et al. (2007), and in Appendix A of Zuur et al. (2009). We start showing how to do linear regression in Brodgar, discuss some of the options, present output for a particular data set and mention further options.

#### 5.1.1 A simple linear regression analysis

In this section, we will use the Argentinean zoobenthic data; see Chapters 4 and 28 in Zuur et al. (2007) for a biological explanation. The data can be found in the Excel file [www.brodgar.com/Argentina.xls](http://www.brodgar.com/Argentina.xls). The four species *L. acuta*, *H. similis*,

*U. uruguayensis* and *N. succinea* are the response variables and the other variables are explanatory variables. Once the data are imported, click on the “Univariate” main menu button (Figure 3.1). The window in Figure 5.1 appears.

The Argentinean zoobenthic data set consists of 4 species measured at 30 sites during Autumn and Spring of 1997. And there are 4 sediment variables. Results in Chapter 4 in Zuur et al. (2007) showed that there is collinearity between the sediment variables. Therefore, we only use medium sand and mud as continuous explanatory variables, and time (season) and transect as nominal variables. Clicking the “Go” button in Figure 5.1 gives Figure 5.2; the window for linear regression. In the first tab labelled “Variables” one can select the response and explanatory variables. It is also possible to use weights. Each of these points is discussed next.

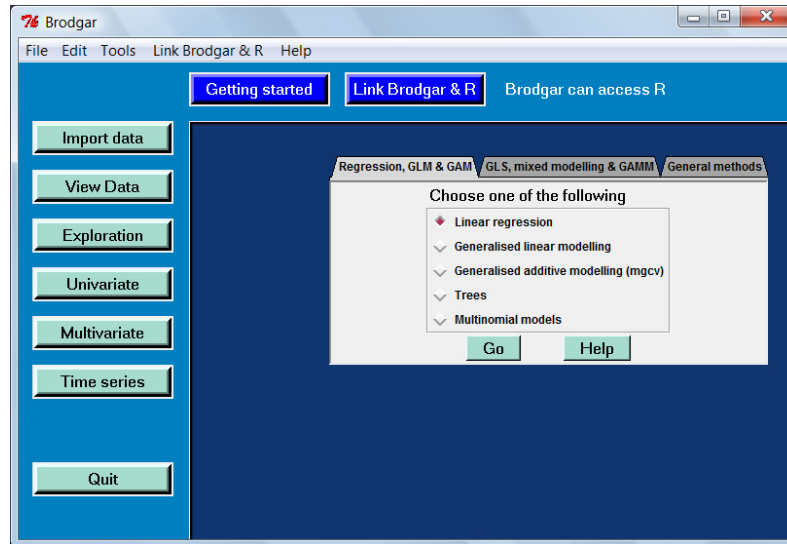


Figure 5.1. Univariate analysis menu. Clicking the “Go” button gives Figure 5.2.

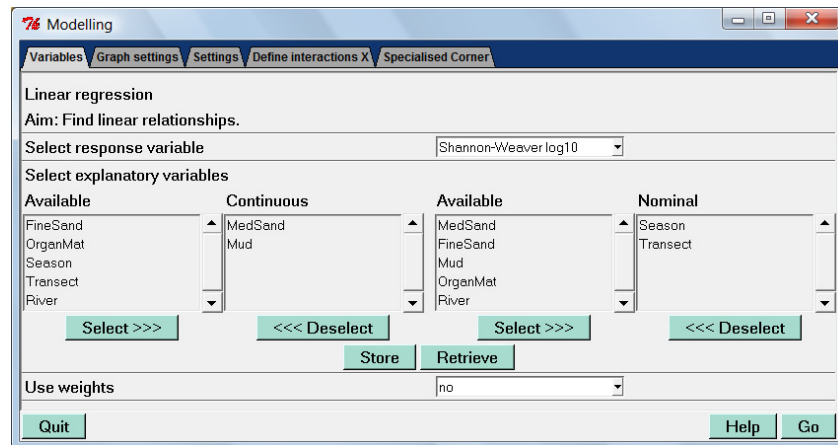


Figure 5.2. Selection of a response variable (top row) and explanatory variables in linear regression. The “Continuous” box is for continuous explanatory variables and the “Nominal” box for nominal (categorical) explanatory variables. Choosing the variables the wrong way around, may cause error messages. But even if no error message is generated, ensure you select the right variables in the appropriate box!

### 5.1.2 Select response variable

Here, one has to choose a response variable. In this case, the options are: *L. acuta*, *H. similes*, *U. uruguayensis* and *N. succine*. Brodgar also allows the user to select a diversity index as response variable. These indices can be selected in the “Select response variable” box (they are enumerated above the response variables). If one of these diversity indices is selected, Brodgar will base the calculation of the diversity index on all response variables (as selected in the data import process). We decided to use the Shannon-Weaver diversity function (with base 10).

### 5.1.3 Select explanatory variables and nominal variables

In this step, one has to select explanatory variables. All available explanatory variables (as selected in the data import process) are enumerated. If no explanatory variables are selected, a linear regression model with only an intercept is fitted. We selected medium sand and mud as continuous explanatory variables, and season and transect as nominal explanatory variables.

Our selections are shown in Figure 5.2. Clicking the “Go” button in Figure 5.2 applies the linear regression model. Before presenting the output, we discuss a few things.

### 5.1.4 Use weights

The following information was taken from the R help file (start R and type: `?lm`):

```
weights: an optional vector of weights to be used in
the fitting process. If specified, weighted least
squares is used with weights 'weights' (that is, mini-
mizing 'sum(w*e^2)'); otherwise ordinary least squares
is used.
```

If weights are to be used in Brodgar, the user should first select the weighting variable as an explanatory variable in the data import process, and then select it in the box for “Use weights” in Figure 5.2. Please make sure that the variable defining the weights does not contain missing values, negative values or zeros!

### 5.1.5 Graphs settings and general settings

In the tab “Graph settings” (Figure 5.3), one can specify the title, *x*-label, *y*-label, graph type, the name of the graph, and which validation graphs should be plotted. Brodgar and R will save the graphs in your project directory under the default name: `graph1.jpg`, but this can easily be changed. Graphs can also be exported to WMF or EPS format, which can be imported into Word or WordPerfect (see Chapter 4 of this manual for further details).

Possible validation graphs are the default graph, a plot of the residuals versus the fitted values, a QQ-plot of the standardised residuals, square root transformed standardised residuals versus fitted values (scale plot), Cook’s distances, hat (or leverage) values for each observation, the change in the fitted values if the  $i^{\text{th}}$  observation is omitted, a histogram of the residuals, partial plots (these are discussed later in this chapter), and the change in regression parameters if the  $i^{\text{th}}$  observation is omitted. Studentised and standardised residuals can also be plotted versus the fitted values. If only one explanatory variable is selected, Brodgar will automatically plot the fitted values and observed values in one graph.

In the “Settings” panel (lower window in Figure 5.3), one can choose to save the residuals or fitted values (we discuss later how to access them). Important model validation graphs are residuals versus each explanatory variable, and it is optional to add a smoother to these plots to aid visual interpretation of the graphs.

Automatic stepwise model selection can be selected by changing the “no” for “Forward/backward selection”. Possible options are “backwards”, “forward”, “backwards and forwards selection”. All these options apply stepwise linear regression using the AIC.

It is also possible to drop one explanatory variable in turn, and apply an *F*-test. This is executing the function `drop1` from R. Note that this is not doing sequential testing! It is a useful tool to obtain one *p*-value for a nominal variable with more than two levels. To select this option, change the “no” for “Forward/backward selection” into “drop1”.

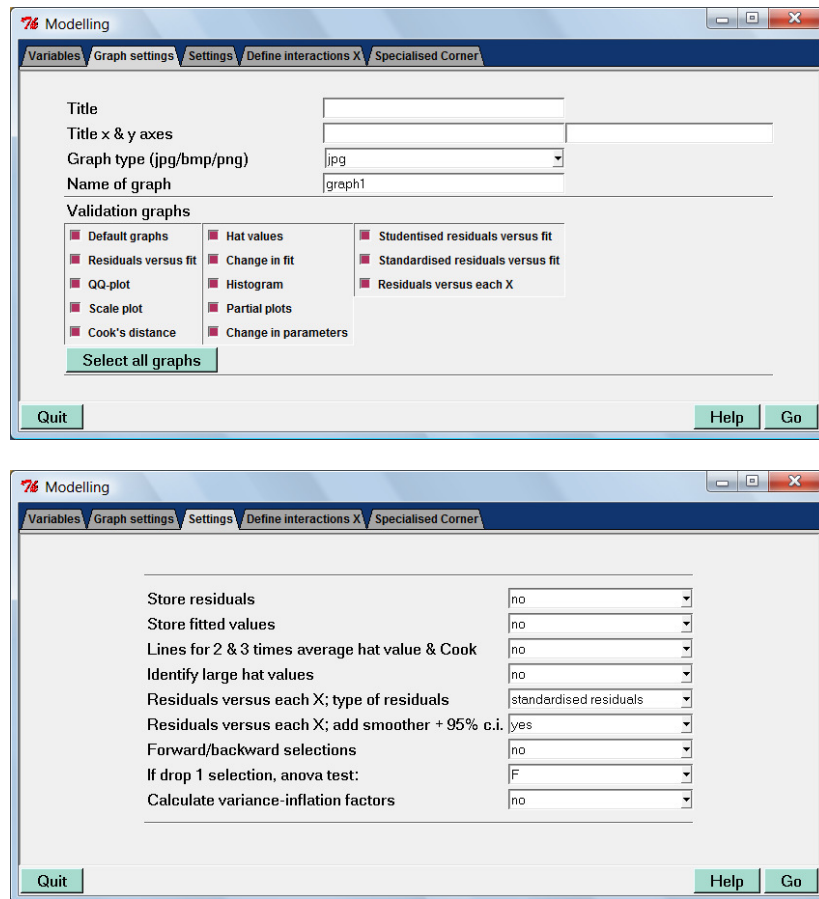


Figure 5.3. More options for linear regression; graph settings, and (general) settings. Newer versions of Brodgar may have a slightly different arrangement of the options.

### 5.1.6 Results linear regression

Clicking the “Go” button in Figure 5.2 results in the window in Figure 5.4. This graph is the default output of the linear regression function in R and Brodgar. The upper left graph in Figure 5.4 shows a plot of the fitted values versus residuals. Look for any pattern, and specifically for an increase (or decrease) in spread of residuals for increasing values of the fitted values, as this would indicate violation of the homogeneity assumption (in which case a transformation or application of GLM should be considered). The upper right panel shows the QQ-plot of the standardised residuals. If all points are approximately on a straight line, then normality of the residuals can be assumed. The lower right panel contains a scatterplot of

leverage versus residuals, with contour lines for the Cook's distance values. It is also possible to obtain a graph with only the Cook distances; just press a couple of times on the “Next graph” button. Cook values larger than 1 indicate influential observations. The lower left panel in Figure 5.4 contains a scatterplot of square root transformed standardised residuals versus the fitted values; see Chapter 5 in Zuur et al. (2007) for an explanation.

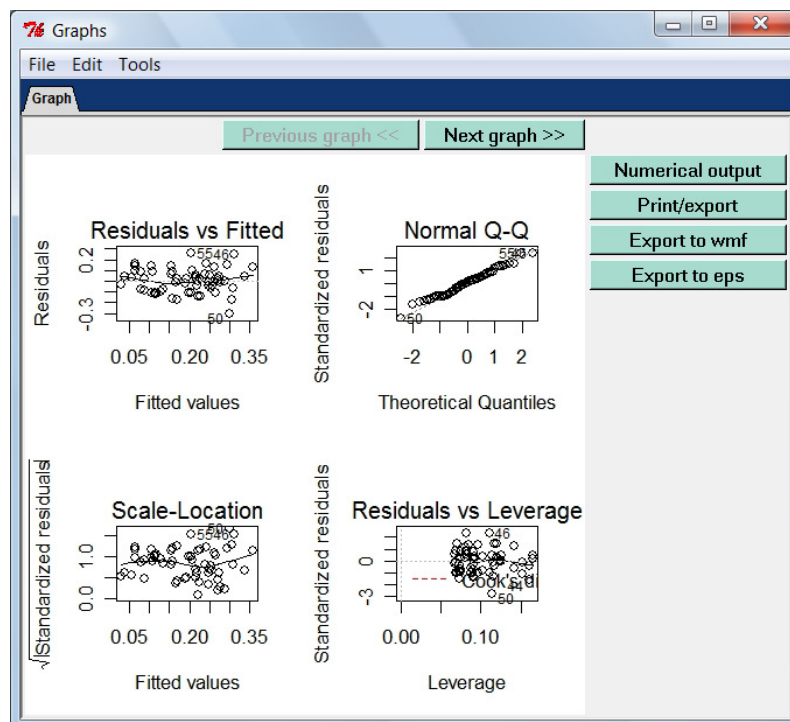


Figure 5.4. Graphical output from a linear regression analysis.

### **Numerical output**

Numerical information can be obtained from the menu in Figure 5.4; just click the “Numerical output” button. A text editor will open with the following information:

```
#####
#### LINEAR REGRESSION NUMERICAL OUTPUT ####
#####
Model is given by f1:
Y1 ~ 1 + MedSand + Mud + as.factor(Season) + as.factor(Transect)

Call:
lm(formula = f1, data = dataz, weights = XW, na.action = na.omit)

Residuals:
      Min       1Q   Median       3Q      Max
-0.299305 -0.091526  0.002258  0.072666  0.257453

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.124396    0.104771  -1.187   0.2403
MedSand         0.001274    0.014171   0.090   0.9287
Mud             0.007269    0.003551   2.047   0.0455 *
as.factor(Season)2  0.049265    0.030159   1.633   0.1082
as.factor(Transect)2 0.191408    0.041030   4.665 2.07e-05 ***
as.factor(Transect)3 0.139169    0.040354   3.449   0.0011 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1168 on 54 degrees of freedom
Multiple R-Squared: 0.3589, Adjusted R-squared: 0.2996
F-statistic: 6.047 on 5 and 54 DF, p-value: 0.0001632

Analysis of Variance Table
Response: Y1
              Df Sum Sq Mean Sq F value    Pr(>F)
MedSand       1  0.00815  0.00815    0.5972   0.4430
Mud           1  0.01955  0.01955    1.4330   0.2365
as.factor(Season) 1  0.03641  0.03641    2.6683   0.1082
as.factor(Transect) 2  0.34841  0.17420  12.7680 2.88e-05 ***
Residuals     54  0.73677  0.01364
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
No weights were used
```

The linear regression model we are fitting here is of the form:

$$H_i = \alpha + \beta_1 \times \text{MedSand}_i + \beta_2 \text{Mud}_i + \text{factor(Season)}_i + \text{factor(Transect)}_i + \varepsilon_i$$

where  $i = 1, \dots, 60$ ,  $H_i$  the Shannon-Weaver index at observation  $i$  and  $\varepsilon_i$  are the residuals. Estimated values, standard errors,  $t$ -values and  $p$ -values are given above and the reader is referred to Zuur et al. (2007) for a statistical explanation. The only point that perhaps needs some explanation is the way Brodgar and R deal with nominal variables. The function “factor(Season)” makes sure that R knows that Season is a nominal variable. The first level of Season (Autumn) is set to zero, and the second level (Spring) is denoted by `as.factor(Season)2`. Its estimated value is 0.049265. The same is done for transect; the first level (transect A) is set to 0, `as.factor(Transect)2` represents transect B, and `as.factor(Transect)3` are the observations from transect C. The first level is set to 0 to avoid collinearity (if an observation is not from transect B or C, it must be from transect A). As a result, the intercept represents the intercept itself, but also level 1 of transect (transect A) and level 1 of season (Autumn).

The interpretation of the nominal variables is as follows. If a sample is from transect B, its fitted value will be 0.191408 higher compared to a sample from transect A. A sample from transect C has a fitted value that is 0.139169 higher than a sample from transect A. Similar statements can be made for samples from transect B and C in Spring.

Results indicate that mud is weakly significant at the 5% level, and the two transect levels are significantly different from the baseline level. Season and medium sand are not significantly related to the index function. The measure of fit  $R^2$  is relative small; the model explains around 35% of the variation in the data.

The second part of the numerical output contains the ANOVA table, and it shows the  $F$ -value (and  $p$ -value) for adding an extra explanatory variable. Hence, adding mud to a linear regression model that already contains medium sand does not significantly improve the model. But adding transect to a model that already contains medium sand, mud and season does result in a significant model improvement. The problem with this table is that it depends on the order of how the explanatory variables were imported. It may be better to drop only one explanatory variable from the full model, and apply an  $F$ -test (see the next paragraph).

As explained in Appendix A of Zuur et al. (2009), there are multiple options for model selection. The first option we illustrate is hypothesis testing using the drop1 function, and the second option is the AIC.

### **Drop 1**

Hypotheses testing procedures can be used to drop non-significant explanatory variables. The drop1 option applies the model with all selected explanatory variables, and in turn, drops one explanatory variable and applies an  $F$ -test. The results are also presented in an ANOVA table, but it is crucial to realise that these test results do not depend on the order of the variables.

To run this test in Brodgar, close the window with the numerical output, and also close the window with graphical output (Figure 5.4). Go to the window with “Settings” in Figure 5.3, and for “Forward/backward selection” change the “no” to “drop1”; ensure that on the next line the  $F$ -test is selected (for GLMs with no dispersion parameter this should be the Chi-square statistic). Click on the “Go” button, and again go to the numerical output. At the end of the window with numerical output, you will find:



```
#####
####   OUTPUT OF SELECTION PROCEDURE   ####
The selection procedure cannot cope with missing values.
If you have missing values, re-apply linear regression but
remove rows with missing values.

Single term deletions

Model:
Yl ~ 1 + MedSand + Mud + as.factor(Season) + as.factor(Transect)
              Df Sum of Sq      RSS      AIC F value    Pr(>F)
<none>                0.737 -251.990
MedSand              1 0.0001104    0.737 -253.981    0.0081 0.92867
Mud                  1    0.057    0.794 -249.505    4.1903 0.04554
as.factor(Season)    1     0.036    0.773 -251.096    2.6683 0.10819
as.factor(Transect)  2     0.348    1.085 -232.756   12.7680 2.88e-05
```

The advantage of this test is that it gives one  $p$ -value for the categorical variable Transect. Note that the  $p$ -values for MedSand and Mud are identical to the ones obtained by the  $t$ -statistic; this is not necessarily the case for GLMs.

### Using the AIC

We now show how to apply and interpret the second option; backwards selection using the AIC. Close the window with the numerical output, and also close the window with graphical output (Figure 5.4). Go to the window with “Settings” in Figure 5.3, and for “Forward/backward selection” change the “no” to “backwards”. The end if the numerical output file now reads:

```
#####
####   OUTPUT OF SELECTION PROCEDURE   ####
The selection procedure cannot cope with missing values.
If you have missing values, re-apply linear regression but
remove rows with missing values.

Start:  AIC=-251.99
Yl ~ 1 + MedSand + Mud + as.factor(Season) + as.factor(Transect)

              Df Sum of Sq      RSS      AIC
- MedSand      1 0.0001104    0.737 -253.981
<none>                0.737 -251.990
- as.factor(Season)  1     0.036    0.773 -251.096
- Mud            1     0.057    0.794 -249.505
- as.factor(Transect)  2     0.348    1.085 -232.756

Step:  AIC=-253.98
Yl ~ Mud + as.factor(Season) + as.factor(Transect)

              Df Sum of Sq      RSS      AIC
<none>                0.737 -253.981
- as.factor(Season)  1     0.036    0.773 -253.087
- Mud              1     0.066    0.803 -250.803
- as.factor(Transect)  2     0.366    1.103 -233.782
```

```

Call:
lm(formula = Yl ~ Mud + as.factor(Season) + as.factor(Transect),
    data = datazz, weights = XW, na.action = na.omit)

Coefficients:
            (Intercept)              Mud      as.factor(Season)2
as.factor(Transect)2  as.factor(Transect)3
-0.119708              0.007137              0.049265
0.192788              0.138020

```

The first step of the backward selection routine indicates that medium sand should be dropped first from the model; it has the lowest AIC. The optimal model, as selected by the AIC, contains mud, season and transect. You now need to close the windows with numerical output and graphical output, remove medium sand from the list of selected explanatory variables in Figure 5.2, and refit the model. The numerical output of this model shows that it is optional to remove season as it is not significant at the 5% level (the AIC tends to be conservative).

You should now apply a graphical model validation; inspect the residuals for homogeneity, normality and independence. All necessary graphs can be obtained by clicking on the “Next graph” button in Figure 5.4. The important graphs are residuals versus fitted values (upper left panel in the first graph, or the second graph), independence (last graphs), and normality (upper right panel in the first graph or the histogram in graph 8).

### 5.1.6 Interactions

It is possible to specify interactions between continuous variables, between continuous and nominal variables, and between nominal variables. To define interactions between for example mud and Transect, click on the “Interaction” tab in Figure 5.2; it gives Figure 5.5. Select Mud from the left panel (it is a continuous explanatory variable), click the ‘:’ button, and select Transect from the right panel (it is nominal). The interaction term will be of the form:

```
Mud:as.factor(Transect)
```

which is the programming code for interactions in R. To add a second interaction term, click the ‘+’ button, add select more variables. For example:

```
Mud:as.factor(Transect)+Mud:as.factor(Season)
```

To remove all interactions, press ‘Clear’. We did not select any interactions for the Argentinean data set. Appendix A in Zuur et al (2009) contains a series of options how to deal with interactions. If you select interactions, make sure that the main terms are selected as explanatory variables in the “Variables” tab.

