Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 18 Offprint

Statistical Modelling

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

Chapter 18

Statistical Modelling

As we said in Chapter 2, the principal methodologies of statisticians are functional methods that aim to explain response variables in terms of a set of explanatory variables (also called *predictors),* i.e. the typical regression situation. In this book we have been concentrating on structural methods of particular importance for ecological data analysts – this is mainly because of the large numbers of response variables observed in ecological studies, and the ensuing need for dimension reduction. In this chapter we intend to give a short overview of the sort of functional methods that are in use today when there is one response variable of interest, emphasising that this is a very brief description of a topic that deserves a book by itself. We start with several variants of linear regression, gathered together under the collective title of generalized linear models. These approaches all achieve a mathematical equation that links the mean of the response variable with a linear function of the explanatory variables. We shall also give some glimpses of two alternative nonparametric approaches to modelling: generalized additive modelling, which replaces the linear function of the predictors with a much freer set of functional forms, and classification and regression trees, which take a radically different approach to relating a response to a set of predictors and their interactions, in the form of a decision tree.

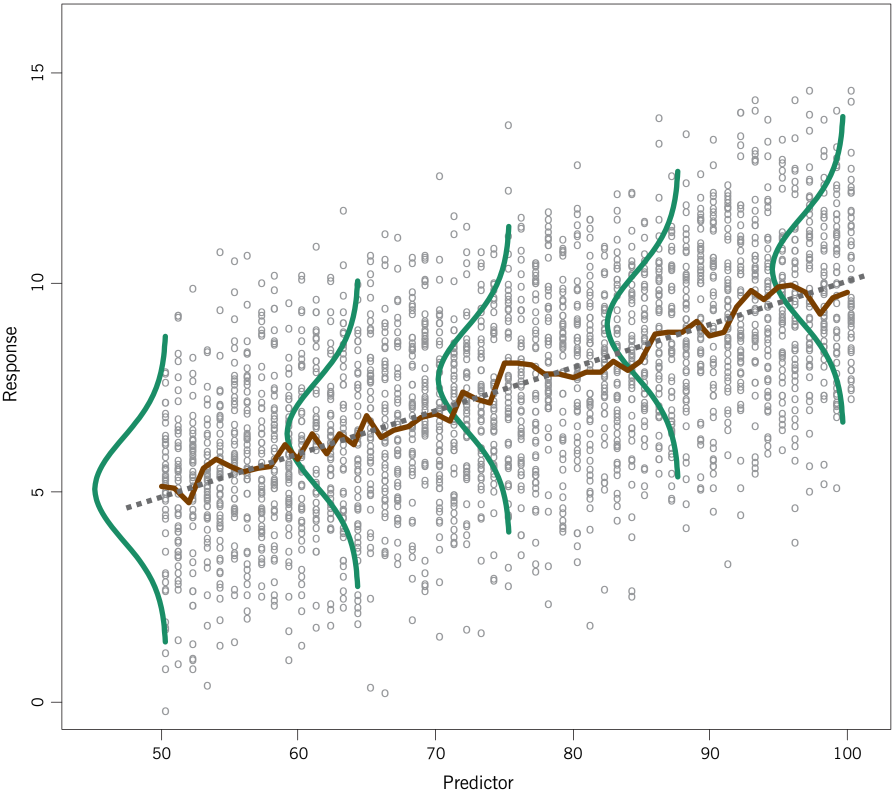
Contents

Multiple linear regression

Multiple regression is a model for the *conditional* mean of a response variable y given a set of explanatory variables x_1, x_2, …, x_m. To explain this statement and the assumptions of the regression model, we consider the simple case when there is only one explanatory variable x. Exhibit 18.1 shows an explanatory variable x (which could be depth, for example) that can take on the integer values from 50 to 100, and for each value of x there are many values of y (which could be pollution). The brown line trajectory connects the means of y in every subsample of points corresponding to a given value of x. For each x we can imagine the total population of values of y, and each of these populations has a probability distribution, called a *conditional distribution* because it depends on x. Each of these conditional distributions has a mean and a variance and if we connected the means of all these conditional distributions (as opposed to the sample means that are connected by the brown lines) we would have what is called the *regression* of y on x, denoted by $\mu(x)$. Multiple linear regression has the following assumptions:

1. The regression function (i.e., means of the conditional distributions for all values of x) is linear, in this case a straight line: $\mu(x) = a + bx$.

Exhibit 18.1:
Many observations of a response variable y for different integer values of a predictor x. Every set of y values for a given x is a sample conditional distribution, with a mean and variance, and the brown trajectory links the conditional sample means. Multiple linear regression assumes that the data are from normal distributions conditional on x (a few are shown in green), with conditional means modelled as a straight line and with constant variances



2. The conditional distribution for any *x* is normal, with mean equal to a linear function of *x* (i.e., assumption 1), and variance equal to a constant value across all the *x* values (we say that the variances are *homoscedastic,* as opposed to *heteroscedastic* if the variances change with *x*).

In Exhibit 18.1 the estimated linear regression function is shown by a dashed line, as well as a few examples of the conditional distributions – since the *y* values are on the vertical axis the distribution functions are shown on their sides. In this example there would be a conditional distribution for each value of *x* and the regression line is the assumed model for the means of these distributions, called the *conditional means.*

There are more than 2,000 sample points in Exhibit 18.1 and we would seldom get such a large sample – rather, we would get a sample of size 50, say, as shown in Exhibit 18.2, but the assumptions of linearity and variance homogeneity remain exactly the same. The analysis estimates the linear regression relationship, as shown, which is used firstly for interpreting the relationship between

