

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 10 Offprint

Regression Biplots

First published: December 2013

ISBN: 978-84-92937-50-9

Supporting websites:

www.fbbva.es

www.multivariatestatistics.org

© the authors, 2013

© Fundación BBVA, 2013

Fundación **BBVA**

Regression Biplots

In the previous chapter, displays of samples were obtained in a scatterplot with spatial properties (hence often called a *map*), approximating given distance or dissimilarity matrices. Then some types of variables were added to the display, specifically zero/one categorical variables (e.g., presences of species, sediment categories) and count variables (e.g., species abundances). In this chapter we continue with this theme of adding variables to a plot of samples, including continuous variables in their original form or in fuzzy-coded form. When samples and variables are displayed jointly in such a scatterplot, it is often called a *biplot*. This designation implies that a certain property holds between the two sets of points in the display in terms of the scalar products between the samples and variables. In this chapter we consider the simplest form of biplot, the regression biplot, which will serve two purposes: first, to give a different geometric interpretation of multiple regression; and second, to give a basic understanding of all the joint displays of samples and variables that will appear in the rest of this book.

Contents

Algebra of multiple linear regression	127
Geometry of multiple linear regression	128
Regression biplot	132
Generalized linear model biplots with categorical variables	133
Fuzzy-coded species abundances	135
More than two predictors	135
SUMMARY: Regression biplots	138

The multiple linear regression model postulates that the expected value of a response variable Y (i.e., the mean of Y) is a linear combination of several explanatory variables x_1, x_2, \dots, x_p :

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p \tag{10.1}$$

Algebra of multiple linear regression

For example, using the data of Exhibit 1.1, consider the regression of species labelled d on depth, pollution and temperature. The model is estimated as:

$$E(d) = 6.271 + 0.148 \times \text{depth} - 1.388 \times \text{pollution} - 0.043 \times \text{temperature} \quad (10.2)$$

Notice that, for the moment, we do not comment on whether this type of linear model of a count variable on three environmental variables would be sensible or not, because d is not an interval variable – we will return to this point later.

Since the coefficients in (10.2) depend on the units of the variables, we prefer to consider the regression using all variables in comparable units. Usually this is done by standardization of the variables, so that they are all in units of standard deviation. Let us denote these standardized variables (i.e., centred and normalized) with an asterisk, then the regression model becomes:

$$E(d^*) = 0.347 \times \text{depth}^* - 0.446 \times \text{pollution}^* - 0.002 \times \text{temperature}^* \quad (10.3)$$

The constant term now vanishes and the coefficients, called *standardized regression coefficients*, can be compared with one another. Thus it seems that pollution has the strongest influence on the average level of species d , reducing it by 0.446 of a standard deviation for every increase of one standard deviation of pollution. The effect of temperature is minimal and, in fact, is nonsignificant statistically ($p = 0.99$), while depth and pollution are both significant ($p = 0.039$ and $p = 0.010$, respectively), so we drop temperature and consider just the regression on the other two variables, which maintains the value of the coefficients, but slightly smaller p -values: $p = 0.035$ and $p = 0.008$, respectively:

$$E(d^*) = 0.347 \times \text{depth}^* - 0.446 \times \text{pollution}^* \quad (10.4)$$

Geometry of multiple linear regression

When referring to the multiple regression model, it is often said that a *hyperplane* is being fitted to the data. For a single explanatory variable this reduces to a straight line in the familiar case of simple linear regression. When there are two explanatory variables, as in (10.4), the model is a two-dimensional plane in three dimensions, the third dimension being the response variable d^* – a view of this plane in three dimensions is given in Exhibit 10.1, with standardized depth* and pollution* forming the two horizontal dimensions and d^* the vertical one. Notice how the plane is going down in the direction of pollution, but going up in the direction of depth, according to the regression coefficients (see the web site of the book which shows a video of this three-dimensional image). Notice too the lack of fit of the points to the plane – the value of R^2 for the regression is 0.442, which means that 44.2% of the variance of d is being