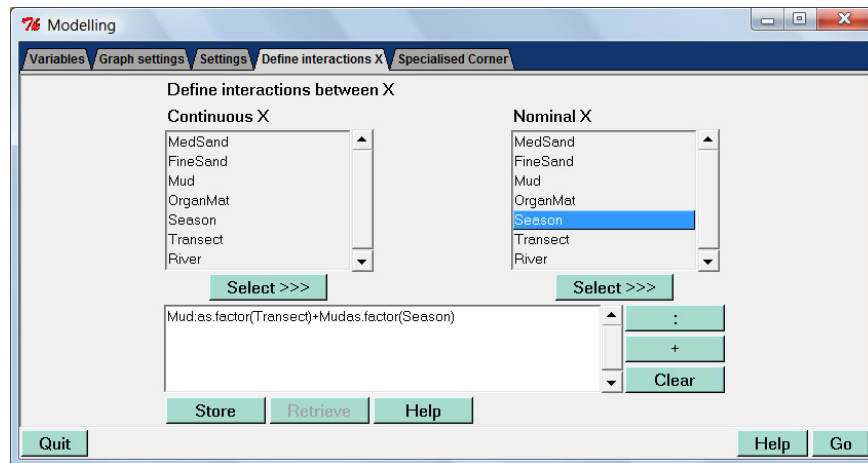


Figure 5.5. Selecting interactions in Brodgar. The text in the panel reads: `Mud:as.factor(Transect)+Mud:as.factor(Season)`. Do not make any mistake with the `:` or the `+`. Note that you must select any terms in the interaction also as a main term in Figure 5.2!

### 5.1.7 Specialised corner

Brodgar versions 2.3.6 and higher have a new tabsheet called the “Specialised corner”. It allows the user to fit a full model, and a nested model. An  $F$ -test is used to compare the models. It is discussed later in this chapter.

### 5.1.8 Visualise the model

Communicating the results of a linear regression model to a non-scientific audience is challenging (it is even challenging if the audience does have a scientific background!). One graph tells more than all the tables coming out of a linear regression analysis, and therefore Brodgar version 2.6.1 and higher have a new tabsheet called “Visualise the model”. This option is only available for linear regression and generalised linear modelling. It allows the user to fit a model with two or more explanatory variables, and create a graph showing what the model is doing. An example is given in Figure A.3 in Zuur et al. (2009). The model that was fitted is of the form:

$$\text{ABUNDANCE} = \alpha + \beta \times \text{LOGAREA} + \text{factor}(\text{GRAZE}) + \varepsilon$$

If you are familiar with this type of models, you will know that this model contains five parallel lines; one for each grazing level. In the next section, we show how to make such a graph.

## 5.2 Regression examples from Zuur et al. (2007; 2009)

In this section, we show how to carry out some of the linear regression models presented in Zuur et al. (2007; 2009).

### 5.2.1 Section 20.4 in Zuur et al. (2007).

In Section 20.4 of Zuur et al. (2007), a linear regression model is applied on the decapod data. These data were also discussed in Chapter 4 of this manual. The response variable is the richness of the decapod families, and the explanatory variables in the linear regression model are T1m, square root transformed chlorophyll-a and location (nominal). Import the data again, apply the square root transformation on chlorophyll-a via the “Import” – “Change data” – “Variables and Transformation” buttons. Do not forget to click the “Save Changes and Finish Data Import Process” button!

Go to the linear regression menu, and ensure that you have the same selection of variables as in Figure 5.6. Clicking the “Go” button gives the same numerical output as on page 379 in Zuur et al. (2007). If you obtain different results, check that in the import process, only the 12 decapod families are selected as response variables, and that chlorophyll is square root transformed.

Besides the numerical output on page 379, you should also get all the graphs in Figure 20.4.

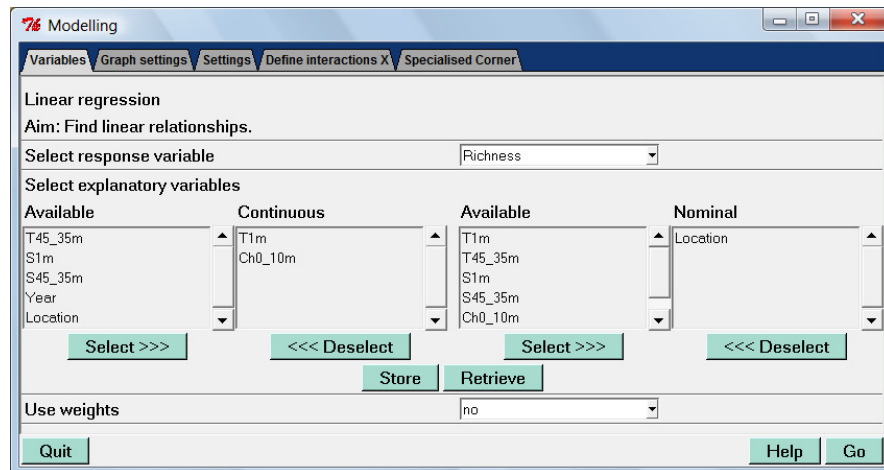


Figure 5.6. Selection of variables to reproduce the linear regression results on page 379 in Zuur et al (2007) for the decapod data. Note that richness is selected as response variable.

### 5.2.2 Section 26.4 in Zuur et al. (2007)

The vegetation data can be found in the file [www.brodgar.com/Vegetation.xls](http://www.brodgar.com/Vegetation.xls). Import the data (note that the first column contains data, not labels!) and select all columns up to ROCK as response variable, and all variables from ROCK onwards are explanatory variables. Go to the linear regression menu and select the variables as in Figure 5.7. The estimated values are given on page 473 in Zuur et al. (2007). The graphical model validation windows are in Figure 26.6 and 26.7. To produce Figure 26.9, select “standardised residuals” for the option “Store residuals in Figure 5.3. Run the linear regression model in Brodgar; it gives Figure 5.8. Note the menu option “Tools”. There are two options under this menu; “View residuals and “Export residuals to Excel”. The first option opens a text file with the (standardised) residuals, the second option copies the residuals to the clipboard, and you can then paste them into, for example, Excel. Provided the selected variables in the linear regression model do not contain any missing values, you can paste the residuals next to the column SAMPLEYR in Excel, and make a scatter-plot.

You can also paste the residuals in Brodgar’s own spreadsheet via “Import” – “Change Data”. If there are missing values in the selected variables in the linear regression model, the residuals will be of shorter length than the original data. Either add NAs to the column with residuals (which is a pain) or remove the missing values from the data (either in Excel or via “Import” – “Change Data” and remove rows) before doing the linear regression analysis in Brodgar.

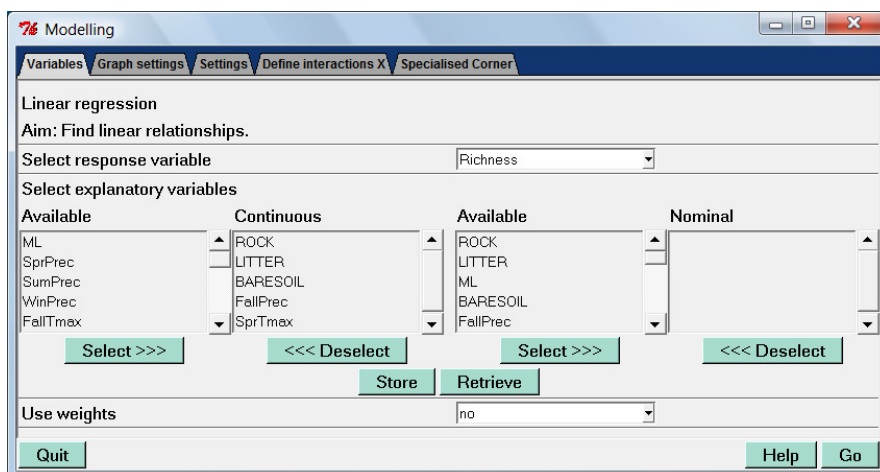


Figure 5.7. Our selection of variables for the vegetation data. Estimated parameters are given on page 473 in Zuur et al. (2007). Note that there are small numerical differences.

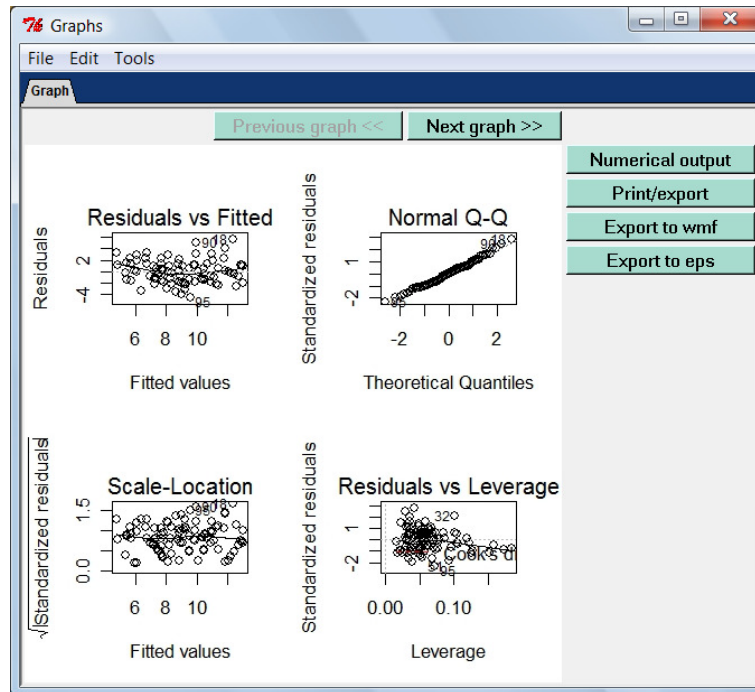


Figure 5.8. Output from the linear regression model applied on the vegetation data. Note the menu option “Stored Information” at the top of the window. You can select “View residuals and “Export residuals to Excel” from this option.

### 5.2.3 Section 7.5 in Zuur et al. (2009)

The data used in Section 7.5 in Zuur et al. (2009) are a subset of the data analysed in Cruikshanks et al. (2007). The original research sampled 257 rivers in Ireland during 2002 and 2003.

One of the aims was to find a different tool for identifying acid-sensitive waters, which currently uses measures of pH. The problem with pH is that it is extremely variable within a catchment, and depends on both flow conditions and underlying geology. As an alternative measure, the Sodium Dominance Index (SDI) is proposed as an indicator of the acid sensitivity of rivers. SDI is defined as the contribution of sodium ( $\text{Na}^+$ ) to the sum of the major cations. Of the 257 sites, 192 were non-forested and 65 were forested.

In section 7.5, pH is modelled as a function of SDI, whether a site is forested or not, and altitude.

The data are in the file [www.brodgar.com/SDI.xls](http://www.brodgar.com/SDI.xls); select pH as response variable, and SDI, forested and altitude as explanatory variables. We will not use the spatial locations. Apply a logarithmic transformation with base 10 on altitude.

In the linear regression menu, ensure that pH is the response variable and select SDI, altitude and forested (nominal) as explanatory variables, see Figure 5.9. Zuur et al. (2009) also included a 3-way interaction, and therefore all 2-way interactions must be included, see Figure 5.10. The text in the window reads as:

```
SDI:as.factor(Forested)+Altitude:as.factor(Forested)+
SDI:Altitude+
SDI:Altitude:as.factor(Forested)
```

The first three components are the 2-way interactions and the last line contains the 3-way interaction term. An alternative would have been:

```
SDI*Altitude*as.factor(Forested)
```

This notation includes all lower interactions automatically. Note that you must still select the main terms in Figure 5.9.

The numerical output (not presented here) shows that the 3-way interaction is not significant. Hence, you can proceed to a backward selection using the AIC, or the drop1 function. It is also interesting to extract the residuals and plot them against their spatial locations (in for example Excel).

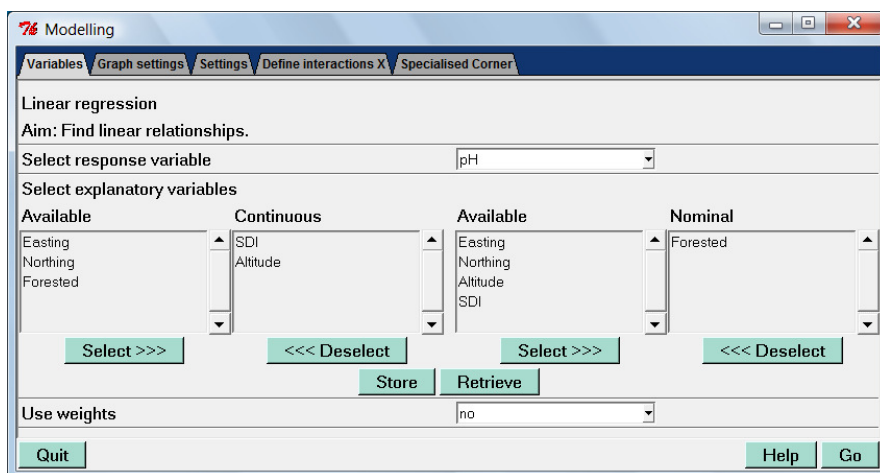


Figure 5.9. Selection of main terms for the Irish pH data.

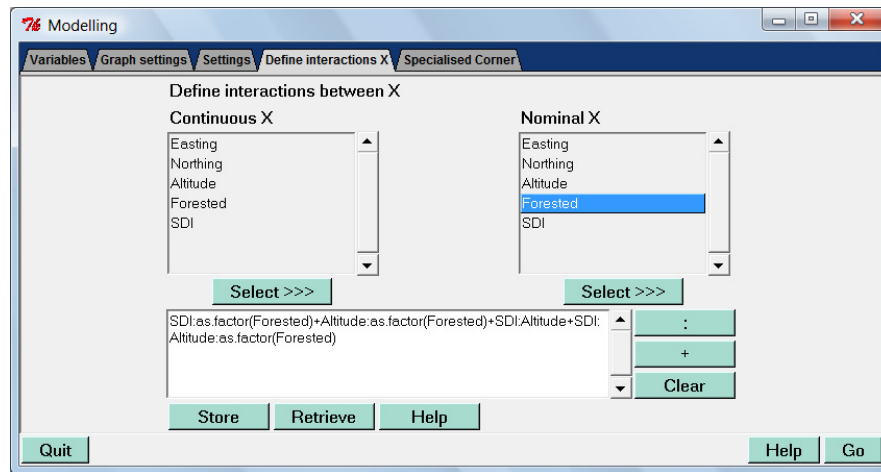


Figure 5.10. Selection of interaction terms for the Irish pH data. The text in the window reads (spaces are optional): SDI : as.factor(Forested) + Altitude : as.factor(Forested) + SDI : Altitude + SDI : Altitude : as.factor(Forested).

#### 5.2.4 Figure A.3 in Zuur et al. (2009)

In this subsection, we illustrate how to visualise the results of a linear regression model. For a GLM, the clicking process is identical.

Open the the Loyn bird data (Loyn, 1987) in from:

<http://www.highstat.com/ZuurDataMixedModelling.zip>

It is the file Loyn.txt. Import the data (e.g. via Excel), set ABUND as response variable and all other variables as explanatory variables, log-10 transform AREA, DIST and LDIST, and apply a linear regression model using ABUND as response variable, and AREA as continuous explanatory variable and GRAZE as a nominal explanatory variable. It is actually a good exercise to follow all the steps in Appendix A.1 – A.3 in Zuur et al. (2009), but we only show how to visualise the final model. Appendix A.3.1 in Zuur et al. (2009) gives the following numerical output for the optimal model:

Coefficients:

	Estimate	Std. Error	t	value	Pr(> t )
(Intercept)	15.7164	2.7674	5.679	6.87e-07	
AREA	7.2472	1.2551	5.774	4.90e-07	
as.factor(GRAZE) 2	0.3826	2.9123	0.131	0.895992	
as.factor(GRAZE) 3	-0.1893	2.5498	-0.074	0.941119	
as.factor(GRAZE) 4	-1.5916	2.9762	-0.535	0.595182	
as.factor(GRAZE) 5	-11.8938	2.9311	-4.058	0.000174	



Residual standard error: 5.883 on 50 degrees of freedom  
 Multiple R-Squared: 0.727, Adjusted R-squared: 0.6997  
 F-statistic: 26.63 on 5 and 50 DF, p-value: 5.148e-13

The question is: What does this all mean in normal words? The answer is given in Appendix A.3.3; a graph is made in which ABUND is plotted versus AREA, and 5 parallel lines are added. Each line represents the ABUND – AREA relationship for a particular grazing level. The lines are parallel because there is no interaction term in the model. If you are not familiar with the underlying theory, then please read Chapter 5 in Zuur et al. (2007) or Appendix A in Zuur et al. (2009).

We now discuss how to make this graph. Note that you need at least two or more explanatory variables in the model, or else Brodgar will give an error message. In the linear regression window, select AREA as continuous explanatory variable and GRAZE as nominal explanatory variable, see Figure 5.11.

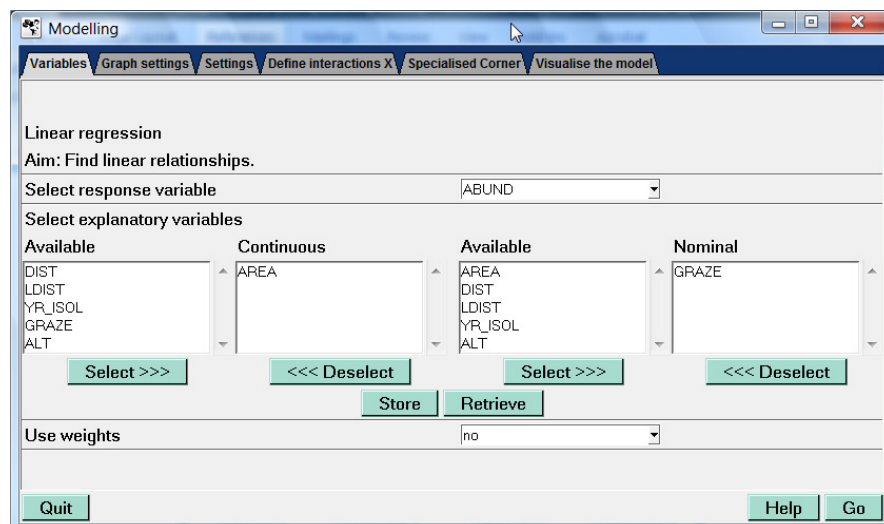


Figure 5.11. Linear regression applied on the bird data from Loyn (1987). AREA (log transformed) and GRAZE are selected as explanatory variables. Note that it is also possible to add interactions.

There is no need to run the model; just select the variables (interactions can be selected as well). Now click on “Visualise the model”, see also Figure 5.12. The clicking process is reasonable simple. First click on “Get current model”. The selected covariates will be shown in the entrybox left of the button. If you want to make changes to the model, then you need to do this from the “Variables” and “Define interactions X” tabsheets; not from this entrybox.

In step 2, you have to specify what you want to plot along the  $x$ -axis, and what you want to do with the other covariates. Click on the button “Specify what to plot”, see Figure 5.13.

By default, the first selected continuous explanatory variable is selected as covariate to be plotted along the  $x$ -axis, which is ok in this case. Brodgar will give an error message if you select a nominal variable for the  $x$ -axis; it has to be one of the continuous explanatory variables.

For GRAZE, we want to have a line for each level, hence stick to the “all levels of factor”. Alternative options are: mean, min, max, first quantile, third quantile, “mean, min and max”, “mean, first quantile and third quantile”, “specific value of factor”, and “all levels of factor”. Note that for a nominal covariate, it does not make sense to use the mean, min, max and quantile options. In this case, we can use all default selections. If you have multiple nominal explanatory variables, then use “all levels of factor” for one of them, and use the “specific value of factor” for all other covariates. Select a specific value that is part of the observed values of the covariate! If male and female are coded as 0 and 1, then you can only use these values as specific values!

If you do not have a nominal covariate in your model, but only continuous explanatory variables, then use the mean, min or max option for each covariate. One of them (but not more than one!) may have the “mean, min and max” option (or the equivalent quantile option).

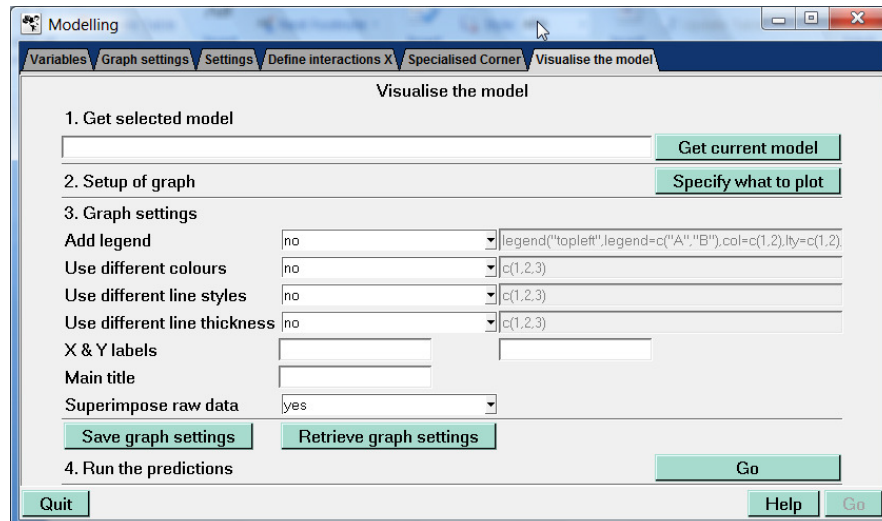


Figure 5.12. Window for visualising the results of the model.