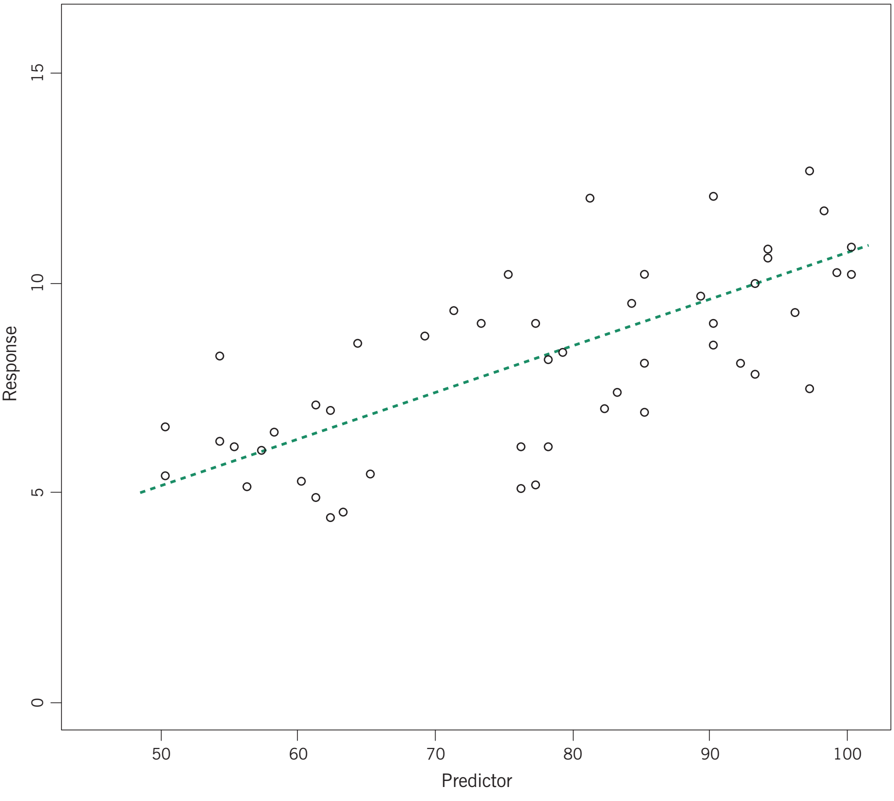


Exhibit 18.2:
A sample of 50 observations of the response y *and predictor* x*, showing the estimated regression line*

response and predictor, and secondly for predicting y for given x. Notice that when a response is predicted from a given value of x, it is the mean of the conditional distribution that is being predicted. In Exhibit 18.2, for example, an x value of 80 predicts a y value of about 7.89. One has then to imagine all the possible y values that constitute the population, possibly infinite, that could be observed for $x = 80$, and then the value 7.89 is the predicted mean of these. Regression analysis includes the technology for setting confidence limits around these predicted means.

Performing the regression on the above sample of 50 observations gives the following regression equation with standard errors (in square brackets) and p-values (in round brackets) associated with each coefficient:

$$\begin{array}{cccccc} \text{mean of } y & = & 0.392 & + & 0.0937x & \quad (18.1) \\ & & [1.587] & & [0.0218] & \\ & & (p = 0.81) & & (p < 0.0001) & \end{array}$$

The statistical conclusion would be that the constant term (i.e. intercept) is not significantly different from zero, while the predictor (i.e. slope) is highly significant. This agrees with the way these data were simulated: the conditional mean of y was set as the linear function $0.1x$ with no constant term, and the confidence interval for the coefficient of x, based on the estimate and the standard error (where the confidence interval is about 2 standard errors about the mean), does include the true value 0.1.

Poisson regression

The regression model can be generalized to the situation where the responses are of different data types, for example count and categorical variables. This general family of methods is called *generalized linear modelling* (GLM, for short). We first consider the case of a count response, which can be modelled using *Poisson regression*. Exhibit 18.3 shows some count data (for example, abundance counts, where only counts of 0 to 5 were observed here) recorded for different values of the predictor x. Theoretically again, we could have an infinite number of count observations for each x value, and the assumption is that for each x the counts follow the natural distribution for count data, the Poisson distribution, with a mean that depends on x (three examples of conditional distributions are shown in Exhibit 18.3, for $x = 50$, 75 and 100). Because a count variable is considered to be a ratio variable, it is the logarithm of the conditional means of the Poisson distribution that is modelled as a linear function: $\log(\mu(x)) = a + bx$ (see Chapter 3 where we discussed relationships of this type, where an additive change in x would imply a multiplicative change in the mean count). Notice in Exhibit 18.3

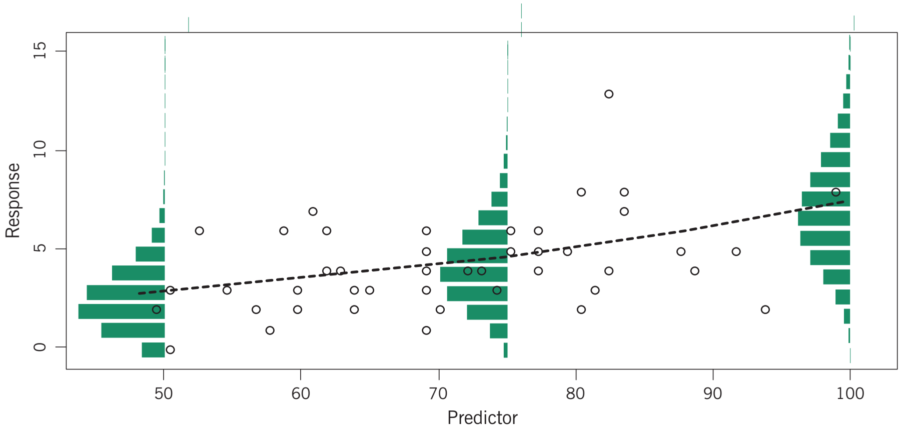


Exhibit 18.3: *Poisson regression, showing some observed response count data, three examples of the conditional distributions for x = 50, 75 and 100, assumed to be Poisson, shown in green, with the dashed line showing the estimated regression relationship of the means, where log mean is modelled as a linear function of the predictor*

that the conditional means are increasing and that the variance of the Poisson distributions is increasing accordingly.

Performing the Poisson regression on the above sample of 50 observations gives the following regression equation with standard errors (in square brackets) and p-values (in round brackets) associated with each coefficient:

$$\begin{array}{rcccc} \log(\text{mean of } y) & = & 0.0822 & + & 0.0183x \\ & & [0.3785] & & [0.0046] \\ & & (p = 0.83) & & (p < 0.0001) \end{array} \qquad (18.2)$$

Again, the statistical conclusion would be that the constant term is not significantly different from zero, while the predictor is highly significant. This agrees with the way these data were simulated: the log of the conditional mean of y was set as the linear function $0.02x$ with no constant term, and the confidence interval for the coefficient of x, based on the estimate and the standard error, does include the true value 0.02. The interpretation of the estimated coefficient 0.0183 is that for every unit increment in x, the log mean is increased by 0.0183, that is the mean is multiplied by $\exp(0.0183) = 1.0185$, or an increase of 1.85%. Notice how the slope of the regression curve is increasing (i.e., the curve is convex) due to the multiplicative effect of the predictor.