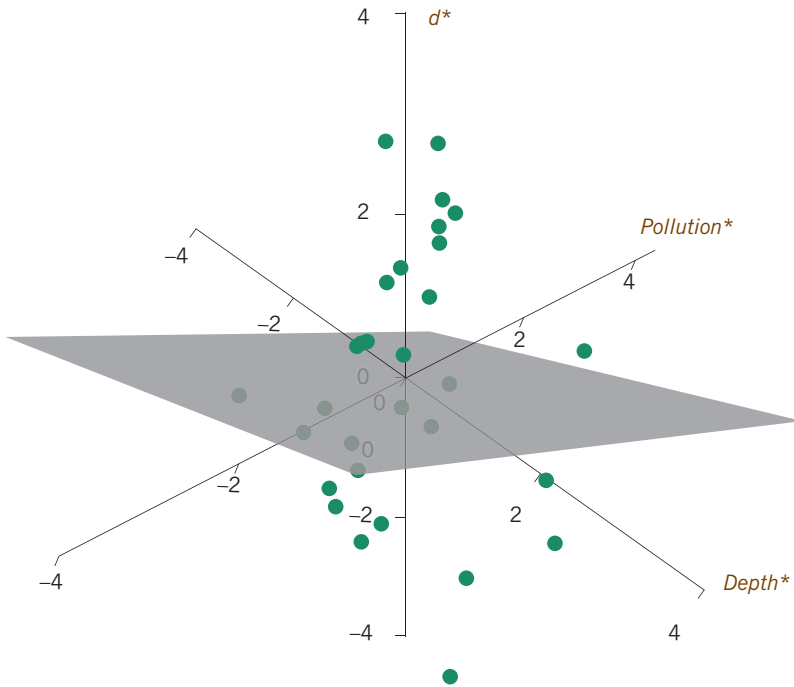


Exhibit 10.1:
 Regression plane defined by Equation (10.4) for standardized response d^* and standardized explanatory variables $pollution^*$ and $depth^*$. The view is from above the plane

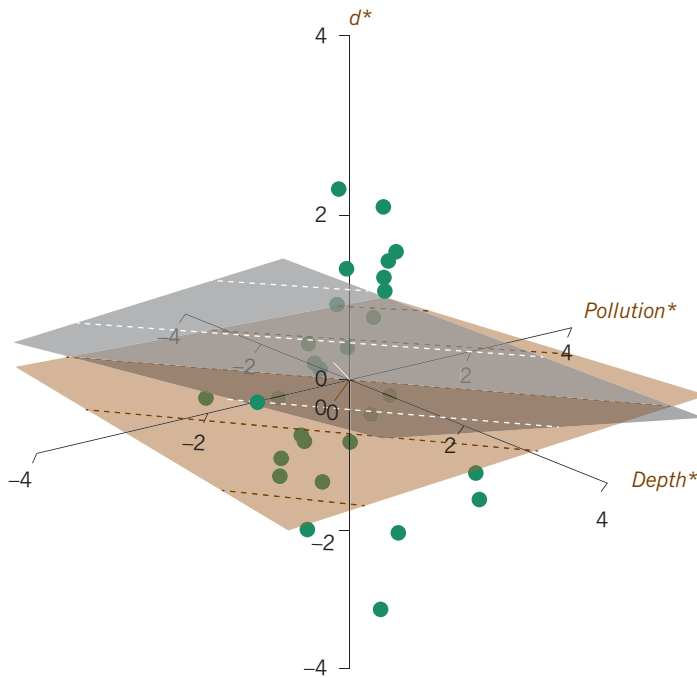
explained, and 55.8% of the variance unexplained and considered residual, or error, variance.

The linearity of the plane means that predictions of the same mean values form parallel straight lines in the plane. From a mountaineer’s point of view, if you are standing on the plane and want to stay at the same height, you need to walk in a straight line. Projecting these parallel straight lines onto the depth–pollution plane gives the *contours*, also called *isolines*, as shown in Exhibit 10.2. Finally, the vector in the depth–pollution plane with coordinates equal to the regression coefficients, $[0.347 \ -0.446]$, called the *gradient*, indicates the direction of steepest ascent in the regression plane, and is perpendicular to the contours. Given the geometry of the regression plane in Exhibit 10.2, it follows that we can do away with the d^* dimension, just like cartographers do, and consider just the depth–pollution plane and the contours of the regression plane, which are perpendicular to the gradient vector. Exhibit 10.3 shows this “ground view” of the model.