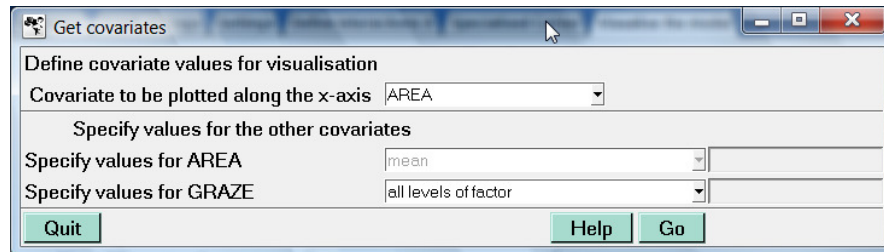


Figure 5.13. Specify what to plot.

In step 3, graph settings can be changed. We advice running the visualisation process first without changing any of these settings. If it works, you can start tidying up the graph. Click on the “Go” button, and Figure 5.14 appears. Each line represents the fitted values for a certain GRAZE level. The problem is that this figure needs axes labels, and a legend, and here is where things get complicated.

Under the “3. Graph Settings” option in Figure 5.12 you can specify a legend, line styles, line thickness, colours of lines and labels, but this has to be done in R language. Before telling what to do, we show the end result, see Figure 5.15. Line type and colours were used, and a legend was added. For sure, this graph looks better than Figure 5.14!

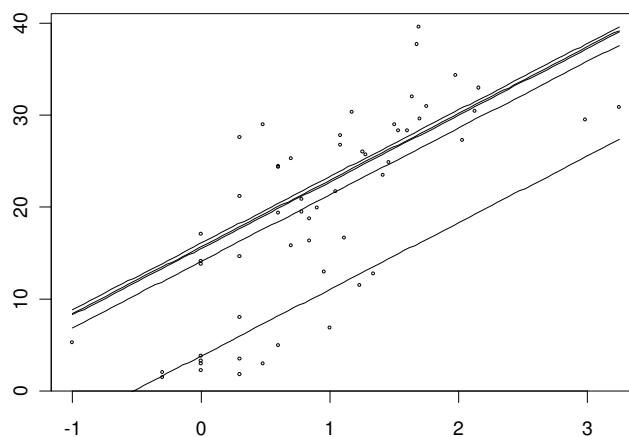


Figure 5.14. Observed and fitted values. The x-axis contains AREA and the y-axis the bird abundance. The points are the observed values and the lines represent the ABUND – AREA relationship for each grazing level.

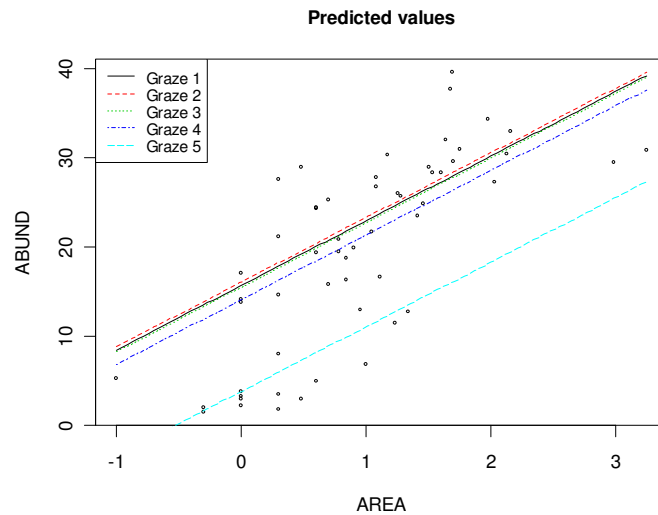


Figure 5.15. As Figure 5.14, but tidied up.

Figure 5.16 shows how we changed the graph settings. The  $x$  and  $y$ -labels, and the main title are trivial. The “Use different colours”, “Use different line types” and “Use different line thickness” options are reasonable simple to adjust. Clicking on “yes” in the corresponding combobox enables the entrybox. The default value for each entrybox is: `c(1,2,3)`. Each number is a different colour, line type or line thickness. Because there are 5 lines (5 grazing levels), you need to change this to: `c(1,2,3,4,5)`. So, one number for each level. The numbers can be the same, or they can differ. For example, you can use `c(1,1,1,2,2)` for colour; this will create three black lines (for levels 1, 2, and 3) and two red lines (for levels 4 and 5). Figuring out which number does what is a bit a matter of trial and error. As quick guide: `c(1,2,3,4,5)` for the colour option gives black, red, green, blue and light blue colours. For line type these numbers give solid, ---, ..., -.- and - - - - lines. For line thickness, 1 is normal and 5 is thick.

The legend bit is the most difficult part to change. It may help to start R, and type `?legend`. See the examples at the end of the help file. The legend in Figure 5.15 was added by changing the default code to:

```
legend("topleft", legend=c("Graze 1", "Graze 2",
  "Graze 3", "Graze 4", "Graze 5"),
  col=c(1,2,3,4,5), lty=c(1,2,3,4,5), cex=0.8)
```

Instead of `topleft`, you can also use `topright`, `bottomright` or `bottomleft`. There are five character strings for the second legend; all strings must be between quotation marks. The `col` argument specifies the colour, and obviously this should match the colours that you choose for the lines. The same holds for the line style

(lty). You can also add `lwd = c(1,2,3,4,5)` for line thickness, but don't forget the comma. The `cex=0.8` specifies the font size of the text in the legend. Whatever you do, one little mistake and R gives an error message. This message will be at the end of the text file that appears.

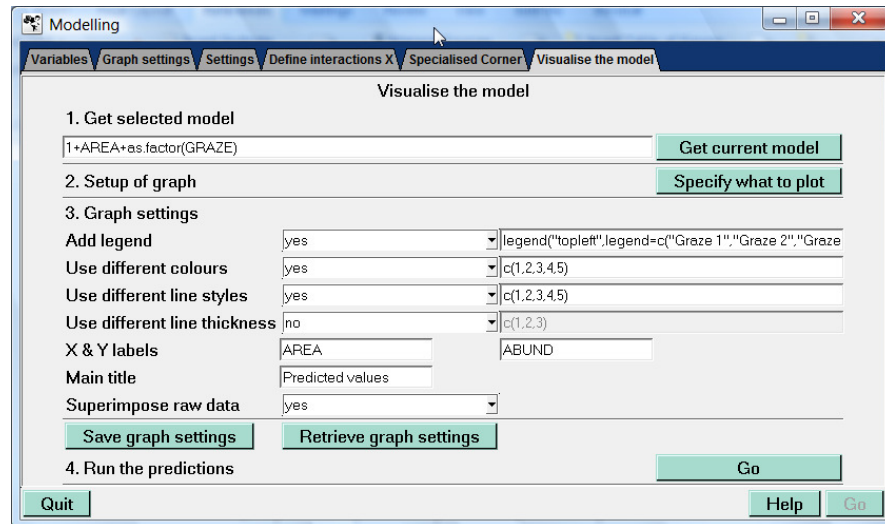


Figure 5.16. Graph settings to produce Figure 5.15.

## 5.3 Generalised linear modelling in Brodgar

A detailed introduction to GLMs for count data (Poisson, quasi-Poisson), and binary and proportional data (binomial) can be found in Chapter 6 in Zuur et al. (2007) and Chapters 8, 9 and 10 in Zuur et al. (2009). The second book also contains information on GLMs with the negative binomial distribution.

The steps to carry out a GLM in Brodgar are nearly identical as in linear regression. To illustrate the clicking process, we use the same data as in Section 9.5 in Zuur et al. (2009). The data set consists of roadkills of amphibian species at 52 sites along a road in Portugal, and were kindly provided by Dr. António Mira (Universidade de Évora, Portugal). The data can be downloaded from: [www.brodgar.com/Roadkills.xls](http://www.brodgar.com/Roadkills.xls). The response variable is TOT\_N (the number of dead animals at a site), and all other variables are explanatory variables.

After importing the data, click on Univariate. Instead of “Linear regression” select “Generalised linear modelling” in Figure 5.1, and click the “Go” button; it gives Figure 5.17. In Section 9.5 in Zuur et al. (2009), only the variable distance to the park, denoted by D\_PARK, is used (the other explanatory variables are used in

later sections). Clicking the “Go” button in Figure 5.17 gives similar graphical and numerical output as in linear regression.

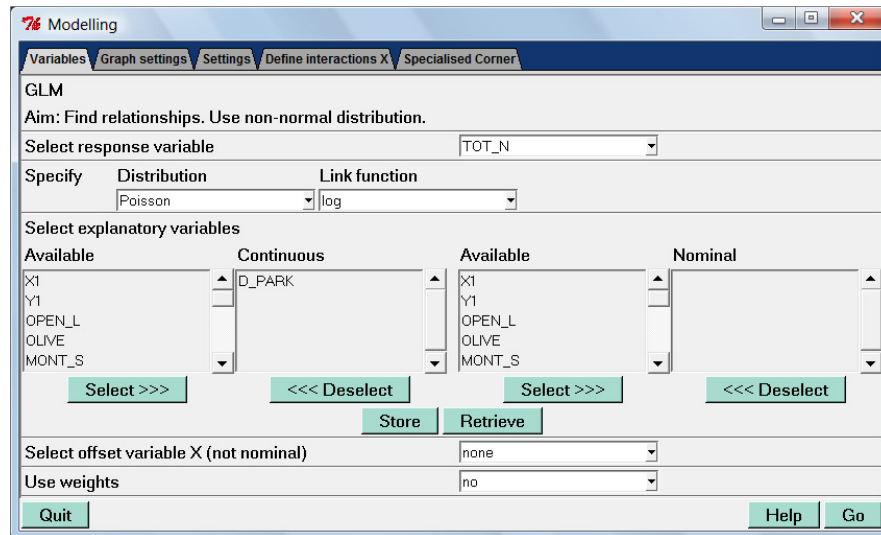


Figure 5.17. Selection of response variable, distribution, link function and explanatory variables. Newer R Brodgar versions also contain a “Visualise the model” tabsheet, see Section 5.2.4.

The numerical output is given below.

```
#####
### NUMERICAL OUTPUT GLM #####
#####
No weights were used

Model is given by f1:
Y1 ~ 1 + D_PARK

Call:
glm(formula = Y1 ~ 1 + D_PARK, family = poisson(link = "log"),
     data = dataz, weights = XW, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.1100 -1.6950 -0.4708  1.4206  7.3337

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.316e+00  4.322e-02  99.87  <2e-16 ***
D_PARK       -1.059e-04  4.387e-06 -24.13  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1071.4 on 51 degrees of freedom
Residual deviance: 390.9 on 50 degrees of freedom
AIC: 634.29

Number of Fisher Scoring iterations: 4

Deviance parameter          = 390.9
n          (null degrees of freedom) = 51
df.residual (residual degrees of freedom) = 50
df          (n-df.residual) = 1

Deviance/df.residual = 7.82
AIC according to formula: -2log(Likelihood) + 2*df = 634.29
```

Because we selected a GLM with Poisson distribution, the overdispersion parameter is set to 1. Hence, the sentence:

```
(Dispersion parameter for poisson family taken to be 1)
```

This does not mean that there is no overdispersion. Dividing the residual deviance by the residual degrees of freedom shows that our assumption is incorrect. This ratio is also presented at the end of the numerical output:

```
Deviance/df.residual = 7.82
```

Because this ratio is larger than 1, you should refit the model with a quasi-Poisson “distribution”. This is done by closing all windows with graphical and numerical output, and change the distribution from Poisson to quasiPoisson in Figure 5.17. The relevant numerical output is now:

```
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.316e+00  1.194e-01  36.156 < 2e-16 ***
D_PARK       -1.058e-04  1.212e-05  -8.735 1.24e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for quasipoisson family taken to be 7.630149)

Null deviance: 1071.4 on 51 degrees of freedom
Residual deviance: 390.9 on 50 degrees of freedom
AIC: 661.56
Number of Fisher Scoring iterations: 4

Deviance parameter          = 390.9
n          (null degrees of freedom) = 51
df.residual (residual degrees of freedom) = 50
df          (n-df.residual) = 1

Deviance/df.residual = 7.82
...
```

The dispersion parameter is 7.63, and all standard errors are automatically adjusted. Note that in quasi-Poisson models, the AIC is not defined. The problem with this model is that it has clear residuals patterns, see Figure 9.6 in Zuur et al. (2009).

If you have multiple explanatory variables, then a stepwise model selection can be applied. For a Poisson (or binomial) GLM it follows the same steps as in linear

regression. Click the “Settings” tab in Figure 5.17. It gives the same window as in the lower panel in Figure 5.3; select backward selection. Alternatively, select the drop1 option, but ensure that the Chi-square test is selected and not the  $F$ -test (this one is needed for quasi-Poisson models). For quasi-Poisson models, model selection should be based on dropping the least significant term. Alternatively, use the negative binomial distribution (see below).

For binary data, quasi-binomial models should not be applied. For proportional data (see below), you can select quasi-binomial models if there is overdispersion.

## 5.4 GLM examples from Zuur et al. (2007; 2009)

### 5.4.1 Section 21.6 in Zuur et al. (2007)

In section 21.6 of Zuur et al. (2007), the presence and absence of the flatfish *Solea solea* was modelled as a function of temperature, salinity, gravel and month (nominal) using a GLM with a binomial distribution. The data were imported and used in Chapter 4 of this manual. To obtain the results presented in Table 21.3, select the variables and distribution as shown in Figure 5.18.

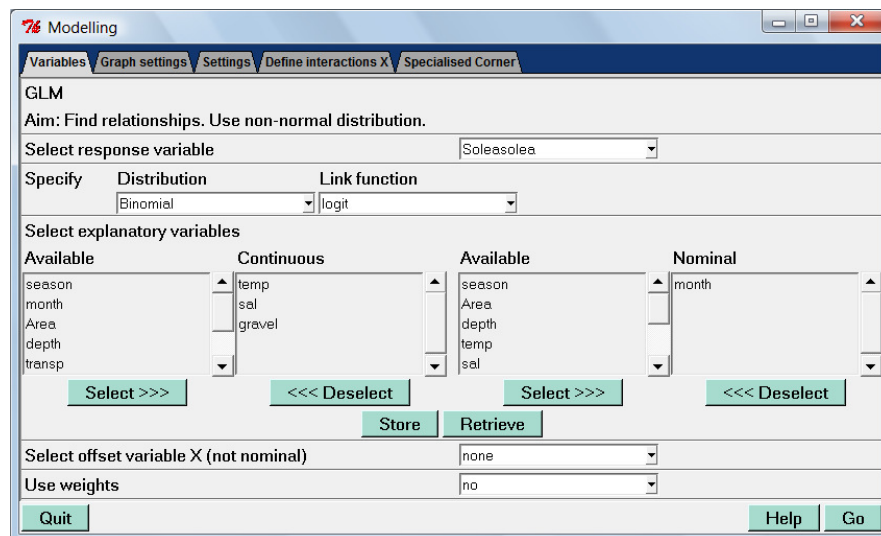


Figure 5.18. Selection of variables, distribution and link function for the *Solea solea* data. Clicking the “Go” button gives the results in Table 21.3 in Zuur et al. (2007). To obtain the results in Table 21.4, click on “Settings”, and apply the drop1 option for “Forward/backward selection”. Make sure that the Chi-square test is selected.



#### 5.4.2 Section 10.3 in Zuur et al. (2009)

In Section 10.3 of Zuur et al. (2009), the results of a binomial GLM applied on proportional data is presented. Vicente et al. (2006) analysed data from a number of estates with wild boar and red deer in Spain. At each estate  $i$ , a group of  $n_i$  animals was sampled. The data set contains information on the tuberculosis (Tb) disease in both species, and on the parasite *Elaphostrongylus cervi*, which only infects red deer. Both variables are recorded as the number of animals that are positive for Tb, or have the parasite *E. cervi*. So, we are modelling the number of animals that test positive, out of  $n_i$  animals. There is also information on the main characteristics of the habitat and management (fencing) at each estate: The percentage of open land, scrubs and pine plantation, number of quercus plants per area, number of quercus trees per area, a wild boar abundance index, a red deer abundance index, estate size (ha) and whether the estate is fenced (1 = yes, 0 = no).

Import the data from the file [www.brodgar.com/TBDeer.xls](http://www.brodgar.com/TBDeer.xls). Select the variable DeerPosProp as response variable; this is the ratio between the number of positive cases and number of observations at a site). Select the following variables as explanatory variables in the data import process: OpenLand, ScrubLand, QuercusPlants, QuercusTrees, RedDeerIndex, EstateSize, fFenced and DeerSampledCervi. The last variable is the number of observations at a site. Go to the GLM menu, and select the variables, distribution, link function and weighting factor (!) as shown in Figure 5.19. Clicking the “Go” button gives the same numerical output as presented in Section 10.3 in Zuur et al. (2009). The numerical output will show that there is overdispersion, and Section 10.3 continuous with a quasi-binomial model. To do this in Brodgar, change the binomial distribution to “quasi-binomial”. The drop1 option can be used to obtain  $p$ -values, but the  $F$ -statistic needs to be selected for quasi-binomial models.

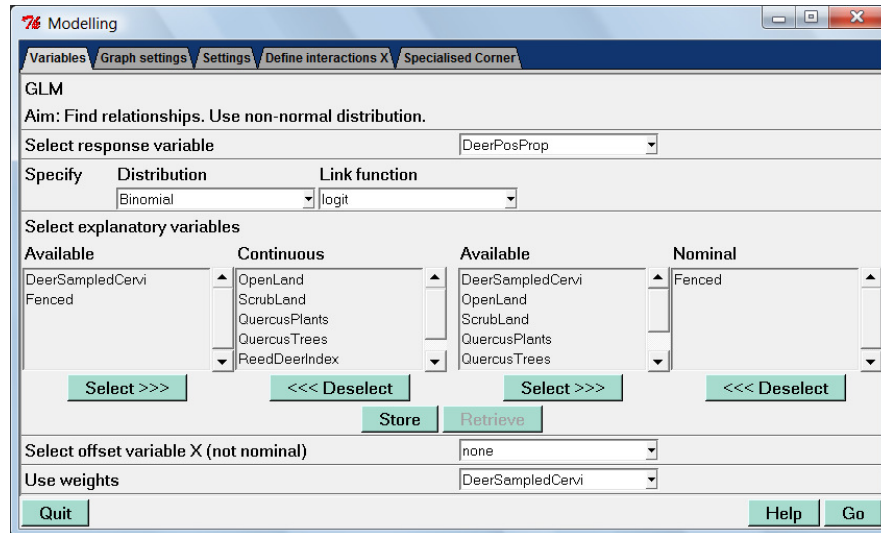


Figure 5.19. Selection of variables, distribution, link function and weights for the deer data. Clicking on the “Go” button produces the output presented in Section 10.3 in Zuur et al. (2009). Note that the number of observations per site is used as a weighting factor, and the proportion of success (positive cases) as the response variable.

#### 5.4.3 Section 9.10 in Zuur et al. (2009)

We return to the roadkill data that was also used in Chapter 4 and in Section 5.3. Zuur et al. (2009) applied a GLM with the negative binomial distribution on these data, and presented results in their Section 9.10.

Import the data again, but apply a square root transformation of the following explanatory variables: POLIC, WAT.RES, URBAN, OLIVE, L.P.ROAD, SHRUB, and D.WAT.COUR, see also Figure 5.20. To apply the GLM with a negative binomial distribution, go to the GLM menu and select the variables, distribution, and link function as in Figure 5.21. If you stick to the default value “Auto” for Theta, then the extra variance term in the negative binomial distribution is automatically estimated (this option was implemented in Brodgar version 2.5.7 and higher). In case of numerical problems, set it to a pre-fixed value.

The numerical output of the negative binomial GLM is fully discussed in Section 9.10 in Zuur et al. (2009).