Logistic regression

As a final example of a generalized linear model, consider a dichotomous response variable, for example presence/absence of a species, and consider observations of this response along with associated predictor values, shown in Exhibit 18.4. Now the

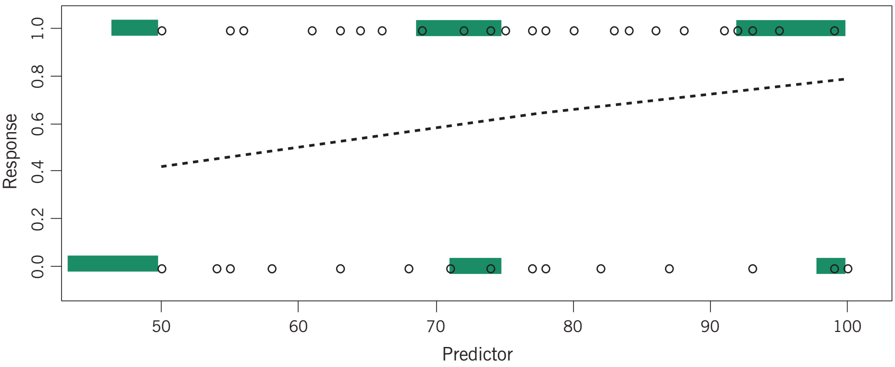


Exhibit 18.4: *Logistic regression, showing some observed dichotomous data, three examples are shown in green of the conditional probabilities of a 1 and a 0, for* x *= 50, 75 and 100 (remember again that these probability distributions are shown on their sides, in this case there are only two probabilities for each distribution, the probability of a 0 and the probability of a 1). The dashed line shows the estimated regression of the means, in this case probabilities, where the logits of the probabilities are modelled as linear functions of the predictor*

observations are only 0s (absences) and 1s (presences) and there could be some repeated observations.

Performing the logistic regression on the above sample of 50 observations gives the following regression equation with standard errors (in square brackets) and p-values (in round brackets) associated with each coefficient:

$$\begin{array}{lcccc} \text{logit}(p) = \log\left(\dfrac{p}{1-p}\right) & = & -3.568 & + & 0.0538x \\ & & [1.523] & & [0.0219] \\ & & (p = 0.019) & & (p = 0.014) \end{array} \qquad (18.3)$$

The statistical conclusion would be that both the constant term and the predictor are significant. This agrees with the way these data were simulated: the conditional mean of y was set as the linear function $-2 + (1/30)\,x$, where $1/30 = 0.0333$, and the confidence intervals for both coefficients do include the true values, but do not include 0.

Explained variance, deviance and AIC

The measure of explained variance in linear regression is well-known and we have used the concept many times in other contexts as well, for example the variance (or inertia) explained by the dimensions of a PCA, LRA, CA or CCA solution. *Deviance* is the generalization of this concept when it comes to GLMs. Without defining deviance mathematically, it functions in the same way: first, there is the concepts of the *full* (or *saturated*) *model*, where the response is fitted perfectly, and the *null model*, where no explanatory variables are fitted at all, with just the constant term being estimated. This difference is used as a measure of total variance and is, in fact, equal to the total variance in the case of linear regression. Deviance is used to measure the difference between models (i.e. hypotheses) in their ability to account for variation in the response variable. For example, if an explanatory variable is introduced in the GLM, deviance measures the amount of variance accounted for by this new variable compared to the null model. If a second variable is then introduced, one can compute the deviance between the one-variable model and the two-variable model to ascertain if it is worth introducing the second variable. Deviance has an approximate chi-square distribution, with degrees of freedom depending on the difference in numbers of parameters between two models. Let us take the logistic regression just performed above as an example. The *null deviance* for a model with no variables turns out to be equal to 68.99, and the *residual deviance*[1] for a model with the single predictor x is equal to 61.81.The deviance is thus the difference 68.99 – 61.81 = 7.18, which is from a chi-square distribution with 1 degree of freedom (the model has one additional parameter), for which the p-value is 0.007. Note that this p-value is not exactly the same as the one ($p = 0.014$) reported for the coefficient in (18.3), which was computed using the normal distribution.

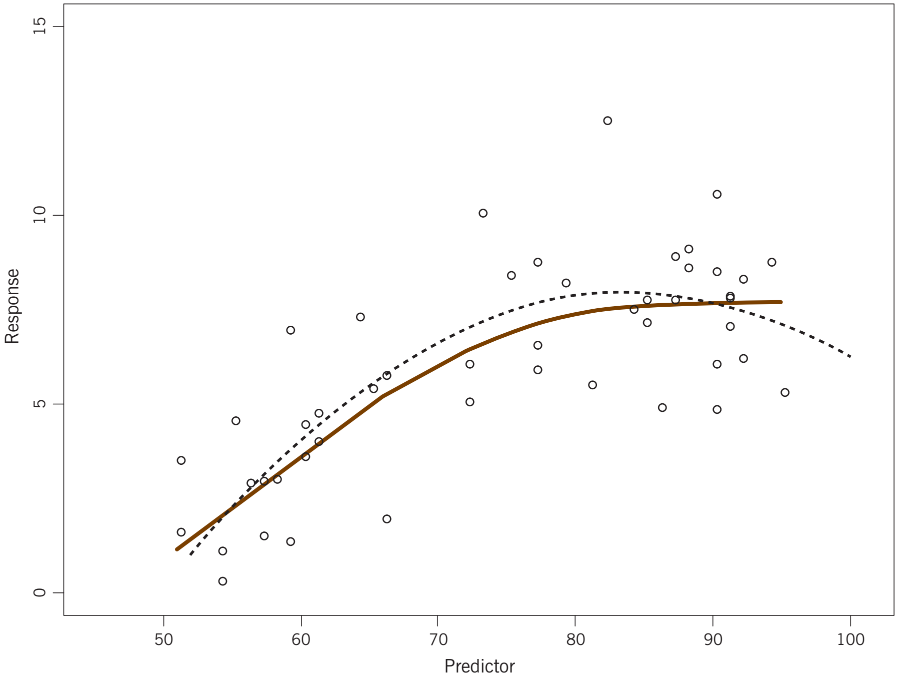
When comparing the performance of alternative models (i.e. testing alternative hypotheses) it is important to consider both their goodness of fit (e.g., analysis of deviance) and complexity, that is the number of explanatory variables (and associated parameters) included in each model. Akaike's information criterion (AIC) is a way of comparing models with different numbers of parameters that combines goodness of fit and complexity considerations. Since adding explanatory variables to a model, whether they are relevant or not, will always explain more variance, i.e. reduce the residual deviance, the AIC criterion adds a penalty for additional parameters equal to two times the number of parameters. Here the constant term is included, hence in our logistic regression model (18.3) the AIC is equal to the residual deviance plus 4 (2 times the number of parameters), i.e. 61.81 + 4 = 65.81. Models with the lowest AIC are to be preferred. For example, if we added another explanatory variable to the model and reduced the residual deviance to 60.50, say, then the AIC would be 60.50 + 2 × 3 = 66.50, and the first model with one parameter is preferable because it has lower AIC.

Nonlinear models

In the above examples, although the link function that transforms the mean of the response changes according to the response variable type, the way the predictors are combined is always a linear function, which is quite a simplistic assumption. Transformations can also be made of the predictors to accommodate nonlinear relationships. For example, Exhibit 18.5 shows another set of observations, and a scatterplot smoother has been added (in brown) indicating the possibility

[1] This is the way the `R` function `glm` reports the deviance values.

Exhibit 18.5: *A scatterplot of a continuous response y and a predictor x, showing a scatterplot smoother (brown line) which suggests a nonlinear relationship. The estimated quadratic relationship is shown by a dashed curve*



that the relationship is not linear, but reaching a peak or possibly an asymptote. In order to take into account the curvature, an additional predictor term in x^2 can be included so that a quadratic is postulated as the regression function (polynomial regression), and the result is as follows:

$$\begin{array}{lccccc} \text{mean of } y = & -39.91 & + & 1.139x & - & 0.0068x^2 \\ & [9.09] & & [0.258] & & [0.0018] \\ & (p < 0.0001) & & (p < 0.0001) & & (p = 0.0004) \end{array} \qquad (18.4)$$

All terms are significant and the confidence intervals for the coefficients all contain the true values used in the simulated formula, which is $-32 + 0.9x - 0.005x^2$. The estimated regression function in (18.4) is shown with a dashed line.

Since we have introduced fuzzy coding, it is interesting to compare the results using this alternative. Fuzzy coding of the predictor variable with three fuzzy categories allows for a curve with one turning point, which is what we need, so three categories were created, x_1, x_2, x_3, and the following regression function resulted:

$$\begin{array}{lcccccc} \text{mean of } y = & 1.03x_1 & + & 7.95x_2 & + & 7.56x_3 & \quad (18.5) \\ & [0.59] & & [0.62] & & [0.61] & \\ & (p = 0.08) & & (p < 0.0001) & & (p < 0.0001) & \end{array}$$

(notice that we show the result without the constant for the three fuzzy dummies). The variance explained is 65.6%, only slightly less than the 66.0% for the quadratic; both have three coefficients and thus the same degrees of freedom in the regression. Exhibit 18.6 shows that with the fuzzy coding the relationship is estimated as two linear functions, because of the triangular membership functions used to create the fuzzy categories. To capture a curved relationship we should use different membership functions, of which there are many possibilities, for example Gaussian (i.e., normal) membership functions. If interest is not only in diagnosing a smooth relationship (like the scatterplot smoother visualizes) but also in testing it statistically, then the section on generalized additive models later in this chapter provides a solution.

Multiple predictors and interactions

Most often there are many explanatory variables, therefore we need a strategy to decide which are significant predictors of the response, and whether they interact, so as to choose the best model (hypothesis) given the data. As an illustration

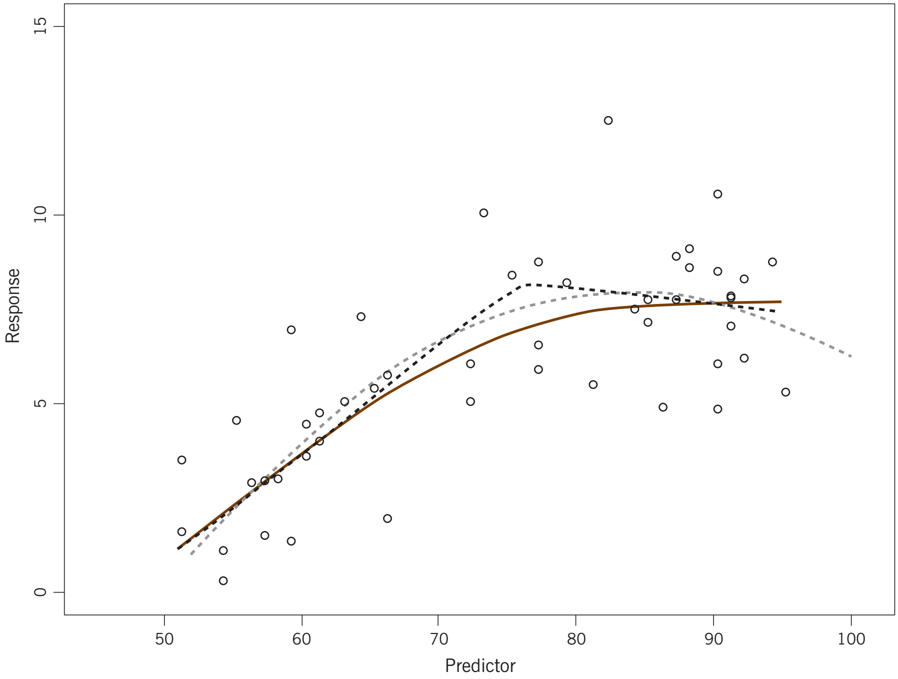


Exhibit 18.6:
Same data as in Exhibit 18.5, with the estimated quadratic relationship in gray, and the relationship according to (18.5) shown by black dashed lines

we return to the "Barents fish" data set studied in Chapters 1, 13 and 15. To convert the abundances of the 30 fish species, effectively a 30-variable response, to a single response, we computed the Shannon-Weaver diversity H' for each of the 89 stations, using the formula:

$$H' = -\sum_j p_j \log(p_j)$$

where p_j is the proportional abundance of species j, $j = 1,\ldots 30$. This well-known diversity index is at its lowest value, 0, when there is only a single species observed in the sample, and reaches its maximum if all species are contained in the same proportion. Amongst possible explanatory variables of diversity we have the bottom depth and temperature at the station trawled. If depth and temperature are entered into a linear regression, then the estimated effects of each variable do not depend on the other variable. In this example the effects of depth and temperature in a model as separate linear terms are not statistically significant. Significant results are found, however, when the interaction term is included, i.e. the product of the two variables, which allows for the possibility that the relationship with temperature depends on the depth and vice versa. The regression, explaining only 6.6% of the variance, but still significant, is:

$$\begin{array}{llllllll} \text{mean } H' = & 0.466 & + & 0.00243\text{ depth} & + & 0.493\text{ temp.} & - & 0.00152\text{ depth} \times \text{temp.} \qquad (18.6) \\ & [0.470] & & [0.00152] & & [0.218] & & [0.00072] \\ & (p=0.32) & & (p=0.11) & & (p=0.026) & & (p=0.038) \end{array}$$