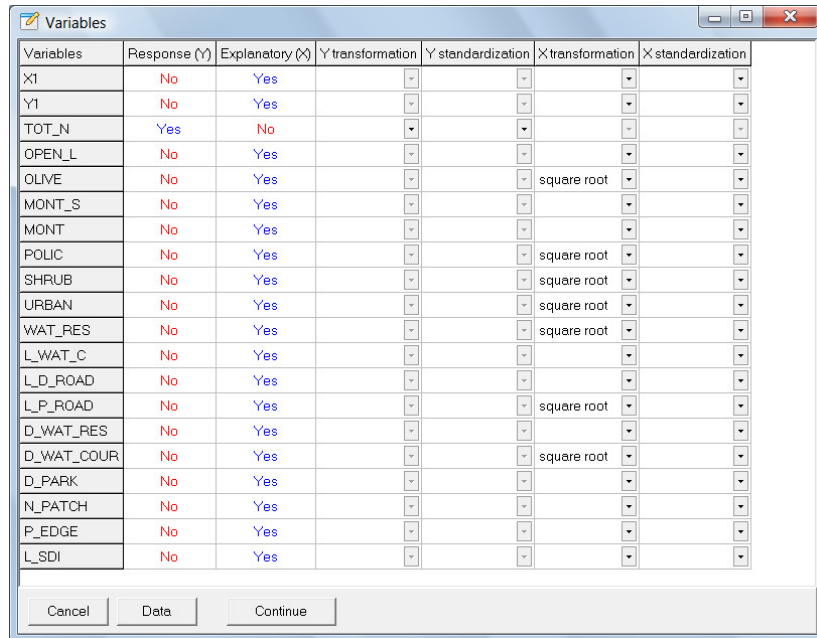


Figure 5.20. The square root transformation is applied on the explanatory variables POLIC, WAT.RES, URBAN, OLIVE, L.P.ROAD, SHRUB, and D.WAT.COUR.

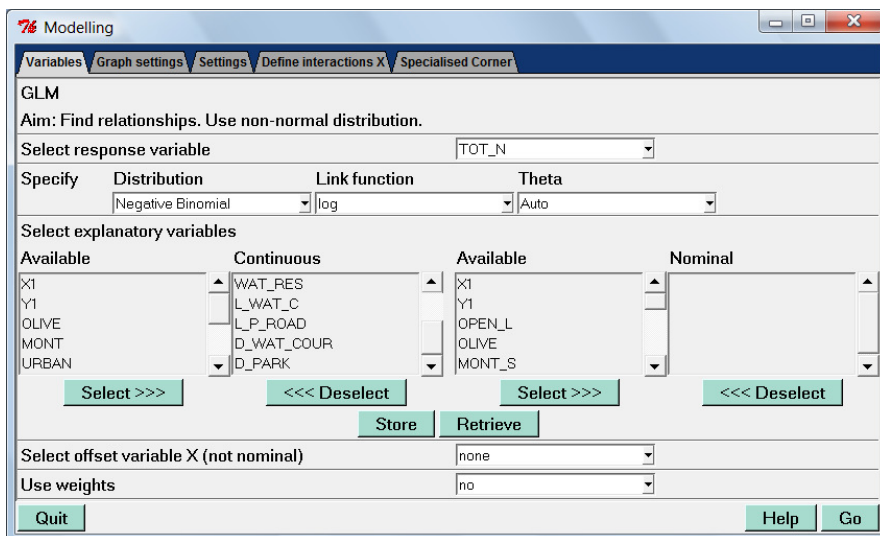


Figure 5.21. Selection of variables, distribution, link function and parameter theta.

## 5.5 Generalised additive modelling in Brodgar

Older versions of Brodgar had two options for GAM, namely:

1. Generalised additive modelling using the gam package based on Hastie and Tibshirani (1990).
2. Generalised additive modelling using the gam function based on the package mgcv from Wood (2006).

Both packages do generalised additive modelling, but the implementation is rather different. The gam package from Hastie and Tibshirani is similar to the SPLUS gam function, and is capable of a backwards selection. However, the mgcv library is useful too as it can estimate the optimal degrees of freedom for a smoother using cross-validation.

In 2007, we decided to use only the gam function from mgcv in Brodgar as it also allows for nested data, auto-correlation and heterogeneity. Note that different versions of the mgcv package in R may give slightly different numerical output (Wood 2006). Additive modelling is generalised additive modelling in which the identity link function and Gaussian distribution are used.

We illustrate GAMs in Brodgar using the squid data set, see also Chapter 4 of this manual. The data are in the file [www.brodgar.com/Squid.xls](http://www.brodgar.com/Squid.xls), GSI is the response variable, and SEX, Location, MONTH and YEAR the explanatory variables. We will fit the following model (see also page 116 in Zuur et al., 2007).

$$\text{GSI}_i = \alpha + f(\text{MONTH}_i) + \text{YEAR}_i + \text{Location}_i + \text{SEX}_i + \varepsilon_i$$

We will also discuss how to add interactions between YEAR, Location and SEX, and between Location and MONTH. A Gaussian distribution is assumed for the error term  $\varepsilon_i$ . To run a GAM (using the mgcv package) in Brodgar, select “Generalised additive modelling (mgcv)” in Figure 5.1 and click on the “Go” button. The window that pops up is in Figure 5.22. The panels are similar to the linear regression menus. Each of the panels and options is discussed next.

### 5.5.1 Selection of variables

You need to select one response variable, at least one smoother, and optionally, parametric terms and nominal variables.

Figure 5.22. Selection of response variable (top row), distribution and link function, smoother (large box), parametric variables and nominal variables. No explanatory variables have been selected yet.

### Response variable

In this case, GSI is the response variable. If you have multiple species, then you can also use a diversity index.

### Explanatory variables – Smoothers, parametric terms and nominals

This is potentially confusing. Brodgar allows you to specify models of the form:

$$Y_i = \alpha + f(X_{1i}) + \beta \times X_{2i} + \text{factor}(X_{3i}) + \varepsilon_i$$

The  $f(X_{1i})$  bit is a smoother,  $X_{2i}$  is modelled as a parametric term (and  $X_{2i}$  needs to be a continuous explanatory variable), and finally,  $X_{3i}$  is modelled as a categorical variable. For the squid data, we want to use a model of the form:

$$\text{GSI}_i = \alpha + f(\text{MONTH}_i) + \text{YEAR}_i + \text{Location}_i + \text{SEX}_i + \varepsilon_i$$

Hence, there is no (continuous) parametric component in this model, only a smoother of month and three nominal explanatory variables. Smoothers should be selected in the big box by clicking the “Add” button. It opens the window in Figure 5.23. We double clicked MONTH in this window in order to select it as a smoother. We will discuss the options later. To close the window, click the “Apply” button.

As to the nominal variables, select them in the lower right box in Figure 5.22.

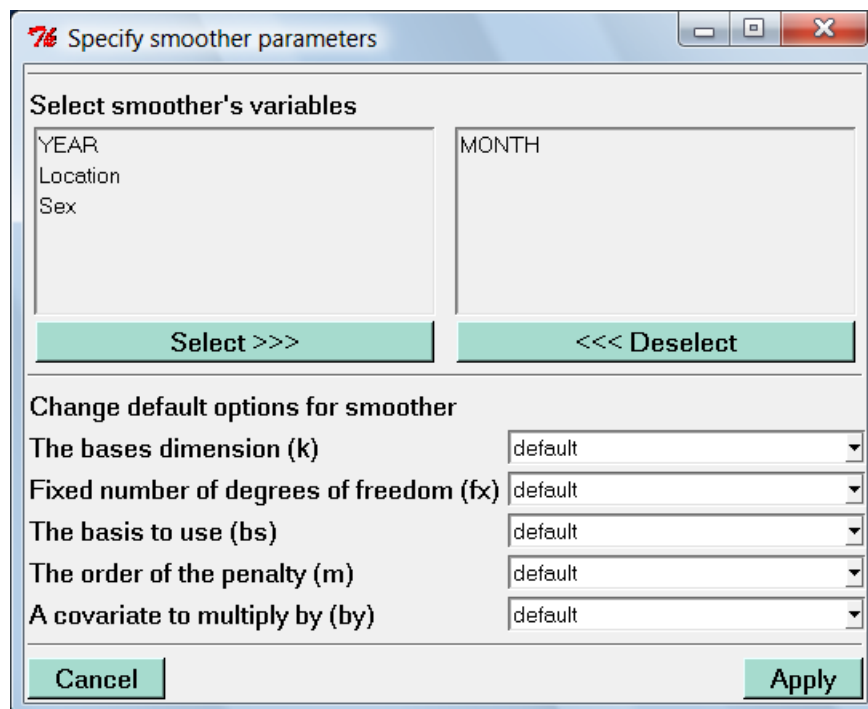


Figure 5.23. Selection of a smoother. We double clicked on MONTH in order to select it. For the moment, leave all options as default. To finalise the selection, click the “Apply” button.

### 5.5.2 Distribution and link function

One has to specify the distribution and link function. See Zuur et al. (2007) for an explanation. We selected the identity link and Gaussian distribution. For count data, it is common to use the Poisson distribution with the log link function. Select ‘QuasiPoisson’ if you want to allow for overdispersion. It is also possible to use the negative binomial distribution with the log-link function. The Binomial distri-

bution is used for 0-1 or proportional data. Only select the 'QuasiBinomial' if you want to allow for overdispersion in the binomial model and have multiple observations per site.

### 5.5.3 Weights

It is possible to use a weighted (generalised) additive model (like in weighted linear regression); create a column in Excel containing the weight for each observation, select it as an explanatory variable in the data import process, and select it as weights in Figure 5.22. Please make sure that the variable defining the weights does not contain missing values, negative values or zeros!

### 5.5.4 Results

Your selection of explanatory variables and other settings should now look like ours in Figure 5.24. Clicking the "Go" button gives the smoother in Figure 7.10B and the numerical output on page 117 in Zuur et al. (2007). The model validation graphs are similar as in linear regression, see Section 5.1 of this manual, and pages 117 – 119 in Zuur et al. (2007).

### 5.5.5 Graphs settings and settings

#### ***Graph settings***

Various trivial graph settings can be changed. The model validation tools for additive modelling contain partial plots showing the partial effects of each smoother, residuals versus fitted values, a QQ-plot of the residuals, hat values, a histogram of the residuals, the response variable (Y) versus fitted value and the response variable versus the linear predictor. For additive modelling, the last two graphs are the same, but they differ for generalised additive models. See the linear regression sections for more details.

#### ***Settings***

Most options under the "Settings" window are graphical settings. The user can add/remove the standard error curves, residuals, rugplots (vertical lines at the  $x$ -axis indicating the values of the observed data) to the partial fits. Residuals and fitted values can be stored or copied to Excel. Values with large hat values can be identified and useful model validation graphs are residuals versus each individual explanatory variable. To guide the eye, smoothers can be added or removed from these graphs.

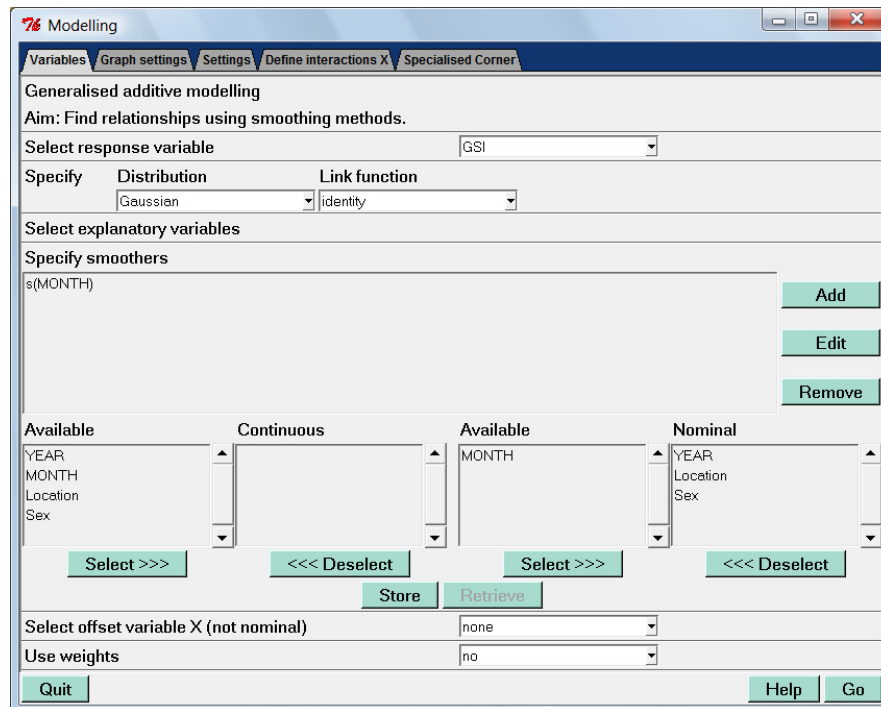


Figure 5.24. Our selection for the smoother, and nominal variables. Clicking the go button gives the smoother in panel 7.10B, and the numerical output on page 117 in Zuur et al. (2007). Note that the gam function from mgcv applies cross-validation to estimate the optimal amount of smoothing for MONTH.

### 5.5.6 Changing options for the smoothers

The bottom part of Figure 5.23 has all kinds of scary looking options. The first two are easy. By default, Brodgar tells the mgcv package to apply cross-validation to estimate the optimal amount of smoothing. Sometimes, it may be better to fix the degrees of freedom to a preset value, e.g. to 4 (default value in SPLUS). To do this, change the second option “fix number of degrees of freedom (fx)” to yes, and set the base dimension to 5 (the degrees of freedom is  $k - 1$ ), see also Figure 5.25.

The last option, “A covariate to multiply by (by)” allows for interaction between a smoother and a nominal variable, and is discussed in the next subsection.

The third option, “Basis to use (bs)” tells the mgcv package what type of smoother to use. Start R, type `library(mgcv)`, and then type: `?s`. This opens a help file with detailed information on the “bs” option:

```
bs: this can be "cr" for a cubic regression spline, "cs" for a cubic regression spline with shrinkage, "cc" for a cyclic (periodic)
```



spline, "tp" for a thin plate regression spline, "ts" for a thin plate regression spline with shrinkage or a user defined character string for other user defined smooth classes. Of the built in alternatives, only thin plate regression splines can be used for multidimensional smooths. Note that the "cr" and "cc" bases are faster to set up than the "tp" basis, particularly on large data sets (although the knots argument to gam can be used to get round this).

The default value in the gam function in mgcv (and therefore Brodgar) is `bs = "tp"`, hence a thin plate regression spline is used. Information on the other smoothers can be found in Wood (2006) and Zuur et al. (2009). The `cc` option is relevant for the MONTH smoother; it ensures that December is connected to January.

The last option we discuss is the "order of the penalty (m)". This is a parameter used in Wood (2006) to make the curves smoother. The help file reads:

m: The order of the penalty for this t.p.r.s. term (e.g. 2 for normal cubic spline penalty with 2nd derivatives). 0 signals autoinitialization, which sets the order to the lowest value satisfying  $2m > d + 1$ , where  $d$  is the number of covariates: this choice ensures visual smoothness. In addition,  $m$  must satisfy the technical restriction  $2m > d$ , otherwise it will be autoinitialized.

To obtain further information on GAM, start R, load the mgcv package from the Packages menu and type: `?gam`.

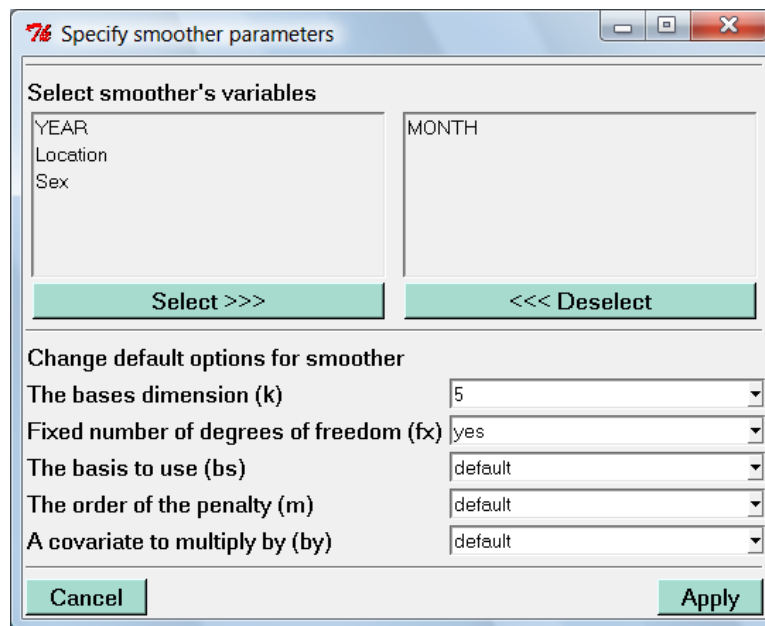


Figure 5.25. Smoother with 4 degrees of freedom. No cross-validation will be applied. The resulting smoother is presented in Figure 7.10A, and the output on page 116 of Zuur et al (2007).

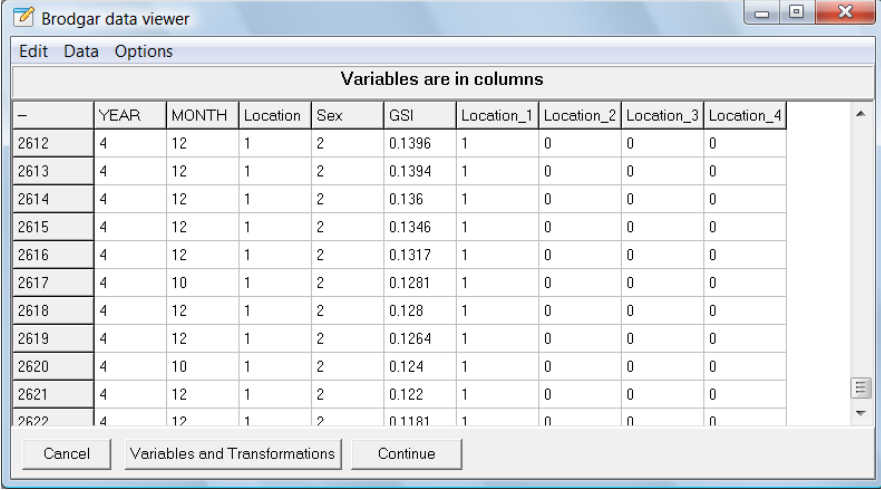
### 5.5.7 Interactions

#### *Interactions between nominal variables*

This is identical as in linear regression. Click on the “Interaction” tab in Figure 5.24. Including interactions between nominal variables, or between a nominal and a parametric term is identical as in linear regression.

#### *Interactions between a smoother and a nominal variable*

Suppose you want to test whether the Month effect differs per location. The by command in Figure 5.23 allows one to select four smoothers for MONTH; one for each location. This process consists of two steps. First, you need to go back to the “Import data” menu, and in “Change Data to be imported” you have to create four new columns with zeros and ones. Let us call these columns Location\_1, Location\_2, Location\_3 and Location\_4. If an observation is from location 2, then Location\_2 should be 1, and Location\_1, Location\_3 and Location\_4 should be zero. Luckily, Brodgar has a build-in function to do this. Go to “Import Data”, and click “Change data to be imported”. Now click on the label Location. It will highlight the entire column Location. Click on “Edit” | “Column” | “Generate dummy variables”, and there they are, see Figure 5.26! Click on “Continue” and “Save Changes and Finish Import Data process”.



	YEAR	MONTH	Location	Sex	GSI	Location_1	Location_2	Location_3	Location_4
2612	4	12	1	2	0.1396	1	0	0	0
2613	4	12	1	2	0.1394	1	0	0	0
2614	4	12	1	2	0.136	1	0	0	0
2615	4	12	1	2	0.1346	1	0	0	0
2616	4	12	1	2	0.1317	1	0	0	0
2617	4	10	1	2	0.1281	1	0	0	0
2618	4	12	1	2	0.128	1	0	0	0
2619	4	12	1	2	0.1264	1	0	0	0
2620	4	10	1	2	0.124	1	0	0	0
2621	4	12	1	2	0.122	1	0	0	0
2622	4	12	1	2	0.1181	1	0	0	0

Figure 5.26. Four extra columns for location. We ended up in this window by clicking “Import Data” | “Change data to be imported”, click on the label Location, and click on “Edit” | “Column” | “Generate dummy variables”. Now click on “Continue” and “Save Changes and Finish Import Data process”.

Return to the GAM menu, and select the nominal variables as before. Carry out the following four steps:

1. Click the “Add” button for adding MONTH as a smoother in Figure 5.22.
2. Double click MONTH so that it is selected as a smoother in Figure 5.23.
3. For the option “A covariate to multiply by (by)“, select Location\_1.
4. Click on Apply in Figure 5.23.

You need to repeat this three more times, and each time select a different dummy for Location. Once you are finished, your selection should match ours in Figure 5.27. To judge whether this model is any better than the model with only one smoother, compare the AICs.

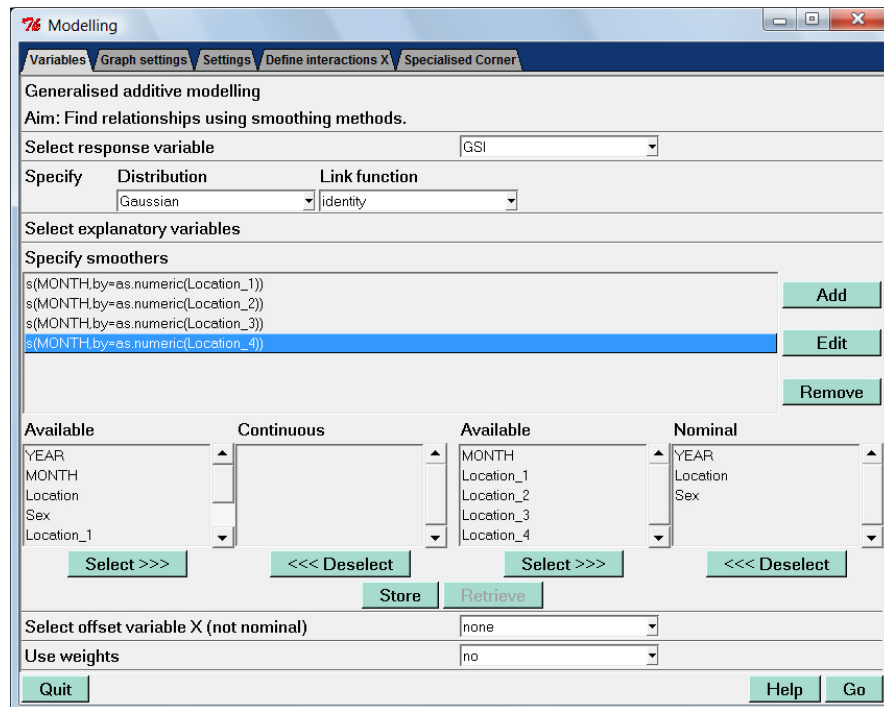


Figure 5.27. Selection of smoother and nominal variables for a GAM that contains an interaction term between the smoother MONTH and the nominal variable Location. Note that there are four smoothers.