The interaction term implies that the relationship with depth varies according to the temperature – notice that we would retain the linear term in depth even though it is insignificant, because the interaction term which involves depth is significant. Exhibit 18.7 shows the linear relationships with depth for three different temperatures that are chosen in the temperature range of the observed data.

Generalized additive models

Generalized additive models (GAM for short) are a very flexible framework for taking care of nonlinearities in the data. The approach is more complex but the benefits are great. Without entering too much into technicalities, we show the equivalent GAM analysis used to estimate the regression of diversity as a function of depth and temperature, in the previous example. If we enter depth and temperature as separate variables, the GAM results show that depth is significant with a clear nonlinear relationship ($p < 0.0001$) but not temperature ($p = 0.25$) – see Exhibit 18.8. In a GAM model the form of the relationship

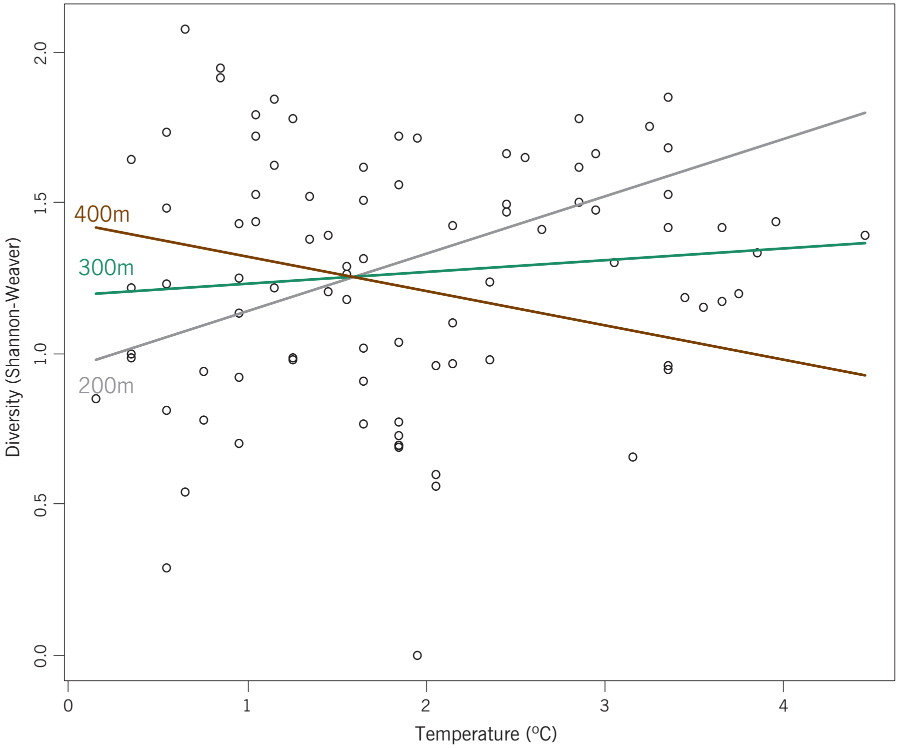


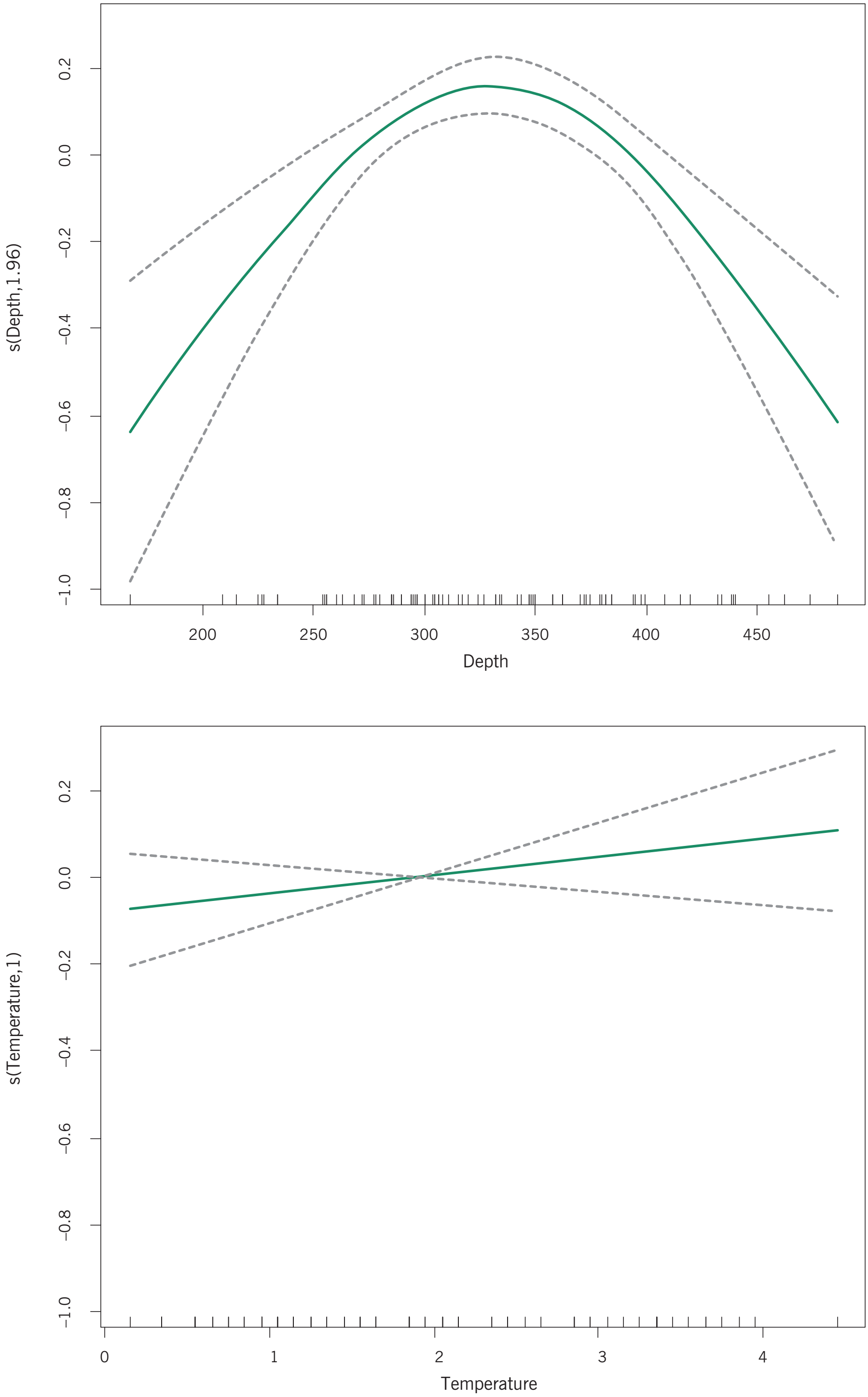
Exhibit 18.7:
Linear regression relationships (18.6) between diversity and temperature that depend on the depth (illustrated for three values of depth), showing an interaction effect (regression lines with different slopes). If the interaction effect were absent, the three lines would be parallel and the effect of temperature would be the same (i.e., not contingent on depth)

is diagnosed for the researcher, and the number of degrees of freedom of the relationship is estimated: 1.96 (say 2) for depth, and 1 for temperature. There is no mathematical model for the regression here, so we cannot write down a formula as before. But, since the depth relationship looks quadratic, we could try adding a quadratic term to the model, and return to using conventional regression:

$$\begin{array}{lcccc} \text{mean } H' = & -2.22 \;+\; 2.15 \text{ depth} & -\; 0.0000326 \text{ depth}^2 & +\; 0.0463 \text{ temp.} & (18.7) \\ & [0.80] \quad [0.0048] & [0.0000071] & [0.0374] & \\ & (p = 0.007)\,(p < 0.0001) & (p < 0.0001) & (p = 0.22) & \end{array}$$

To choose between models (18.6) and (18.7) we can compare the AIC in each case: 94.9 for (18.6) and 79.7 for (18.7). The difference in AIC between the parametric model in (18.7) and the GAM model summarized in Exhibit 18.8, which has an AIC of 79.5, is tiny. Model (18.7) could thus be further improved by dropping the insignificant temperature term:

Exhibit 18.8:
Generalized additive modelling of diversity as smooth functions of depth and temperature: depth is diagnosed as having a significant quadratic relationship, while the slightly increasing linear relationship with temperature is non-significant. Both plots are centred vertically at mean diversity, so show estimated deviations from the mean. Confidence regions for the estimated relationships are also shown



$$\begin{array}{lccccc} \text{mean } H' = & -2.11 & + & 2.13 \text{ depth} & - & 0.0000324 \text{ depth}^2 \\ & [0.80] & & [0.0048] & & [0.0000071] \\ & (p = 0.01) & & (p < 0.0001) & & (p < 0.0001) \end{array} \qquad (18.8)$$

in which case the AIC is 79.3, and explained variance is 19.9%.

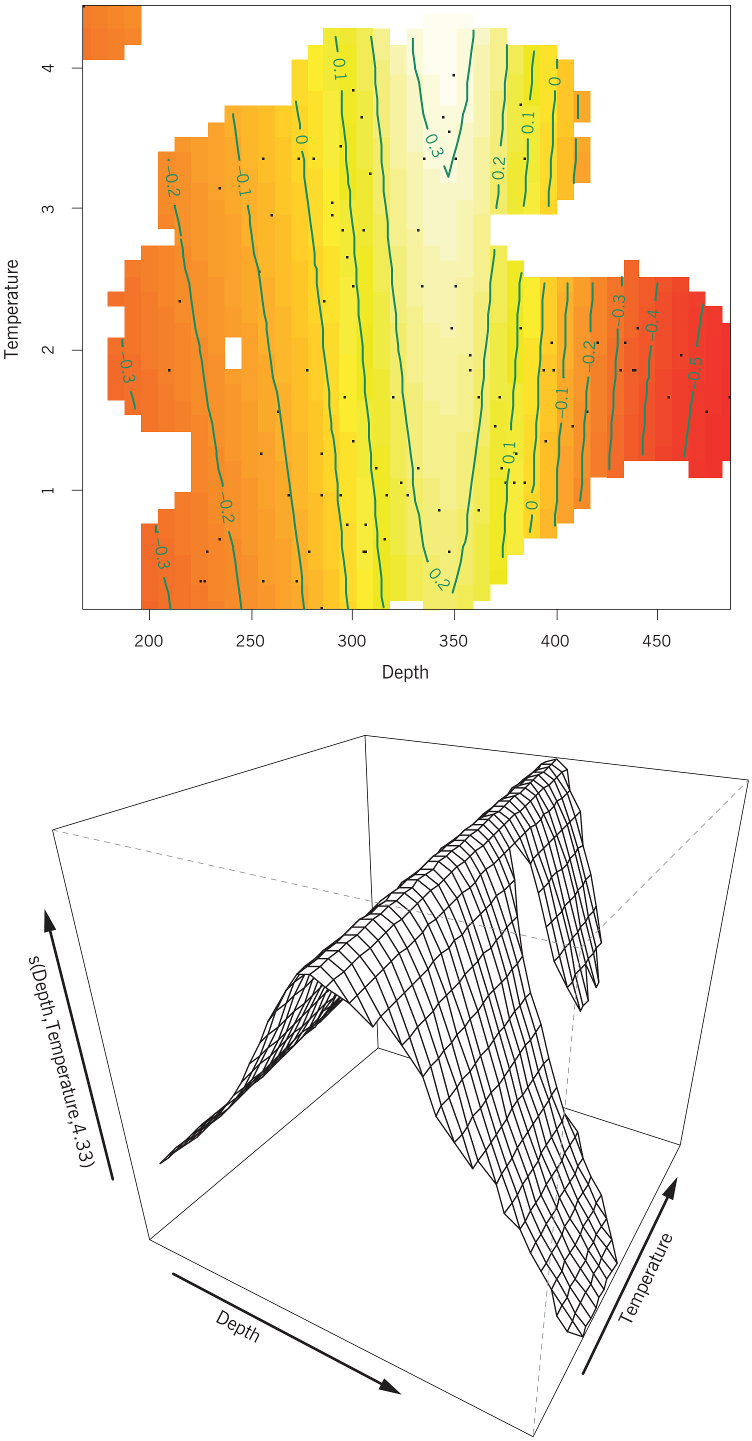
As a final illustration of the power of GAMs, we can make a model with a smooth interaction of depth and temperature. This has an even lower AIC value 70.7, and now to visualize the diagnosed relationship requires making a contour plot of the model values in the space of the depth and temperature variables, or making a perspective plot in three dimensions – see Exhibit 18.9.To test whether the interaction is significant we can compare the residual deviances for the model shown in Exhibit 18.8 (85.04) and the one in Exhibit 18.9 (83.67), i.e. a difference of only 1.37 units of deviance, which is not significant.[2] All these results and considerations lead us to the conclusion that the parametric model (18.8) with depth modelled as a quadratic is the one of choice – it has few parameters, is a function that can be easily interpreted and computed and does almost as well as several competing models that are more complex. Here we have demonstrated how GAM can help to suggest a nonlinear model for a regression. We will return to GAM modelling in Chapter 20 where we show that it is a convenient and flexible approach for taking into account the effect of spatial position.