### 5.5.8 Specialised corner

Let us return to the situation in which we only have one smoother for MONTH and the three nominal variables YEAR, LOCATION and SEX, see Figure 5.22. And suppose we want to compare the following two models with each other:

$$\text{GSI}_i = \alpha + f(\text{MONTH}_i) + \text{YEAR}_i + \text{Location}_i + \text{SEX}_i + \varepsilon_i$$

$$\text{GSI}_i = \alpha + f(\text{MONTH}_i) + \varepsilon_i$$

Clearly, these models are nested, and we can therefore compare them with an  $F$ -test. Select the smoother and three nominal variables as in Figure 5.22, and click the tab “Specialised corner” in Figure 5.22. The window in Figure 5.28 appears. Click on both “Get current” buttons, and remove the three nominal variables from the nested model (lower box), see Figure 5.28 for our settings. Clicking the second “Go” button applies a GAM on the full, and on the nested model, and at the end of the numerical output, an ANOVA table is presented in which both models are compared.

The screenshot shows the 'Specialised Corner' window with the following settings:

- Define full model:**
  - Distribution: Gaussian
  - Link function: identity
  - Explanatory part: `1+s(MONTH)+as.factor(YEAR)+as.factor(Location)+as.factor(Sex)`
  - Buttons: Add, Get current, Go
- Define nested model for ANOVA comparison:**
  - Distribution: Gaussian
  - Link function: identity
  - Explanatory part: `1+s(MONTH)`
  - Choose the test statistic: F
  - Buttons: Add, Get current, Go

At the bottom of the window are buttons for Quit, Help, and Go.

Figure 5.28. Using the specialised corner to compare two nested models. The text in the first box (full model) reads: `1 + s(MONTH) + as.factor(YEAR) + as.factor(Location) + as.factor(Sex)`, and the text in the second box for the nested model is: `1 + s(MONTH)`.

## 5.6 GAM examples from Zuur et al. (2009)

### 5.6.1 GAM with multiple smoothers

In Section 3.5 in Zuur et al. (2009), the vegetation data are used to illustrate a GAM with multiple smoothers. These data were also used and imported in Chapter 4 of this manual. The GAM we apply is of the form:

$$\text{Richness}_i = \alpha + f(\text{ROCK}_i) + f(\text{LITTER}_i) + f(\text{BARESOIL}_i) + f(\text{FallPrec}_i) + f(\text{SprTmax}_i) + \varepsilon_i$$

To apply this model in Brodgar, import the data, set all species as response variables and select all other variables as explanatory variables in the data import process. In the GAM menu, select richness as the response variable. Just like Section 3.5, we use the Gaussian distribution (though one may argue that a Poisson distribution is more appropriate as richness is a count).

We also allow the cross-validation method to come up with any degrees of freedom for each smoother, including 0 degrees of freedom. All smoothers with 0 degrees of freedom can be dropped simultaneously. This is done by selecting a smoother with shrinkage, see Section 3.5 in Zuur et al (2009).

The potential confusing thing is that you have to repeat the following process for each smoother:

1. Click on the “Add” button in Figure 5.22.
2. Select the cc smoother under the “basis to use (bs)” option in Figure 5.23.
3. Click on the ”Apply“ button in Figure 5.23.

What you should not do is click on “Add” in Figure 5.22, and then select all smoothers in this window. If you do this, a 5-dimensional smoother will be applied. Once you are finished, your selection should match ours in Figure 5.29.

See Section 3.5 in Zuur et al. (2009) for a discussion of the output and problems like collinearity for these data.

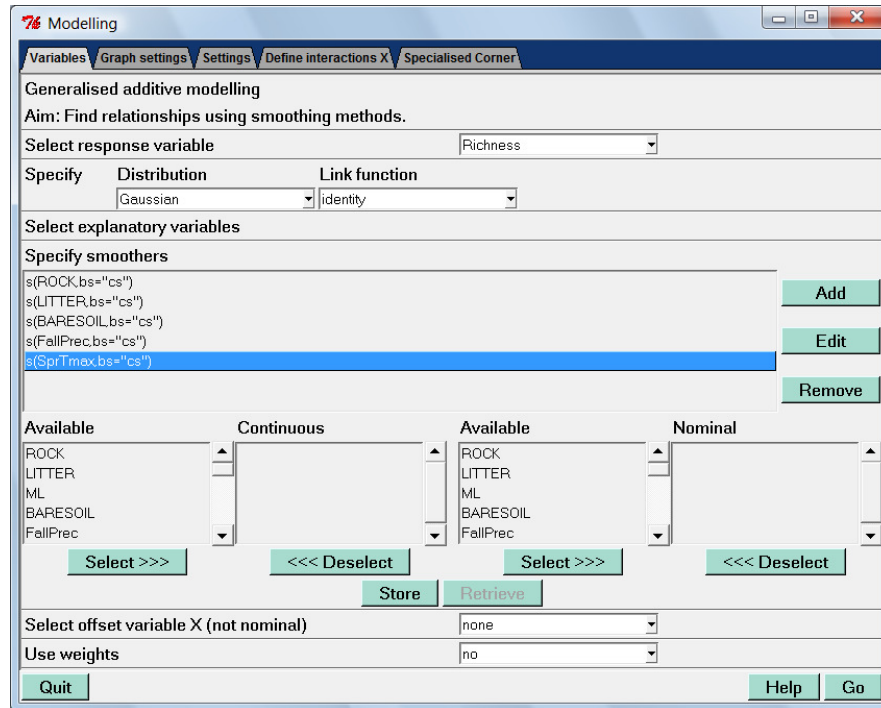


Figure 5.29. Selection of smoothers for the GAM:  $\text{Richness}_i = \alpha + f(\text{ROCK}_i) + f(\text{LITTER}_i) + f(\text{BARESOIL}_i) + f(\text{FallPrec}_i) + f(\text{SprTmax}_i)$ .

### 5.6.2 GAM with an offset

In Section 9.11 in Zuur et al (2009), data from Penston et al. (2008) are used. Plankton tows were taken approximately weekly at two depths (0 meter and 5 meter) at five stations for two years. In the original paper, numbers of nauplii and copepodids were analysed in two separate univariate analysis where production week (time expressed in weeks since March 2002, when the local farms stocked their cages with lice-free, juvenile fish), station and depth were the covariates. There are five stations, labelled as A, C, E, F and G. In Section 9.11, copepodids were used.

One of the problems with these data is that for each sample, a different volume was used. To deal with this, a GLM (or GAM) with an offset variable can be used. See Section 9.11 for the underlying theory. To do this in Brodgar, add an extra column to the original data in Excel, and calculate the (natural) logarithm of volume. Import the new data into Brodgar. In case of any errors in Brodgar, avoid importing data generated by formulae; copy and paste the log-transformed data in Excel as values before importing them into Brodgar (we have seen this problem

on only a very few computers). In Brodgar, select the logarithmic transformed volume as an explanatory variable during the data import process, and in the GAM or GLM with Poisson, quasi-Poisson or negative binomial distribution, select this variable in the offset option in Figure 5.22. Do not select volume, or the log volume as an explanatory variable in the GAM or GLM.

## 5.7 Regression and classification trees in Brodgar

To illustrate how to run tree models in Brodgar, we use the data from Chapter 24 in Zuur et al. (2007). The chapter looks at possible effects of wind farms on birds. The original data consisted of echoes with their associated characteristics as stored by the radar software, plus information from the observer who verified the type of object associated with that of the echo. This information was divided into different categories:

- General information (e.g., date, time).
- Information recorded by the observer at the same moment as the echo was made (e.g., species, flock size, flight altitude).
- Echo appearance information (e.g., echo dimensions, reflectivity).
- Echo position (e.g.,  $x$  and  $y$  coordinates in the radar plane, distance from radar).
- Echo movement (e.g., speed).

Some of these variables represent the same ecological information and had high ( $>0.8$ ) correlations. Using common sense and statistical tools like correlations and principal component analysis, we condensed the number of original variables into a subset of important characteristics.

The verified dataset consists of 659 cases divided over 16 groups. In Chapter 24, 9 different birds groups were classified with at least 10 observations resulting in 629 observations. The nine groups in the analysis were auks, air clutter, water clutter, divers, geese and swans, gulls, sea ducks, ships and terns.

The data are given in the file [www.brodgar.com/radar\\_tree\\_data.xls](http://www.brodgar.com/radar_tree_data.xls). Import the data (note that the first two columns should not be imported as these contain alpha-numerical values). Set the variable g9 as response variable and EPT, TKQ, TKT, AVV, VEL, MXA, AREA, MAXREF, TRKDIS, MAXSEG, ORIENT, ELLRATIO, ELONG, COMPACT, CHY, MAXREF1, MINREF and SDREF all as explanatory variables. Click on “Trees” in Figure 5.1, and the resulting graph is presented in the top panel in Figure 5.30.

Selecting variables is similar as in linear regression. The only extra option is whether you want to do regression or classification trees. In this case, we want the latter one. Clicking the “Go” button gives a classification tree. To obtain the tree in Figure 24.3, set the cp value in the lower panel in Figure 5.30 to 0.0323.

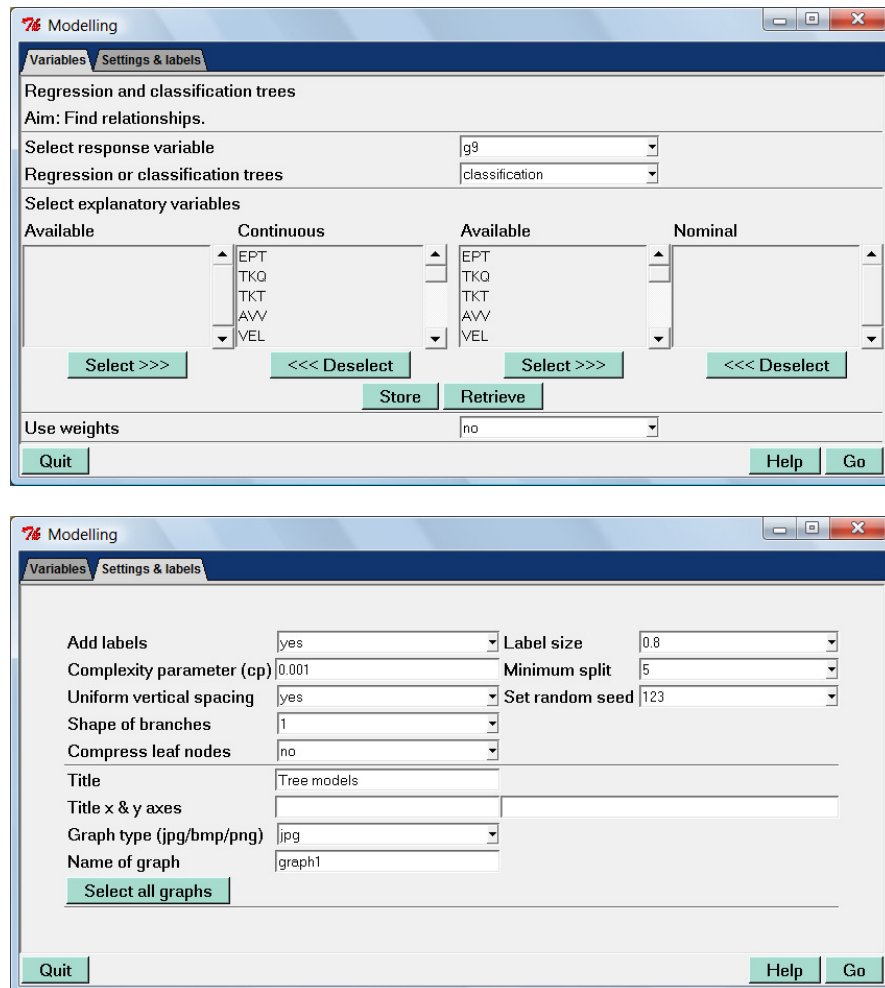


Figure 5.30. Univariate tree models in Brodgar.

### Settings and labels

- **Add labels.** Labels are the classification rules. Try to use short variable names. It might be an option to omit them.
- **Label size.** Increase or decrease the size of the labels.
- **Complexity parameter (cp).** Choose the complexity parameter cp. A good choice of cp is the leftmost value for which the mean (dot) lies below the horizontal line, in the cp plot.



- Minimum split. If there are lots of branches, it might be an option to make the tree smaller by only allowing new branches if there are at least  $x$  samples in a group. This is a sort of pre-pruning.
- Add info N. This provides information on the group sizes.
- Uniform vertical spaces (source: R help files). If 'yes', uniform vertical spacing of the nodes is used. This may be less cluttered when fitting a large plot onto a page. The default is to use a non-uniform spacing proportional to the error in the fit.
- Shape of branches (source: R help files). This controls the shape of the branches from parent to child node. Any number from 0 to 1 is allowed. A value of 1 gives square shouldered branches, a value of 0 give V shaped branches, with other values being intermediate.
- Compress leaf nodes (source: R help files). If 'no', the leaf nodes will be at the horizontal plot co-ordinates of '1:nleaves'. If 'yes', the R routine attempts a more compact arrangement of the tree. The algorithm assumes 'uniform vertical spaces=yes'. The result is usually an improvement even when 'uniform vertical spaces=no'.
- Set random seed. By setting the random seed, the cross-validation will produce the same results each time it is applied.
- Other points are the titles, graph name and graph type; see Sections 5.1 and 5.2 for details.

## 5.8 Mixed effects modelling in Brodgar

A non-technical introduction to mixed effects modelling is given in Chapter 8 of Zuur et al. (2007), and in Chapter 5 of Zuur et al. (2009). In Chapter 8 in Zuur et al. (2007), a series of linear regression models were applied on the RIKZ data (marine benthic data set). The response variable was species richness and the explanatory variable was NAP (height of a site compared to average sea level). The data were sampled at 9 beaches (5 samples per beach) and the question is whether there is a relationship between richness and NAP.

### 5.8.1 The random intercept model

We start with the random intercept model, see model 5 on page 128 in Zuur et al. (2007). The equation for a random intercept model is reproduced below.

$$\text{Model 5} \quad Y_{ij} = \alpha + \beta \times \text{NAP}_{ij} + a_j + \varepsilon_{ij}$$

where  $a_j \sim N(0, \sigma_a^2)$  and  $\varepsilon_{ij} \sim N(0, \sigma^2)$

$Y_{ij}$  is the richness of observation  $i$  on beach  $j$ , where the index  $j$  takes values from 1 to 9, and  $i$  from 1 to 5 (there are 5 observations per beach). The model

states that there is one intercept  $\alpha$  and one slope  $\beta$ . This part of the model is called the *fixed part*. On top of this, there is a random intercept  $a_j$ , which adds a certain amount of random variation to the intercept at each beach. The random intercept is assumed to follow a normal distribution with expectation 0 and variance  $\sigma_a^2$ .

We show how to apply this model in Brodgar for the RIKZ data. These data were used in earlier chapters of this manual, and can be downloaded from the URL: [www.highstat.com/RIKZ.xls](http://www.highstat.com/RIKZ.xls). Open the data in Excel, import them to Brodgar, select the columns labelled as C1 to I5 as response variables (these are all the zoobenthic species), and all other columns as explanatory variables. Once the data are imported, click on the “Univariate” main menu button. Select the tab “GLS, mixed modelling & GAMM”. The window in Figure 5.31 appears. As the name of the tab suggest, you can select from generalised least squares (GLS), linear mixed effects modelling and generalised additive mixed modelling (GAMM). We discuss GLS and GAMM later. Select “Linear Mixed Effects Models” and click on the “Go” button. The resulting graph is presented in Figure 5.32.

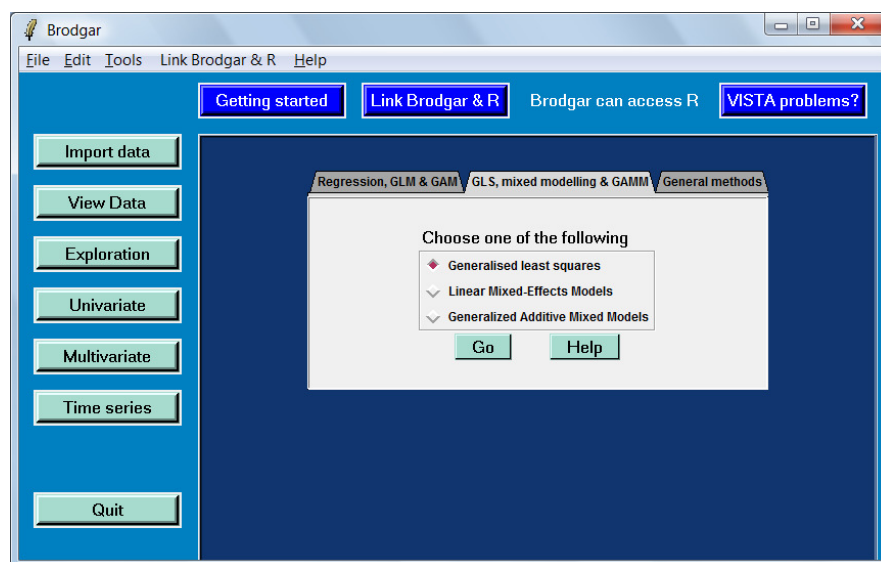


Figure 5.31. Options under the “GLS, mixed modelling & GAMM” tab.

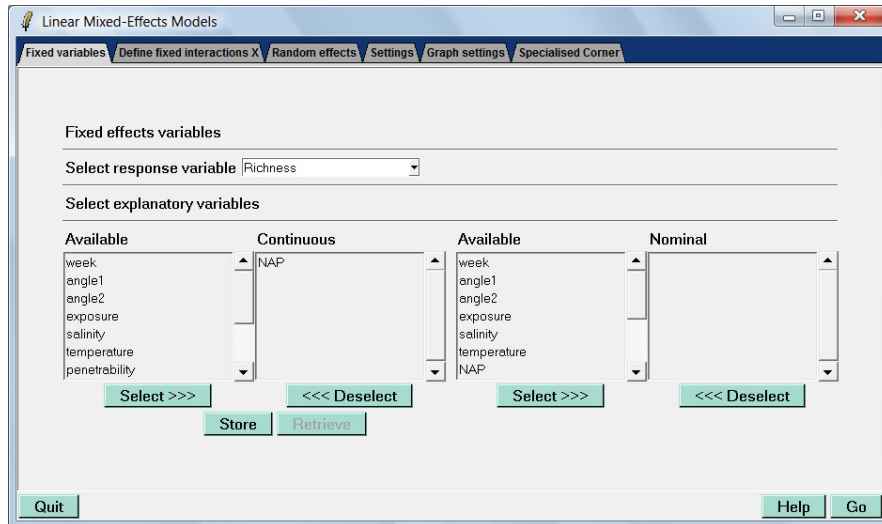


Figure 5.32. Selection of the response variable and the fixed effects part of the model. We selected species richness as the response variable (ensure that only the 75 species are selected as response variables in the data import process!) and NAP as continuous explanatory variable.

In the panel labelled “Fixed variables” in Figure 5.32, you need to select a response variable (we selected species richness) and the fixed part of the model (we selected NAP). The clicking process is identical as in linear regression, see Section 5.1. If multiple explanatory variables are selected, you can also add interactions and this can be done in the panel labelled “Define fixed interactions”. The clicking process is identical as in linear regression, see Section 5.1.7. However, in this particular example, there is no scope for interactions as we only use one explanatory variable.

Click on the tab labelled “Random effects” in Figure 5.32; the resulting window is presented in Figure 5.33. From the box “Select grouping factors”, select the variable Beach, and click on the “Add” button. Note that the white box on the lower left side contains (by default) a “1”. Keep it in. Clicking on the “Go” button will execute the random intercept model. Internally, Brodgar sets up ascii files with R commands, runs R in BATCH mode, and applies the mixed effects model using the lme function from the nlme package. When R is finished, all results are saved in text files and Brodgar shows the results, see Figure 5.34. The graph shows the (normalised) residuals versus fitted values, and this graph can be used to assess the homogeneity assumption. The button “Next graph” leads to a histogram of the normalised residuals, and can be used to assess the normality assumption. The numerical output can be obtained by clicking the “Numerical output” button. It is discussed in detail on pages 128 – 129 in Zuur et al. (2007). Note that the anova table at the end of the numerical output applies sequential testing.