Classification trees

We close this chapter on statistical modelling by showing a completely different approach to modelling a continuous or categorical response variable, by constructing a type of decision tree with the goal of predicting the continuous response variable (regression trees) or categorical response category (classification trees). We consider the latter case first, and take as example the presence/absence of polar cod *(Boreogadus saida)* in a sample. In the data matrix there are 21 samples with polar cod and 68 without, so the response data consist of 21 ones and 68 zeros. Applying a classification tree algorithm, with two predictors, depth and temperature, produces the tree model of Exhibit 18.10. The 89 samples are notionally fed down the tree and are split by the decisions at each branch, where each decision indicates the subsample that goes to the left hand side. For example, samples going to the left at the top of the tree satisfy the condition

[2] Here we have not entered into the aspect of the degrees of freedom for this comparison of GAM models, nor how p-values are computed. In GAM the degrees of freedom are not integers, but estimates on a continuous scale. Hence, comparing models leads to differences in degrees of freedom that are also not whole numbers – in this particular case the degrees of freedom associated with the deviance difference of 1.37 are 1.01, close enough to 1 for all practical purposes.

Exhibit 18.9:
Contour plot (upper) and perspective plot (down) of the diagnosed interaction regression surface of depth and temperature, predicting the deviations from mean diversity. The concave relationship with depth is clearly seen as well as the slight relationship with depth. The difference between the model with or without interactions is, however, not significant



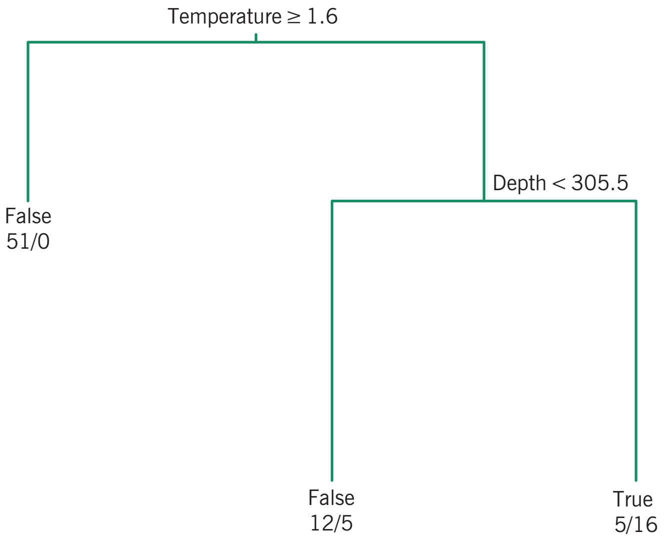


Exhibit 18.10: *Classification tree model for predicting the presence of polar cod. The one branch which predicts their presence gives the rule: temperature < 1.6 °C and depth ≥ 306 m. This rule correctly predicts the presence of polar cod in 16 samples but misclassifies 5 samples as having polar cod when they do not*

that temperature is greater than or equal to 1.6 °C, while the others for which temperature is less than 1.6 °C go to the right. Of the 89 samples, 51 go to the left, and all of them have no polar cod, so the prediction is False (i.e., no polar cod). The remaining 38 samples that go to the right are optimally split into two groups according to depth, 305 m or less to the left, and 306 m or more to the right. Of 38 samples, 17 go to the left and of these 12 have no polar cod, so False is predicted, while 21 go to the right, and a majority has polar cod so polar cod is predicted (True). The final branches of the tree, where the final predictions are made, are called *terminal nodes*, and the objective is to make them as concentrated as possible into one category.

The beauty of this approach is that it copes with interactions in a natural way by looking for combinations of characteristics that explain the response, in this case the combination of lower temperature (lower than 1.6 °C) and higher depths (greater than or equal to 306 m) is a prediction rule for polar cod, otherwise no polar cod are predicted.

As a comparison, let us perform a logistic regression predicting polar cod, using depth and temperature. Both variables are significant predictors but result in only 12 correct predictions of polar cod presence. The misclassification tables for the two approaches are given in Exhibit 18.11.

Regression trees

The same style of tree model can be constructed for a continuous response. In this case the idea is to arrive at terminal nodes with standard deviations (or

Exhibit 18.11: *Comparison of misclassification rates for the classification tree of Exhibit 18.10, compared to that for logistic regression, using the same predictors. The classification tree correctly predicts presence and absence in 79 of the 89 samples, while logistic regression correctly predicts 74*

		Classification tree		Logistic regression	
		Truth		*Truth*	
		Polar cod	No polar cod	Polar cod	No polar cod
Predicted	Polar cod	16	5	12	6
	No polar cod	5	63	9	62

any other appropriate measure of variability for the response) as low as possible. As an example, we return to the diversity response, this time choosing time latitude and longitude coordinates as the predictors in order to classify the samples into regions of homogeneous diversity. The result is given in Exhibit 18.12.