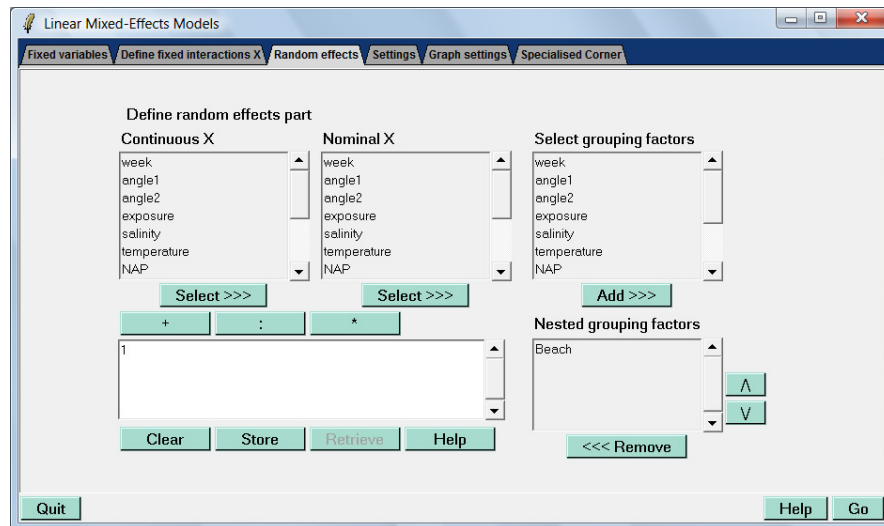


Figure 5.33. Application of a random intercept model. Beach is selected as random intercept.

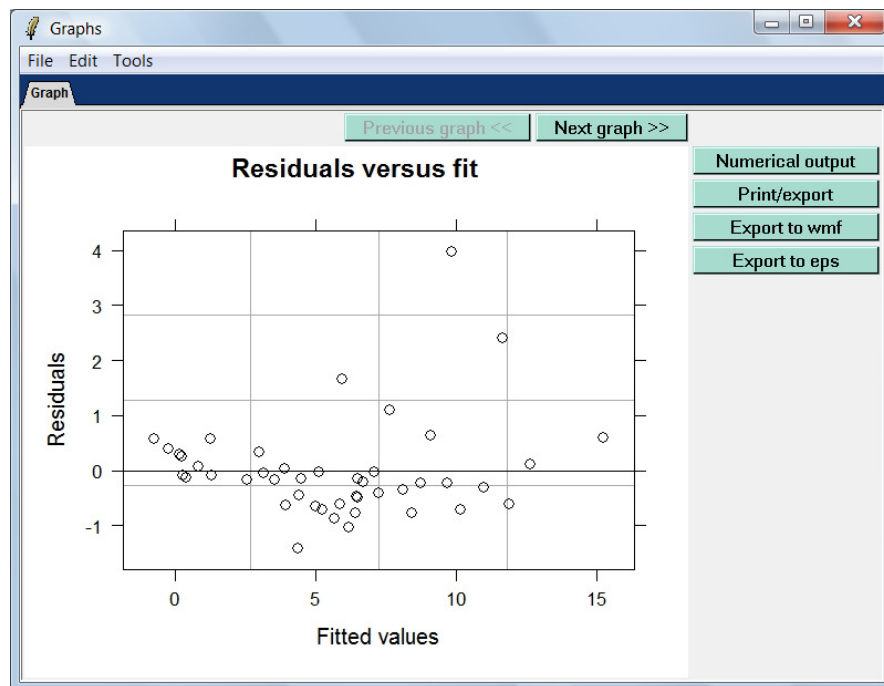


Figure 5.34. Output from the lme function in R. The first graph shows residuals versus fitted values, and can be used to assess homogeneity.

### 5.8.2 Settings

Various settings can be changed; click on the “Settings” tab in Figure 5.35. Some settings are trivial (e.g. the number of digits), others are important. An important option is the “Model fitting method”; by default this is restricted maximum likelihood (REML). A detailed explanation when to use REML and when ML (maximum likelihood) can be found in Chapter 5 of Zuur et al. (2009). See their 10-step protocol in Chapter 4.

The options in Figure 5.35 also allow one model heterogeneity, or temporal and spatial correlation structures, and these are discussed later in this chapter. For the moment, keep all settings as they are.

Linear Mixed-Effects Models	
Fixed variables Define fixed interactions X Random effects <b>Settings</b> Graph settings Specialised Corner	
Model fitting method by REML or ML	Restricted log-likelihood
Allowing for heterogeneity	no
Variance structure for heterogeneity	<input type="text"/> Add
Allowing for correlation	no Spatial
Correlation structure	<input type="text"/> Add
Number of iterations for the EM algorithm	default
Maximum number of evaluations of the objective function	default
Nominal variables contrasts	Treatment
Number of digits in the numerical output	7
Confidence intervals in numerical output	No
Type of residuals	normalized
Spatial variables	no
Spatial distance metric	euclidean
Calculate Moran's I statistics	no
Store Retrieve Clear	
Quit Help Go	

Figure 5.35. Options for mixed effects models. By default, REML estimation is used. We assume that residuals are homogenous, and independent; hence, the “no” for “Allowing for heterogeneity” and “Allowing for correlation”. The contrast for categorical variables is the so-called “Treatment” option. This means that the first level is used as baseline.

### 5.8.3 The random intercept and slope model

Model 5 is a mixed effect model with a random intercept. Hence, the regression line is allowed to randomly shift up or down. Extending this process to a model in which not only the intercept is allowed to vary randomly but also the slope follows

the same principle. This is called a random intercept and slope model. The mathematical formulation of the model is given by:

$$\text{Model 6} \quad Y_{ij} = \alpha + a_j + \beta \times \text{NAP}_{ij} + b_j \times \text{NAP}_{ij} + \varepsilon_{ij}$$

where  $\varepsilon_{ij} \sim N(0, \sigma^2)$  and  $a_j \sim N(0, \sigma_a^2)$  and  $b_j \sim N(0, \sigma_b^2)$

The equation was taken from page 129 in Zuur et al. (2007). This is the same formulation as model 5, except for the term  $b_j \times \text{NAP}_{ij}$ . It allows for random variation of the slope at each beach.

To fit this model in Brodgar, go to the “Random effects” window in Figure 5.33, click on the “+” button and double click on NAP. The window should now look like the one in Figure 5.36. Clicking the “Go” button applies the random intercept and slope model. The numerical results are given on page 129 – 130 in Zuur et al. (2007); note the small differences in estimated values (this is probably a typing mistake in the book).

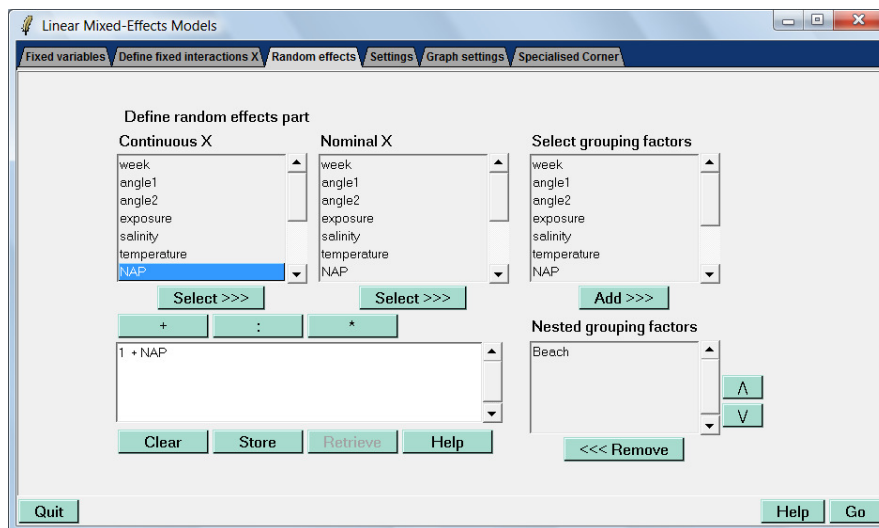


Figure 5.36. Application of a random intercept and slope model. The variable Beach is selected as “Nested grouping factor” and the lower left box reads: 1 + NAP.

#### 5.8.4 Specialised corner for mixed effects models

It is interesting to compare a mixed effects model that contains the random intercept Beach, with a model that does not contain the random intercept Beach.

These models are nested, and therefore the likelihood ratio test can be applied. The null hypothesis is whether the variance for the random intercept Beach is 0.

To do this test in Brodgar, select Richness and NAP as in Figure 5.32. Also select the random intercept Beach as in Figure 5.33. In the “Settings” menu, ensure that REML estimation is selected. Now click on the “Specialised corner” tab in Figure 5.35. We get the window in Figure 5.37. Click on both “Get current” buttons, and then remove the code for the “Random effects part” for the nested model. Your settings should look like ours. Note that both models have the same fixed structure, and REML is selected. The sub-model is actually a GLS (see below) because no random intercept is specified.

Click the second “Go” button, and after Brodgar is finished, the end of the file with numerical output shows the following output:

```
===== ANOVA table for the submodels =====
      Model df      AIC      BIC    logLik    Test  L.Ratio p-value
Submodel1   1  4 247.4802 254.5250 -119.7401
Submodel2   2  3 258.2010 263.4846 -126.1005 1 vs 2 12.72075 4e-04
```

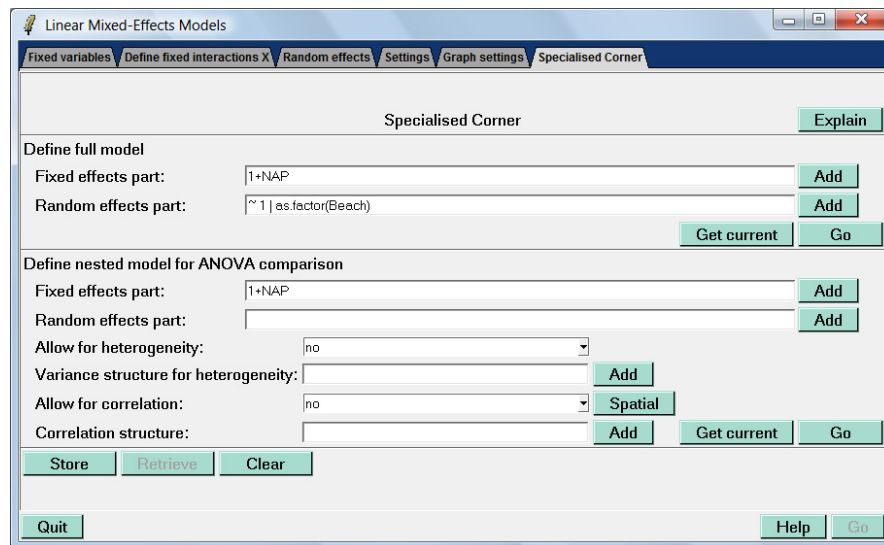


Figure 5.37. Specialised corner menu to compare two models with the same fixed effects, but with different random effects (or vice versa). In this case, we compare a model with a random intercept Beach, with a model that does not contain a random intercept (a GLS model).

The likelihood ratio statistic is  $L = 12.72$ . The R code that generated the  $p$ -value assumes that  $L$  is Chi-square distributed with 1 degree of freedom. The problem is that we are testing on the boundary (the variance of the random intercept is 0, with the alternative that the variance is larger), and therefore the  $p$ -value needs to be adjusted, see Chapter 5 in Zuur et al. (2009). To obtain the correct  $p$ -value, start R, and type:

```
> 0.5 * (1 - pchisq(12.72, 1))
```

followed by an enter.

To compare a model with a random intercept, and a model with a random intercept and slope, use a similar structure:

1. Specify the random intercept and slope model as explained in Figure 5.36.
2. Go to the „Specialised corner“ and click both “Get current“ buttons.
3. Change the random structure of the sub-model to:
 

```
~ 1 | as.factor(Beach).
```
4. Click the second “Go“ button.

An example is given in the next subsection.

### 5.8.5 Mixed effects modelling example from Zuur et al. (2009)

In Chapter 19 of Zuur et al. (2009), mixed effects modelling is applied to model the density of *Paenibacillus larvae* in honey bees as a function of the number of bees in the hive, presence or absence of American Foulbrood (AFB, an infectious disease) and hive identity. The following model was applied:

$$LSpobe_{ij} = \alpha + \beta_1 \times BeesN_{ij} + \beta_2 \times Infection_{ij} + \beta_3 \times BeesN_{ij} \times Infection_{ij} + a_i + \varepsilon_{ij}$$

$LSpobe_{ij}$  is the logarithmic transformed density of spores,  $BeesN_{ij}$  the number of bees in hive  $i$ , and  $Infection_{ij}$  is a nominal variable indicating whether a hive is infected (1) or not (0). The term  $a_i$  is the random intercept for hive. Full details can be found in Sections 19.1 and 19.2 in Zuur et al. (2009).

Import the data from [www.highstat.com/BeesHives.xls](http://www.highstat.com/BeesHives.xls). The first column contains data. Set SpoBee as response variable and apply a  $\log(Y+1)$  transformation on it. The variables Hives, BeesN and Infection are explanatory variables. It is useful to apply a data exploration to reproduce Figures 19.2 and 19.3. Also apply a linear regression to model SpoBee (log transformed) as a function of BeesN, Infection, and their interaction. You should also be able to produce Figure 19.4 (store the residuals, copy and paste them in the spreadsheet via “Change Data” under the main menu “Import”, and make a boxplot under data exploration). We now apply mixed effects modelling and follow the 10-steps protocol as outlined in Section 19.3 in Zuur et al. (2009).



**Step 1 of the protocol**

In this step, linear regression is applied. See the previous paragraph.

**Steps 2 – 6 of the protocol**

First, we need to refit the linear regression model without any special variance structure using the GLS function. This gives us a reference model. We cannot use the ordinary linear regression function from Section 5.1, as it uses ordinary least squares. Instead, use the GLS option in Figure 5.31, select the response and the two explanatory variables (BeesN is continuous, Infection is nominal), and their interaction. Ensure that the default estimation method, “Restricted log-likelihood”, alias REML, is selected (Settings – “Model fitting method by REML or ML”). The relevant numerical output is:

```
Generalized least squares fit by REML
Model: Y1 ~ 1 + BeesN + Infection + BeesN:Infection
Data: dataz2
AIC BIC logLik
252 263 -121
```

Coefficients:

	Value	Std.Error	t-value	p-value
(Intercept)	2.64	0.526	5.02	0.0000
BeesN	0.00	0.000	-1.73	0.0876
Infection1	3.65	0.875	4.17	0.0001
BeesN:Infection1	0.00	0.000	-1.20	0.2353

This model gives us an AIC of 252, but does not take into account the explanatory variable Hive. To use Hive as a random intercept, close the numerical output, and GLS menu, and click on “Linear Mixed Effects Models” in Figure 5.31. Select exactly the same fixed structure, namely BeesN as a continuous explanatory variable, Infection as a nominal explanatory variable, and their interaction. See the top and middle panels in Figure 5.38. To select Hive as a random intercept, click on “Random effects” and select Hive for “Nested grouping factors”, see the lower panel in Figure 5.38. Also ensure that the default estimation method, REML, is used. If you click on any of the three “Go” buttons in Figure 5.38, the mixed effects model in Equation (19.1) in Zuur et al. (2009) is applied. The relevant part of the numerical output is given below.

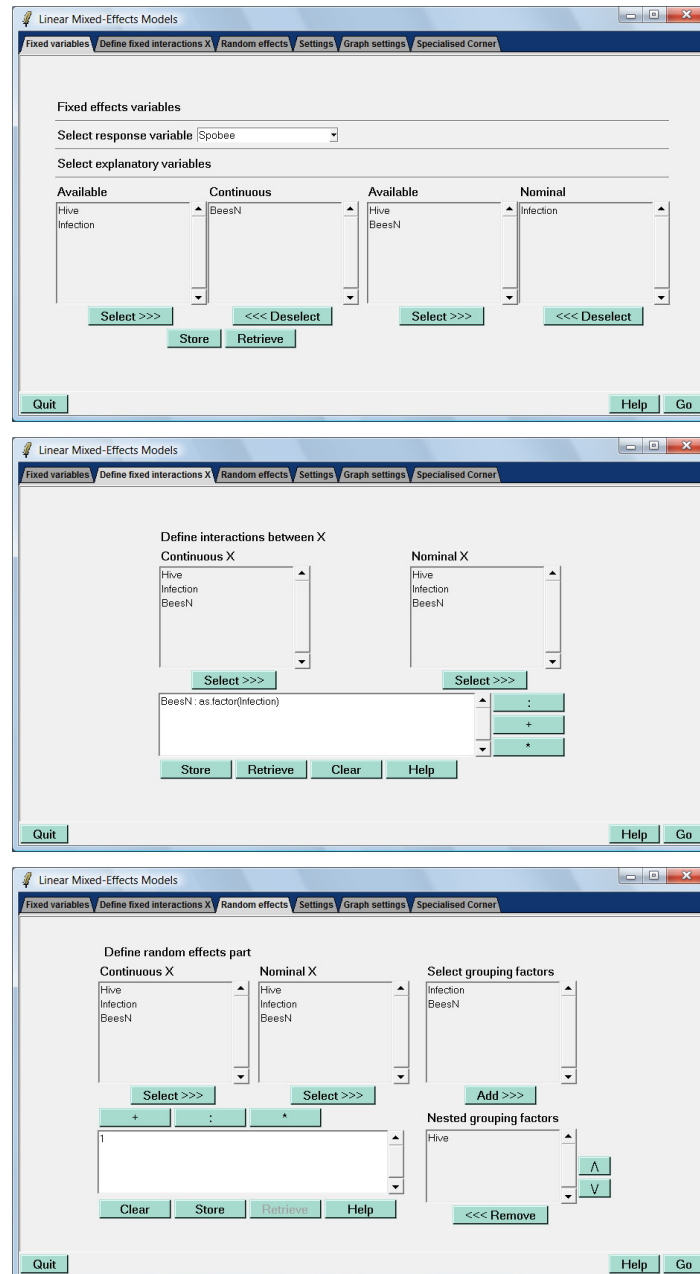


Figure 5.38. Top panel: Selection of response and explanatory variables for the fixed part of the mixed effect model. Middle panel: Selection of interaction for the fixed part of the model. Lower panel: Selection of the random intercept.

```

Linear mixed-effects model fit by REML
Data: dataz2
    AIC BIC logLik
    175 188  -81.5

Random effects:
Formula: ~1 | Hive
      (Intercept) Residual
StdDev:      0.967      0.337

Fixed effects: Yl ~ 1 + BeesN + Infection + BeesN:Infection
              Value Std.Error DF t-value p-value
(Intercept)   2.64    0.928  48   2.848  0.0065
BeesN          0.00    0.000  20  -0.983  0.3374
Infection1     3.65    1.542  20   2.365  0.0283
BeesN:Infection1 0.00    0.000  20  -0.679  0.5049

```

These two models (with and without the random intercept) are nested, and Zuur et al. (2009) compared them using a likelihood ratio test (though the  $p$ -value has to be corrected because we are testing on the boundary). This test can be carried out in the specialised corner. To apply this test in Brodgar, close the graphical and numerical output, and click on the “Specialised corner” tab in Figure 5.38. We obtain Figure 5.39. Remove the code from the “Random effects part” of the nested model. Your selection should be identical as ours in Figure 5.39.

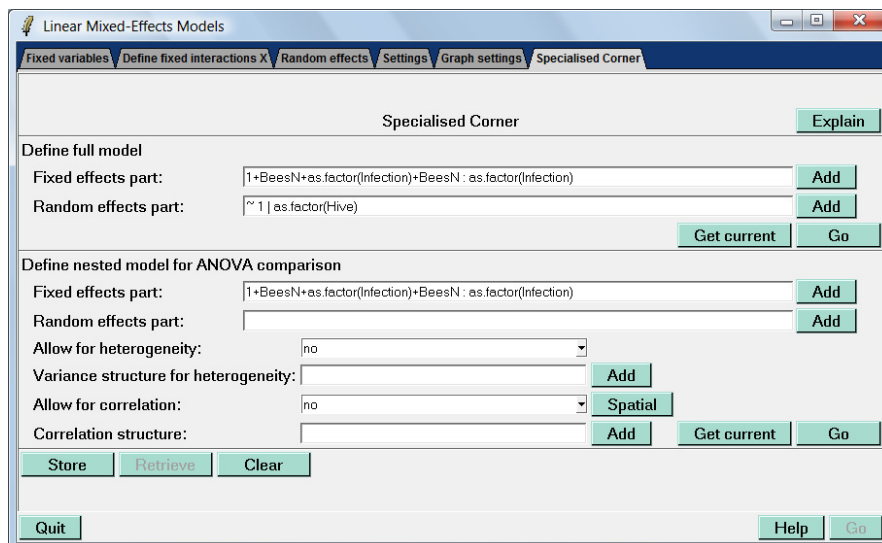


Figure 5.39. Specialised corner for the honey bee example. We removed the code from the “Random effects part” of the nested model. Ensure that REML estimation is used.

The output reads as follows.

```
#####
####          NUMERICAL OUTPUT          ####
#####

===== First submodel =====

Linear mixed-effects model fit by REML
Data: dataz2
      AIC      BIC    logLik
175.0128 188.3299 -81.50642

Random effects:
Formula: ~1 | Hive
      (Intercept)  Residual
StdDev:   0.9666872 0.3373335

Fixed effects: Y1 ~ 1 + BeesN + Infection + BeesN:Infection
              Value Std.Error DF   t-value p-value
(Intercept)  2.643551 0.9281956 48   2.8480536 0.0065
BeesN        -0.000012 0.0000127 20  -0.9829792 0.3374
Infection1    3.646261 1.5420563 20   2.3645445 0.0283
BeesN:Infection1 -0.000016 0.0000234 20  -0.6790599 0.5049
Correlation:
              (Intr) BeesN  Infct1
BeesN        -0.971
Infection1   -0.602  0.585
BeesN:Infection1 0.528 -0.543 -0.932

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3      Max
-2.2650454 -0.4431810  0.1416440  0.5021554  1.6554665

Number of Observations: 72
Number of Groups: 24

----- ANOVA table for the submodel -----
              numDF denDF   F-value p-value
(Intercept)      1     48 123.94402 <.0001
BeesN             1     20   8.64062 0.0081
Infection         1     20 22.82242 0.0001
BeesN:Infection   1     20   0.46112 0.5049

===== Second submodel =====

Generalized least squares fit by REML
Model: Y1 ~ 1 + BeesN + Infection + BeesN:Infection
Data: dataz2
      AIC      BIC    logLik
251.5938 262.6914 -120.7969
```

```

Coefficients:
              Value Std.Error   t-value p-value
(Intercept)   2.643551 0.5264205   5.021748 0.0000
BeesN         -0.000012 0.0000072  -1.733209 0.0876
Infection1     3.646261 0.8745678   4.169214 0.0001
BeesN:Infection1 -0.000016 0.0000133 -1.197333 0.2353

Correlation:
              (Intr) BeesN  Infct1
BeesN         -0.971
Infection1    -0.602  0.585
BeesN:Infection1 0.528 -0.543 -0.932

Standardized residuals:
              Min          Q1          Med          Q3          Max
-2.1491390 -0.6020178 -0.2018275  0.5485022  2.6988650

Residual standard error: 0.9686788
Degrees of freedom: 72 total; 68 residual
----- ANOVA table for the submodel -----
Denom. DF: 68
              numDF  F-value p-value
(Intercept)      1 385.3355 <.0001
BeesN             1  26.8632 <.0001
Infection         1  70.9537 <.0001
BeesN:Infection   1   1.4336 0.2353

===== ANOVA table for the submodels =====
              Model df  AIC    BIC   logLik   Test  L.Ratio p-value
Submodel1 1 6 175.01 188.32 -81.506
Submodel2 2 5 251.59 262.69 -120.796 1 vs 2 78.58097 <.0001

```

The first part of the output is for the full model, the second part is for the nested model, and the last ANOVA table compares them. Note that we are testing on the boundary again. To correct the *p*-value, start R, and type:

```
> 0.5 * ((1-pchisq(78.58, 1)) + (1 - pchisq(78.58, 2)))
```

and press enter. For other data sets replace the 78.58. All output is explained in Chapters 5 and 19 in Zuur et al. (2009).

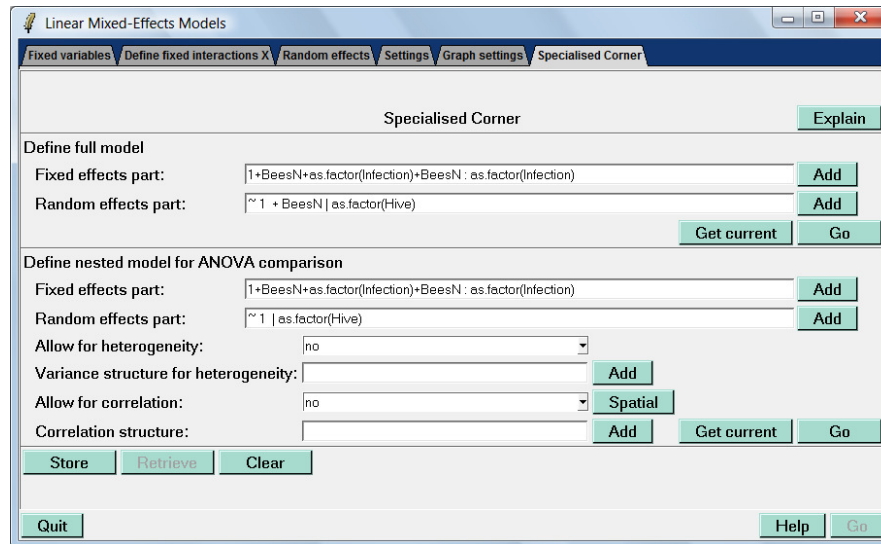


Figure 5.40. Settings to compare a random intercept and slope model versus a random intercept model.

### Step 7 – 9 of the protocol

In the second part of the protocol, we look for the optimal fixed structure, using the selected random structure from steps 1 – 6. There are three tools to assess the significance of explanatory variables:

1.  $t$ -values.
2.  $F$ -statistic using the ANOVA table (this test depends on the order of variables!).
3. The likelihood ratio test.

The first two are part of the output presented above. The likelihood ratio test can be carried out via the “Specialised corner”. Select the explanatory variables (and their interaction) in the “Variables” tab and the “Interaction” tab as described above. Ensure that ML estimation is selected in the “Settings” tab. Select the desired random effects in the “Random effects” tab. In the specialised corner, click both “Get current” buttons. Do not change the random part of the nested model; both models should have exactly the same random structure! Drop terms from the fixed effects part of the nested model (note that in the current model, you can only drop the interaction). If you have a model with multiple 2-way interactions, then the “Store” and “Retrieve” buttons may save some time.

## 5.9 GLS to model heterogeneity

Chapter 4 in Zuur et al. (2009) contains a detailed explanation of GLS. Using a squid data set, the weight of the testicles was modelled as a function of the mantel length and month. The data were clearly heterogeneous, and various variance structures were applied, namely:

- The varFixed variance
- The varIdent variance
- The varPower variance
- The varExp variance
- The varConstPower variance
- The varComb variance

These variance structures are also discussed in detail in Pinheiro and Bates (2000). We do not discuss the underlying theory here. The data can be downloaded from [www.highstat.com/SquidTestisWeight.xls](http://www.highstat.com/SquidTestisWeight.xls). The data were published in Smith et al. (2005); see also Section 4.1 in Zuur et al. (2009) for a more detailed biological explanation. Note that this is a different squid data set than the one used in Chapter 4 of this manual.

Import the data, select Testisweight as response variable, and DML (length) and Month as explanatory variables (ignore or deselect the variable year).

First apply a linear regression model in which testis weight is modelled as a function of DML (continuous), Month (factor), and their interaction. Note that there is considerable heterogeneity. Compare your graphical output with Figure 4.2A in Zuur et al. (2009). The Brodgar clicking process for this is explained in Section 5.1 of this manual.

We are now going to model the heterogeneity in the residuals by using models of the form:

$$\varepsilon_i \sim N(0, \sigma^2 \times DML_i) \quad i = 1, \dots, 768 \quad (4.2)$$

$$\varepsilon_{ij} \sim N(0, \sigma_j^2) \quad j = 1, \dots, 12 \quad (4.4)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2 \times |DML_{ij}|^{2\delta}) \quad (4.5)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2 \times |DML_{ij}|^{2\delta_j}) \quad (4.6)$$

$$\text{var}(\varepsilon_{ij}) = \sigma^2 \times e^{2\delta \times DML_i} \quad (4.7)$$

$$\text{var}(\varepsilon_{ij}) = \sigma^2 \times (\delta_1 + |DML_{ij}|^{\delta_2})^2 \quad (4.8)$$

$$\text{var}(\varepsilon_{ij}) = \sigma^2 \times (\delta_{1j} + |DML_{ij}|^{\delta_{2j}})^2 \quad (4.9)$$

$$\text{var}(\varepsilon_{ij}) = \sigma_j^2 \times e^{2\delta \times DML_{ij}} \quad (4.10)$$

We used the same equation numbers as in Zuur et al. (2009). The variance defined by Equation (4.2) is the varFixed variance, (4.4) is the varIdent, (4.5) is the varPower, (4.6) is also the varPower but with multiple parameters  $\delta_j$ , (4.7) is the varExp, (4.8) and (4.9) are the varConstPower, and (4.10) is the varComb variance structure. All these equations are fully explained and illustrated in Chapter 4 of Zuur et al. (2009).

To implement any of these variance structures in Brodgar, click on the “Univariate” main menu button, and then the “Mixed modelling, GLS & GAMM” tab, so that you get the window in Figure 5.31. Select the GLS option, and click the “Go” button. A similar menu appears as for linear regression. Select DML as continuous explanatory variable, Month as a nominal variable, and add the interaction term between these two variables, see also Figure 5.41.

To select one of the variance structures, click on the “Settings” tab in Figure 5.41; you get the window in Figure 5.42. The first option “Model fitting method by REML or ML” allows the user to select restricted maximum likelihood (REML) or maximum likelihood (ML). See Chapters 4 and 5 in Zuur et al. (2009) for a detailed discussion when to use which one. The second line, “Allowing for heterogeneity”, can be used to model any of the variance structures mentioned above. Change the “no” to one of the following options:

- Fixed weights
- Constant variance(s)
- Power of a variance covariate
- Exponential of a variance covariate
- Constant plus power of a variance covariate
- Combination of variance functions
- Custom

The “Fixed weights” option implements the fixed variance structure in Equation (4.2). However, it is not a variance structure that one would use as it is rather limited. However, if you want to do this one in Brodgar, change “no” to “Fixed weights” and for “Variance structure for heterogeneity”, ensure that the box contains the following text: DML. Do not type the dot; see our snapshot of part of the GLS settings menu in Figure 5.43. Note that there is a space (blank) in front of DML. Instead of typing it in, you can also use the “Add” button.



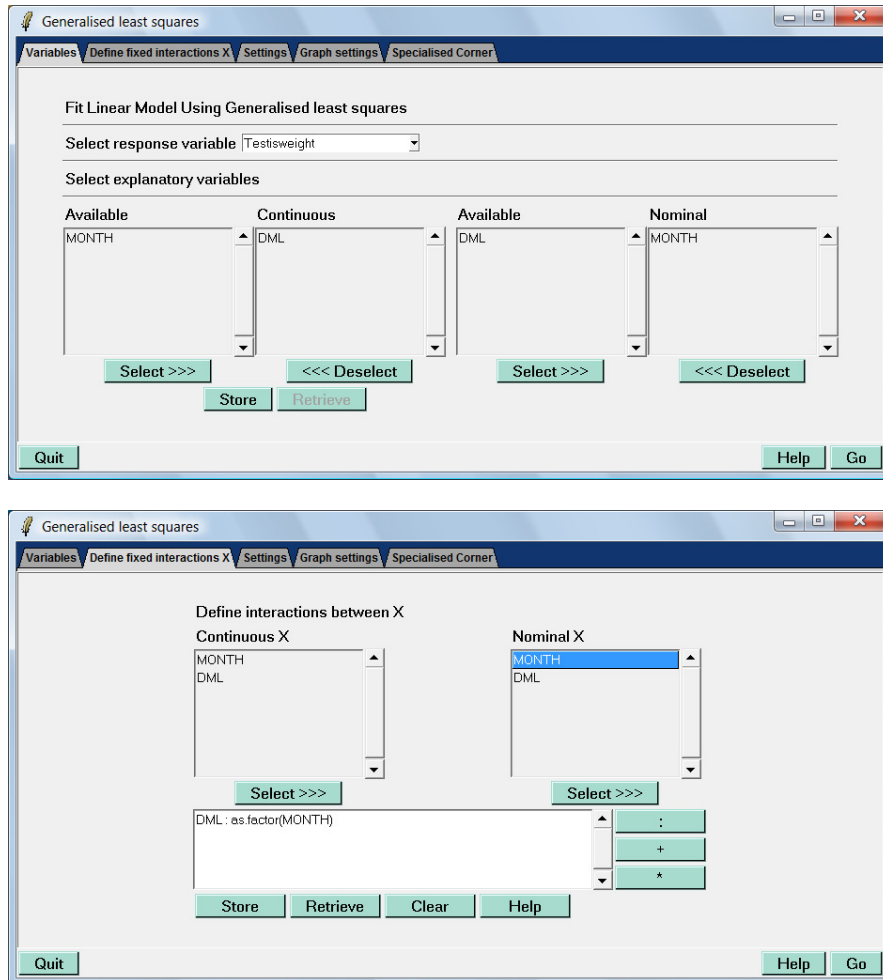


Figure 5.41. Selection of explanatory variables and interaction for a GLS model applied on the testis weight of squid.

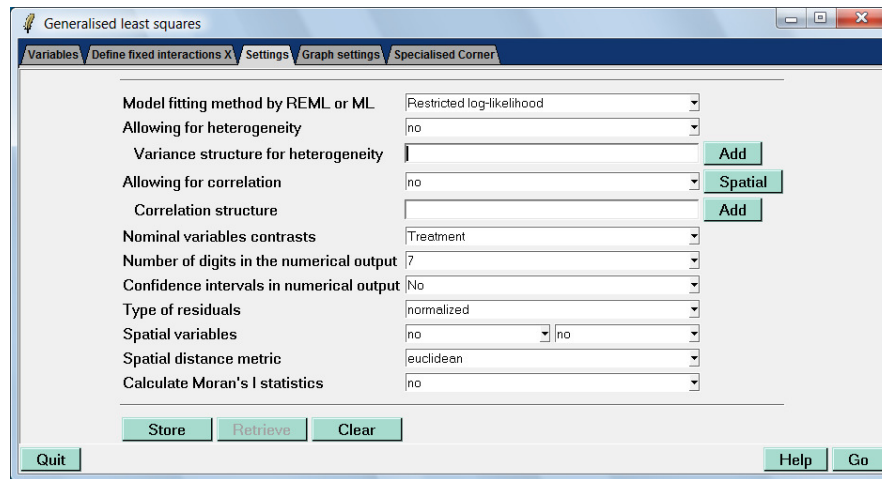


Figure 5.42. Options for the GLS model. The first row allows one to select REML or ML.



Figure 5.43. Part of the “Settings” menu for GLS. The selected options ensure that a GLS with the varFixed variance structure along DML is used, see also Equation (4.2). REML estimation is used as we are looking for the optimal random structure.

Other variance structures can easily be implemented by changing the settings in Figure 5.43. For example, the “Constant variance(s)” implements the varIdent in Equation (4.4). To select this variance structure, change the “no” in Figure 5.42 to “Constant variance(s)”. The difficult bit is the code for the “Variance structure for heterogeneity”. It has to become: `1|factor(MONTH)`, see Figure 5.44. The “Add” button only allows one to select MONTH as a nominal variable; this adds the code as `.factor(MONTH)`. You will have to add the `1|` bit yourself. Clicking the “Go” button gives exactly the same results as on page 76 in Zuur et al. (2009).

The varExp and the varPower functions can be run in a similar way as the varFixed; instead of “Fixed weights”, use “Exponential of a variance covariate” or “Power of a variance covariate” respectively in Figure 5.43. Ensure that DML is selected and not MONTH! To implement the model in Equation (4.6), see our set-

tings in Figure 5.45. You can either use the “Add” button twice (and typing the `|` symbol in between, or just type it all yourself. Do not make any mistakes!

To implement the `varComb` variance structure, change the “no” in Figure 5.42 to “Combination of variance structures”, and type on the next line:

```
varIdent(form =~ 1 | MONTH) , varExp(form =~ DML)
```

It is crucial not to make any mistake (R is capital sensitive)!

Model fitting method by REML or ML	Restricted log-likelihood	
Allowing for heterogeneity	Constant variance(s)	
Variance structure for heterogeneity	1 as.factor(MONTH)	<b>Add</b>

Figure 5.44. Part of the “Settings” menu for GLS. The selected options ensure that a GLS with the `varIdent` variance structure with MONTH is used, see also Equation (4.4). REML estimation is used as we are looking for the optimal random structure.

Model fitting method by REML or ML	Restricted log-likelihood	
Allowing for heterogeneity	Power of a variance covariate	
Variance structure for heterogeneity	DML   as.factor(MONTH)	<b>Add</b>

Figure 5.45. Part of the “Settings” menu for GLS. The selected options ensure that a GLS with the `varPower` variance structure with multiple parameters is used, see also Equation (4.6). REML estimation is used as we are looking for the optimal random structure.

Model fitting method by REML or ML	Restricted log-likelihood	
Allowing for heterogeneity	Combination of variance functions	
Variance structure for heterogeneity	varIdent(form =~ 1   MONTH),varExp(form =~ DML)	<b>Add</b>

Figure 5.46. Part of the “Settings” menu for GLS. The selected options ensure that a GLS with the `varComb` variance structure is used, see also Equation (4.10). REML estimation is used as we are looking for the optimal random structure. The text on the third line reads:

```
varIdent(form =~ 1 | MONTH) , varExp(form =~ DML)
```

All these GLS models with different variance structures can be compared using for example the AIC, see page 81 in Zuur et al. (2009). In this case, the model in Equation (4.6) has the lowest AIC. You can now go to steps 6 – 10 of our protocol, and look for the optimal fixed structure. This can either be done using the  $t$ -values of the estimated parameters (using REML estimation), the  $F$ -statistic (also using REML) from the anova table (at the bottom of the output), or use the likelihood ratio statistic. The later one can be estimated in Brodgar using the specialised corner, see Figure 5.47. If you click on both “Get current” buttons, Brodgar will fill in the fields with your current selected model. If the aim is to apply a likelihood ratio test to test the interaction term, then remove the interaction term from the nested model (don’t forget to remove the + symbol), and also change the estimation method from REML to ML in the “Settings” menu in Figure 5.42. Clicking on the lower right “Go” button in Figure 5.47 gives the following numerical output:

```
===== ANOVA table for the submodels =====
      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Submodell1  1 37 3276.58 3448.409 -1601.294
Submodell2  2 26 3451.54 3572.282 -1699.772 1 vs 2 196.9548 <.0001
```

The specialised corner can also be used to compare a model with a certain variance structure (as defined in the “Settings” menu in Figure 5.42), and a nested model with no variance structure. Ensure that REML estimation is used and that both models have the same fixed structure (main terms and interaction).

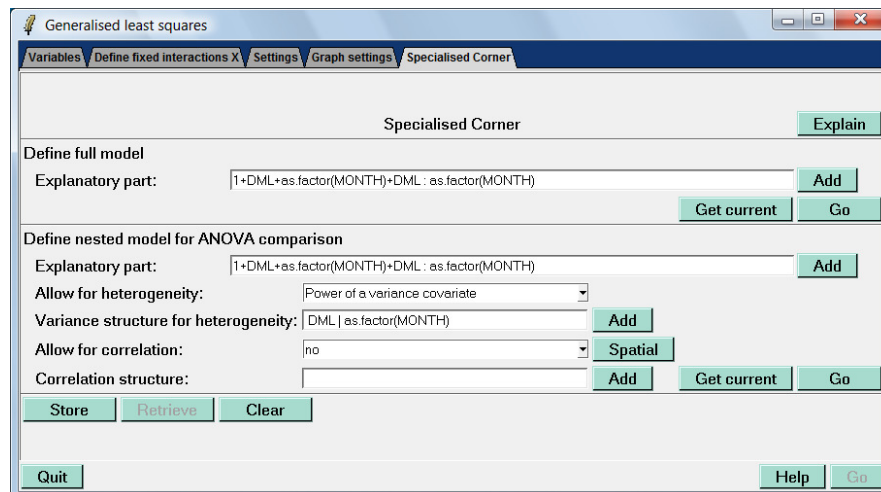


Figure 5.47. The specialised corner. We clicked on both “Get current” buttons. From the nested model, remove the interaction term + DML : as.factor(MONTH). Ensure that under the “Settings” menu, ML estimation is used, and not REML.

## 5.10 GLS example from Zuur et al. (2009)

In Section 4.2 in Zuur et al. (2009), data from a replicate mesocosm experiment were analysed. Biological details and references can be found in Section 4.2.1.

At the start of the experiment, each container was filled with homogenized sediment from mudflats on the Ythan estuary (Scotland, UK). The macrofaunal biomass (*H. diversicolor*) was fixed across the following levels (0, 0.5, 1, 1.5 and 2 grams), and replicated within each biomass level ( $n = 3$ ). The response variable is the concentration of a particular nutrient. We modelled the concentration data from all three nutrients, as a function of one continuous explanatory variable (biomass), and two nominal explanatory variables: enrichment (with or without algae), and a variable identifying the nutrient with the levels NH<sub>4</sub>-N, NO<sub>3</sub>-N and PO<sub>3</sub>-P.

The data can be downloaded from the file [www.highstat.com/Biodiversity.xls](http://www.highstat.com/Biodiversity.xls). The first column contains data; not labels! The response variable is Concentration and biomass is a continuous explanatory variable. Treatment (enrichment) and Nutrient are the categorical explanatory variables. See the comments in the Excel spreadsheet for the interpretation of the values of the categorical variables.