The regression tree partitions the sampling area and can be drawn on the map in Exhibit 18.13. The most diverse area is in the north-west, while the least diverse is in the central western block.

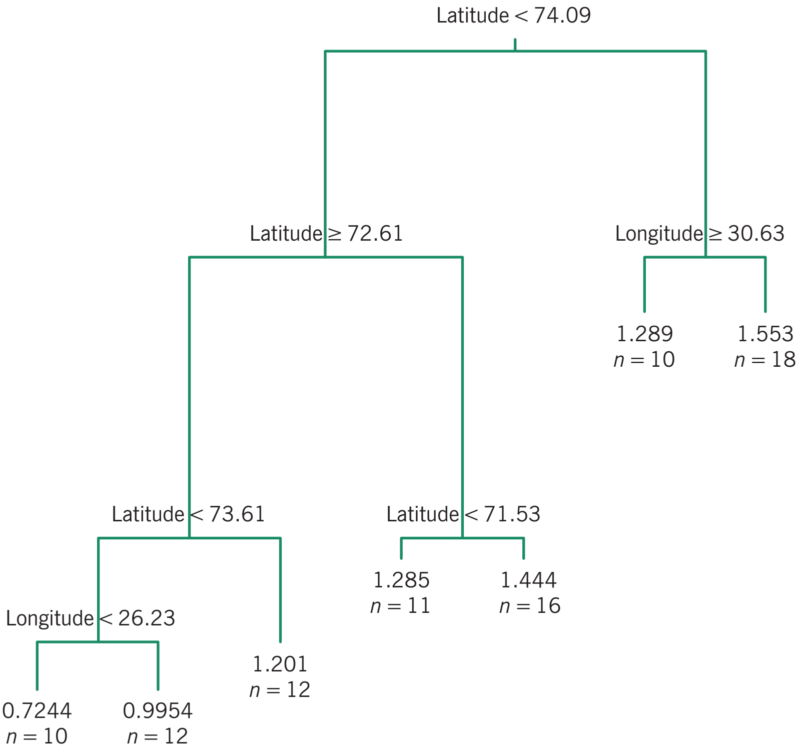


Exhibit 18.12: *Regression tree predicting fish diversity from latitude and longitude of sample positions. The terminal nodes give the average diversity of the samples that fall into them. This tree yields the spatial classification of the sampling region given in Exhibit 18.13*

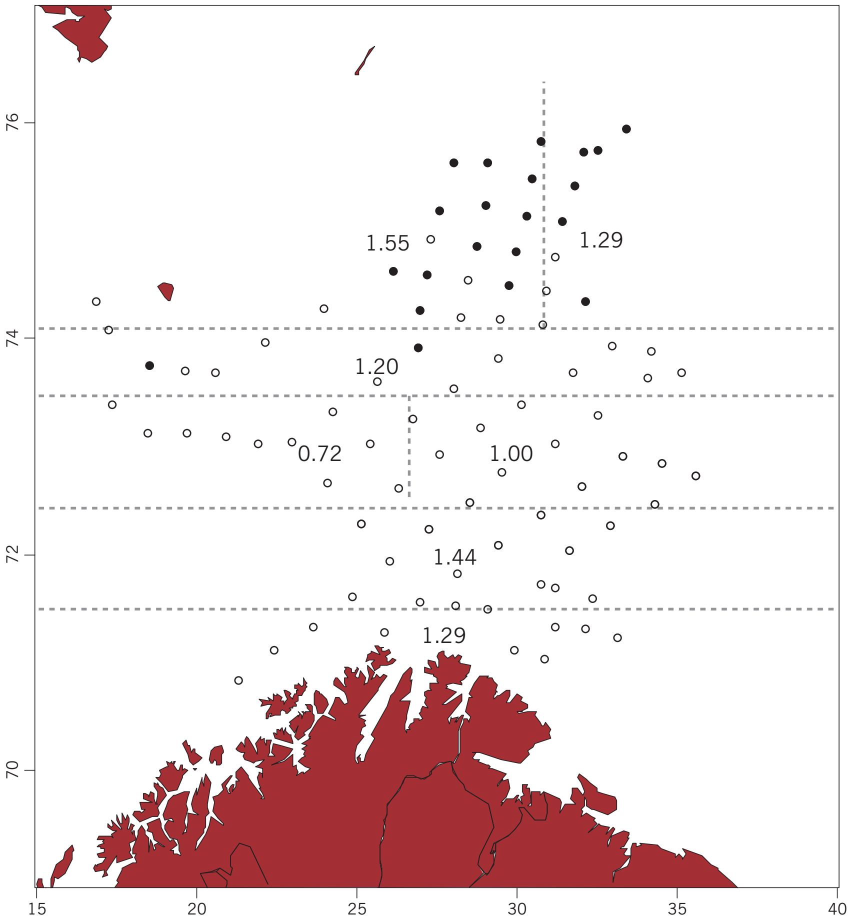


Exhibit 18.13:
Map of Barents Sea showing the locations of the 89 sampling sites (see Exhibit 11.1) and the slicing up of the region according to the regression tree of Exhibit 18.12, and the average fish diversities in each block. Most of the slices divide the latitudes north to south, with just two east-west divisions of the longitudes. Dark locations show the 21 sites where polar cod were found

SUMMARY:
Statistical modelling

1. The family of generalized linear models (GLMs) includes multiple linear regression, Poisson regression, and logistic regression, when the response variable is continuous, count or categorical, respectively, for which the assumed conditional distributions given a set of explanatory variables (or predictors), are normal, Poisson and binomial respectively.

2. Each of these models assumes that a transformation of the mean is a linear function of the explanatory variables. This transformation is called the *link function*. In multiple regression there is no transformation, and the link is thus the identity. In Poisson regression the link is the logarithm, and in logistic regression it is the logit function, or log-odds.

3. To take into account nonlinearities, polynomial functions of the explanatory variables or fuzzy coding into several categories can be used.

4. Generalized additive models (GAMs) are even more general than GLMs, allowing considerable flexibility in the form of the relationship of the response with the explanatory variables.

5. Both GLM and GAM environments allow interaction effects to be included and tested.

6. Classification and regression trees are an alternative that specifically look at the interaction structure of the predictors and come up with combinations of intervals that predict either categorical or continuous responses with minimum error.

List of Exhibits