Apply a data exploration and reproduce Figure 4.7 in Zuur et al. (2009). Also apply linear regression to model Concentration as a function of Biomass, Treatment, Nutrient, all 2-way interactions and the 3-way interaction. The window with the interactions should read:

```
Biomass : as.factor(Treatment) +
Biomass : as.factor(Nutrient) +
as.factor(Treatment) : as.factor(Nutrient) +
Biomass : as.factor(Treatment) : as.factor(Nutrient)
```

Don't forget the select the main terms! As part of the model validation process, Brodgar produces Figure 4.8, which clearly shows heterogeneity. In Section 4.2.2, Zuur et al. (2009) apply four GLS models:

- A GLS with no variance structure.
- A GLS with the varIdent using Nutrient as variance covariate.
- A GLS with the varIdent using Treatment as variance covariate.
- A GLS with the varIdent using Nutrient and Treatment as variance covariate.

All these models can easily be applied in Bodgar. Go to the GLS menu, and select the same explanatory variables and interactions. First run the model without any variance structure (this is just linear regression executed via the GLS menu). Do this with the option “no” for “Allowing for heterogeneity” in Figure 5.42. Write down the AIC! The other three models are executed with the option “Constant variance(s)” and the following variance structure for heterogeneity:

- `1|as.factor(Nutrient)`
- `1|as.factor(Treatment)`
- `1|as.factor(Treatment) * as.factor(Nutrient)`

Just choose one of them, type it in (without the black bullet from Word), run the model, and write down the AIC. Ensure that REML estimation is selected! You should get the AICs presented on page 89 in Zuur et al. (2009). You can also apply a likelihood ratio test to compare the model with a varIdent variance structure and a nested model without a variance structure. Do this in the specialised corner, and ensure that REML is selected and that both models have the same fixed structure. From the sub-model, remove the variance structure, see Figure 5.48

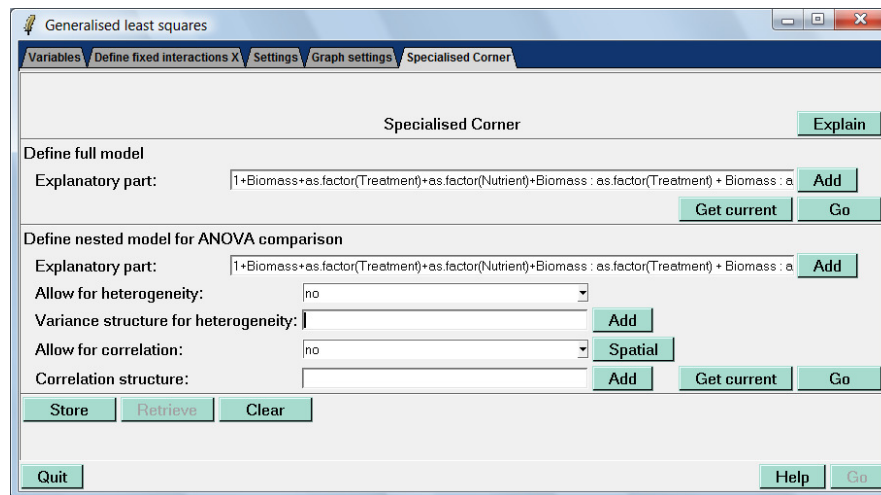


Figure 5.48. Comparison of two GLS models. The full model contains all main terms, 2-way interactions and the 3-way interaction, and the varIdent variance structure with Nutrient and Treatment as variance covariates. The nested model is identical, except that it does not contain a varIdent variance structure. REML estimation is selected under the “Settings” tab.

In Subsection 4.2.3 of Zuur et al. (2009), a 10-step protocol is developed, and it is applied on the diversity data in subsection 4.2.4. The first part of the protocol deals with finding the optimal random structure, and we have already carried out this step. In the second part of the protocol, the optimal fixed structure is determined. We have three options to assess the significance of the terms in the fixed part, namely the *t*-statistics, the *F*-statistic and likelihood ratio test. The first two are part of the output that we already have (assuming REML was used):

...

```
Variance function:
Structure: Different standard deviations per stratum
Formula: ~1 | Treatment * Nutrient
```

```

Parameter estimates:
      0*1      1*1      0*2      1*2      0*3      1*3
1.0000000 0.5722142 1.4163869 9.4464105 0.4997456 1.2528436

Coefficients:
              Value Std.Error   t-value p-value
(Intercept)    2.879333 0.3214829   8.956411 0.0000
Biomass        -0.596000 0.2700997  -2.206593 0.0297
Treatment1     -2.772667 0.3703937  -7.485729 0.0000
Nutrient2      -2.396306 0.5573954  -4.299113 0.0000
Nutrient3      -2.465118 0.3593923  -6.859129 0.0000
Biomass:Treatment1  0.601000 0.3111929   1.931278 0.0564
Biomass:Nutrient2  1.196733 0.4683058   2.555453 0.0122
Biomass:Nutrient3  0.857107 0.3019499   2.838572 0.0055
Treatment1:Nutrient2 16.022361 3.0930643   5.180093 0.0000
Treatment1:Nutrient3  4.592260 0.5702849   8.052571 0.0000
Biomass:Treatment1:Nutrient2 0.846933 2.5986938   0.325907 0.7452
Biomass:Treatment1:Nutrient3 -0.675893 0.4791352  -1.410653 0.1616

...

Residual standard error: 0.7396984
Degrees of freedom: 108 total; 96 residual

----- ANOVA table for the submodel -----
Denom. DF: 96
              numDF   F-value p-value
(Intercept)         1 205.73774 <.0001
Biomass              1  1.22179  0.2718
Treatment            1 14.62898  0.0002
Nutrient             2  1.57754  0.2118
Biomass:Treatment    1  0.26657  0.6068
Biomass:Nutrient     2  4.17802  0.0182
Treatment:Nutrient   2 121.57149 <.0001
Biomass:Treatment:Nutrient 2  1.09043  0.3402

```

Estimated variances,  $t$ -values and  $F$ -statistics (using sequential testing) are presented. Note that the  $F$ -statistic indicated that the 3-way interaction is not significant at the 5% level. The likelihood ratio test required a bit more fiddling. Go to the specialised corner, and:

- Click both “Get current” buttons.
- Remove the 3-way interaction from the nested model.
- Set the estimation technique to ML.
- Ensure that both models have the same random structure.
- Click the second „Go“ button.

This should give the following output:

```

===== ANOVA table for the submodels =====
      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Submodel1  1 18 321.0648 369.3432 -142.5324
Submodel2  2 16 319.4653 362.3794 -143.7327 1 vs 2 2.400507 0.3011

```

Hence, the likelihood ratio test also indicates that the 3-way interaction is not significant. Drop it from the full model (in the Interaction menu tab), and compare

the new full model (that has three main terms and three 2-way interaction terms) to each of the following sub-models:

1. The model without the Biomass : Treatment interaction.
2. The model without the Biomass : Nutrient interaction.
3. The model without the Nutrient : Treatment interaction.

Drop the term that is the least significant, and continue until everything is significant, see pages 93 – 97 in Zuur et al. (2009). You should get the same output. The “Store” and “Retrieve” buttons may save some time!

### 5.11 GAMM (or GLS) and auto-regressive correlation

In Chapter 36 of Zuur et al (2007), time series of water birds are analysed using dynamic factor analysis, chronological clustering and additive mixed modelling (AMM). Here, we reproduce some of the AMM results, and show how to do it in Brodgar.

The data can be downloaded from [www.highstat.com/waterbirdislandseries.xls](http://www.highstat.com/waterbirdislandseries.xls). The first column contains data, not labels. The response variable is the column labelled Birds, and all other variables are explanatory variables. Apply a square root transformation on the birds. As part of a data exploration, try to reproduce Figures 36.2 and 36.3; both graphs are available from the “Exploration” menu.

In Section 36.5, various AMMs are applied. Here, we apply the following model:

$$\text{Birds}_t = \text{constant} + f_1(\text{Rainfall}_t) + f_2(\text{Time}_t) + \varepsilon_t$$

where  $f_1()$  and  $f_2()$  are smoothing functions and  $\varepsilon_t$  is normally distributed noise. The index  $t$  refers to year. Just like Section 36.5, we impose an auto-regressive correlation structure of order 1 (AR1) on the residuals:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \gamma_t$$

The error term  $\varepsilon_t$  is now allowed to be correlated with noise from previous years and  $\gamma_t$  is independently normally distributed noise. R uses the same correlation parameter  $\rho$  for all time series, hence we should use:

$$\varepsilon_{ti} = \rho\varepsilon_{t-1,i} + \gamma_{ti}$$

To fit this model in Brodgar, select the GAMM option in Figure 5.31, and click the “Go” button. Select the smoothers as for ordinary GAM, see our selection in Figure 5.49. The correlation structure can be selected via the “Settings” menu in Figure 5.49.



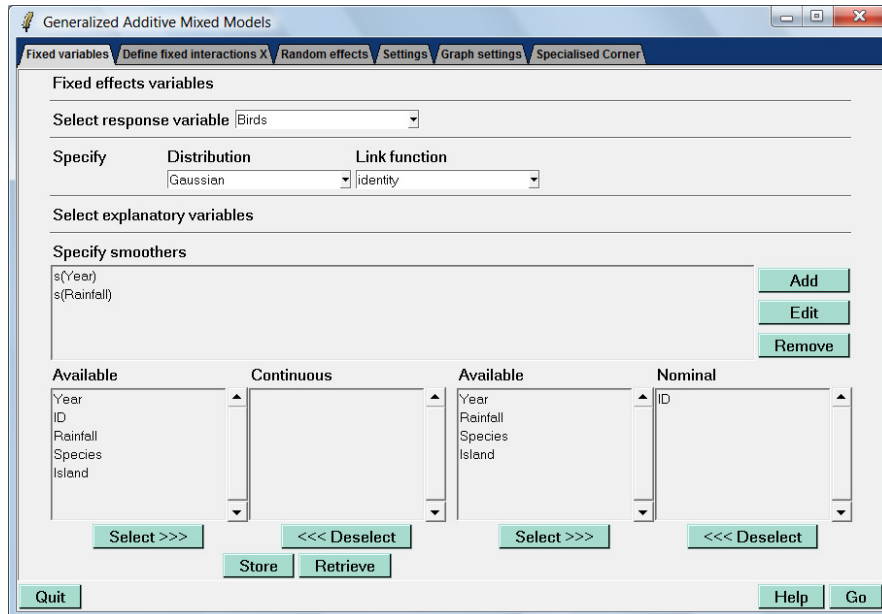


Figure 5.49. Selection of smoothers for the water bird data. Two smoothers are selected, one for Year and one for Rainfall. We also added the categorical explanatory variable ID.

In order to specify the correlation structure, change the “no” for “Allowing for correlation” to “Autoregressive process of order 1”. For most regular spaced data sets, this correlation structure suffices. We also need to specify the correlation structure. If you have only one univariate time series, just type: `Year`. R then knows that the order of the data is specified by the variable `Year`, and uses this to calculate the parameter  $\rho$ . The problem here, is that the correlation is inside a particular time series, hence we have to type: `Year | as.factor(ID)`. An error message will occur if you do not use the `| as.factor(ID)` bit. Note that you cannot use the AR1 correlation structure with multiple observations at the same time. The numerical output produced by the `mgcv` package in R is quite technical, and not everything is equally important. See Chapters 5 and 6 in Zuur et al. (2009) for a detailed explanation of the numerical output of GAMMs. The water bird data are also analysed in detail in Chapter 6 (and 7) of Zuur et al. (2009).

Besides adding a correlation structure, one can also allow for heterogeneity, e.g. a different variance per time series using the `varIdent` variance structure with ID as variance covariate. This can also be done in Figure 5.50, see also the GLS section in this chapter.

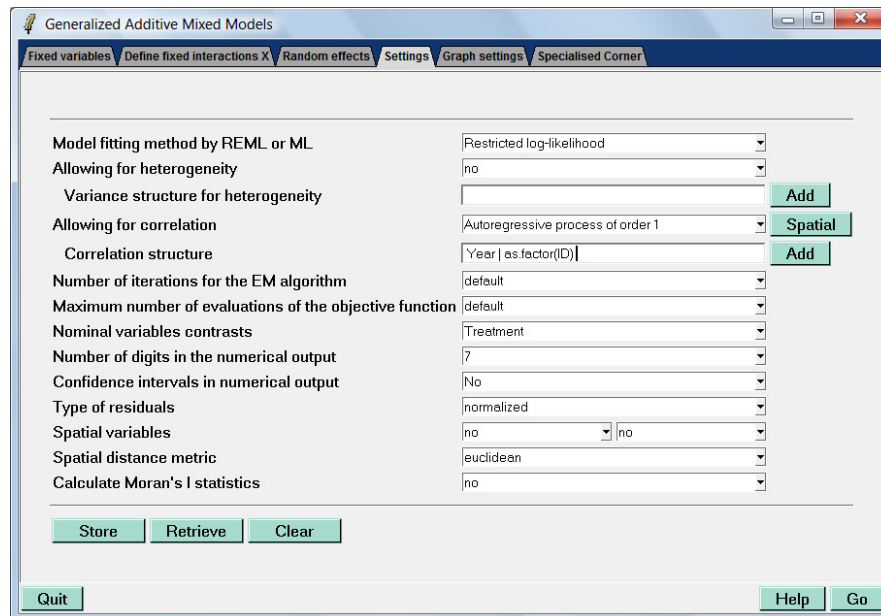


Figure 5.50. Adding a residual auto-correlation structure to the AMM. The “no” for “Allowing for correlation” is changed into “Autoregressive process of order 1”. The correlation structure reads as: `Year | as.factor(ID)`.

## 5.12 GAMM (or GLS) and other correlation structures

In Chapter 6 of Zuur et al. (2009), regular spaced time series are analysed using GLS and additive models. A residual auto-correlation was added to these models using either the AR1 or the ARMA structure. In the previous subsection, we showed how to add the AR1 residual correlation. The clicking process for a GLS is identical, except that you have to do it via the GLS menu.

To use the ARMA correlation structure for the water bird data, change the “no” option for “Allowing for correlation” to “custom”. And on the next line, type (without making a mistake):

```
corARMA(form=~Year | as.factor(ID) , p=2, q=0)
```

It is important to write the correlation function with capital ARMA. The form argument tells R that the correlation is between sequential observations, as defined by the variable Yera, within a particular time series ID. If you have only one univariate time series, omit the `as.factor(ID)`. The  $p$  and  $q$  specify the number of AR and MA parameters respectively. Numerical problems may occur for relative large values of  $p$  and  $q$ . Our settings are shown in Figure 5.51.

Model fitting method by REML or ML	Restricted log-likelihood	
Allowing for heterogeneity	no	
Variance structure for heterogeneity		Add
Allowing for correlation	Custom	Spatial
Correlation structure	corARMA(form=~Year   as.factor(ID) ,p=2,q=0)	Add

Figure 5.51. Snapshot of part of the “Settings” menu to incorporate a ARMA residual correlation structure to an additive model. The correlation structure reads as: `corARMA(form=~Year | as.factor(ID) ,p=2,q=0)`.

In Chapter 7 of Zuur et al. (2009) irregular spaced data were used and different residual correlation structures were implemented in the GLS and additive models, namely:

- Exponential correlation using the function `corExp`.
- Gaussian correlation using the function `corGaus`.
- Linear correlation using the function `corLin`.
- Rational quadratic correlation using the function `corRatio`.
- Spherical correlation using the function `corSpher`.

Each of these options implies a specific mathematical model for the correlation structure. It is relatively easy to implement any of these correlation structures in Brodgar. Just change the “Autoregressive process of order 1” in Figure 5.50 to:

- Exponential spatial correlation for the `corExp` structure
- Gaussian spatial correlation for the `corExp` structure
- Linear spatial correlation for the `corLin` structure
- Rational spatial correlation for the `corRatio` structure
- Spherical spatial correlation for the `corSpher` structure

Note that the correlation structure should still be: `Year | as.factor(ID)`. Ensure that REML estimation is selected, and compare the different correlation structures with, for example, the AIC. The main advantage of these correlation structures is that they can be used to model irregular spaced data.

By default, the spatial correlation structures use a range, but no nugget. If you want to use a nugget in the correlation structure, the clicking process in Brodgar is slightly more complicated. Suppose you want to use the `corExp` correlation structure with a nugget. Click the “Spatial” button in Figure 5.50; you are presented with Figure 5.52. All the options look intimidating, but it is not that difficult. The first part shows the model definitions; all values are filled in and there is no need to change anything. For “Specify spatial variables(s)”, select Year in the first box. If for your own data, you have spatial coordinates in the form of northing and easting (or latitude and longitude), then select them here (one in each box). For the

water bird data, the correlation is inside a specific time series, and therefore we have to select ID for “Grouping factor”. Leave it to “no” if you do not have nested data. The last thing we have to do is selecting the type of correlation structure; choose the “Exponential spatial correlation”. If you now click on “Apply to model and Quit”, Brodgar will use the custom option for the spatial correlation, with the following correlation structure:

```
corExp(value=numeric(0), form=~ Year | ID, nugget=T,
       metric="euclidean", fixed=F)
```

This is fancy R code for a corExp correlation structure that contains a nugget and range. Note that it will estimate the values of the nugget and range. You can also set these to a fixed value using the “Range” and “Nugget” entryboxes in Figure 5.52. Brodgar can even help you choosing these values with the “Get the following values from graph” button. This is an option if numerical (convergence) problems occur.

Figure 5.52. Adding a spatial correlation structure with a nugget to a GLS or additive model.

Model fitting method by REML or ML	Restricted log-likelihood	
Allowing for heterogeneity	no	
Variance structure for heterogeneity		Add
Allowing for correlation	Custom	Spatial
Correlation structure	corExp (value=numeric(0), form=~Year ID, nugget=	Add

Figure 5.53. Snapshot of the updated “Settings” menu for the additive model with a corExp variance structure allowing for a nugget effect. Note that not all code for the correlation structure is visible.

## 5.13 GAMM and random effects

Roulin and Bersier (2007) analysed the begging behaviour of nestling barn owls. They looked how nestlings respond to the presence of the father and of the mother. Using microphones inside and a video outside the nests, they sampled 27 nests, and studied vocal begging behaviour when the parents bring prey. The data were used in Chapter 5 of Zuur et al. (2009) to illustrate mixed effects modelling and additive mixed effects modelling. The mixed effects model was of the form:

$$\begin{aligned} \text{LogNeg}_{ij} = & \alpha + \beta_1 \times \text{SexParent}_{ij} + \beta_2 \times \text{FoodTreatment}_{ij} + \beta_3 \times \text{ArrivalTime}_{ij} + \\ & \beta_4 \times \text{SexParent}_{ij} \times \text{FoodTreatment}_{ij} + \beta_5 \times \text{SexParent}_{ij} \times \text{ArrivalTime}_{ij} + \\ & a_i + \varepsilon_{ij} \end{aligned}$$

$\text{LogNeg}_{ij}$  is the log-10 transformed sibling negotiation for observation  $j$  at nest  $i$ .  $\text{SexParent}_{ij}$  and  $\text{FoodTreatment}_{ij}$  are nominal variables with two levels, and  $\text{ArrivalTime}_{ij}$  is a continuous variable. The second line contains interactions. The term  $a_i$  is a random intercept for nest, and is assumed to be normally distributed with mean 0 and variance  $d^2$ . The residual  $\varepsilon_{ij}$  is assumed to be normally distributed with mean 0 and variance  $\sigma^2$ .

A detailed analysis is presented in Chapter of Zuur et al. (2009), and is not repeated here. We encourage the reader to reproduce Figures 5.4 – 5.7, and the 10 steps of the protocol. In this manual, we show how to fit the following additive mixed effects model (also applied in Zuur et al., 2009):

$$\text{LogNeg}_{ij} = \alpha + \beta_2 \times \text{FoodTreatment}_{ij} + f(\text{ArrivalTime}_{ij}) + a_i + \varepsilon_{ij}$$

The term  $\beta_3 \times \text{ArrivalTime}_{ij}$  has been replaced by  $f(\text{ArrivalTime}_{ij})$ , which is now a smoother.

The data can be found in the file [www.highstat.com/Owls.xls](http://www.highstat.com/Owls.xls). The response variable is in the column NegPerChick, and the explanatory variables are ArrivalTime, SexParent, FoodTreatment and Nest. The last three are categorical. Ignore the variables BroodSize and SiblingNegotiation in this analysis (these are used in

other Chapters in Zuur et al., 2009). Note that the first column contains data, and apply a  $\log(Y+1)$  transformation on the response variable NegPerChick.

To implement the additive mixed effects model with nest as a random effect, go to the GAMM menu, select ArrivalTime as a smoother, SexParent and Food-Treatment as nominal explanatory variables, and add the interaction between these two nominal variables. To select the random effect nest, click on the “Random effects” tab, and select nest as nested grouping factor. Clicking the “Go” button gives exactly the same output as in Section 5.10 in Zuur et al. (2009). A full explanation of all the numerical output is provided in that section as well.

## 5.14 Multinomial logistic regression

Application of this technique in Brodgar is identical as linear regression or logistic regression, except that the response variable needs to be a variable with values 0, 1, 2, 3, etc. Hence, in logistic regression, we use a response variable consisting of 0 and 1, here it contains more levels. An application is given on pages 158 – 161 in Zuur et al. (2007). The data for this example are available from [www.highstat.com/DitchData.xls](http://www.highstat.com/DitchData.xls). The response variable is Site (the first column contains data!). Note that we ignored (perhaps wrongly) the year and month information in the analysis.