The short arrow labelled d is the gradient vector. The dashed line through this vector is called the *biplot axis* for the variable d . Contour lines are perpendicular to the biplot axis. Exhibit 10.3(a) corresponds to the darker “shadow” in Exhibit 10.2 in the depth–pollution plane, where the contours are in units of standard

Exhibit 10.2:

Another view of the regression plane, showing lines of equal height (dashed white lines in the plane) and their projection onto the depth–pollution plane (brown dashed lines in the darker “shadow” of the plane). The view is now from below the regression plane but above the depth–pollution plane. The short solid white line in the regression plane shows the direction of steepest ascent, and its projection down onto the depth–pollution plane is the gradient vector



deviation (sd) of species d (sd of $d = 6.7$). The mean of d , equal to 10.9, corresponds to the contour line through the origin. Calibrating the biplot axis in the original abundance units of d , Exhibit 10.3(b) is obtained.

Now the expected abundance, according to the regression model, can be estimated for any sample by seeing on what contour line it lies, which is achieved by projecting the point perpendicularly onto the biplot axis. For example, the sample shown in Exhibit 10.3(b), with standardized coordinates -0.668 and 1.720 , is on a contour line with value 4.2. The observed value for this sample is 3, so this means that the regression plane lies above the sample point and thus over-estimates its value. The action of projecting the sample point perpendicularly onto the biplot axis is a scalar product operation – just the regression model (10.4), in fact. The scalar product of the gradient vector $[0.347 \ -0.446]$ with the sample point vector $[-0.668 \ 1.720]$ is:

$$0.347 \times -0.668 + (-0.446) \times 1.720 = -0.999$$

which means that the prediction is almost exactly one standard deviation below the mean of d (in Exhibit 10.3(a) it is on the contour line $-1sd$), that is $10.9 - 0.999 \times 6.7 = 4.2$.

REGRESSION BIPLOTS

(a)

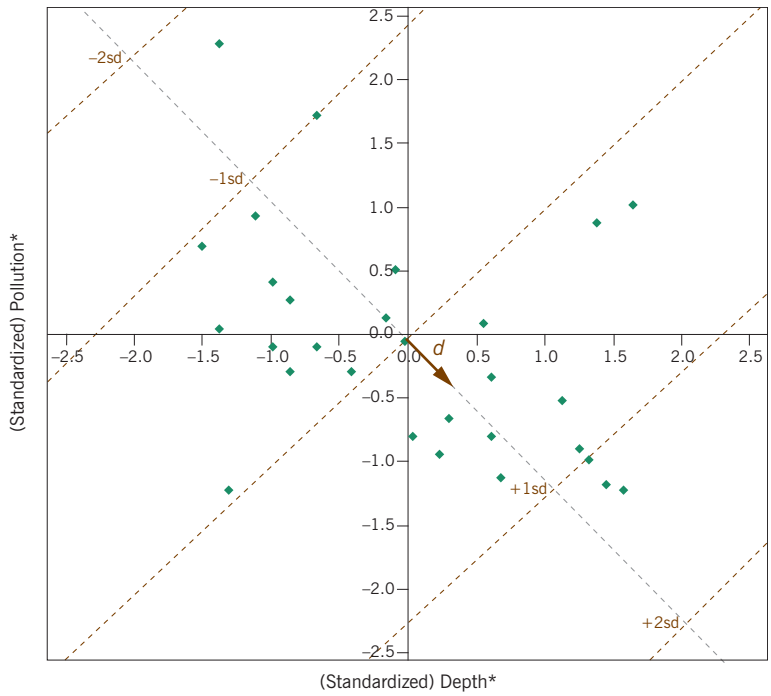
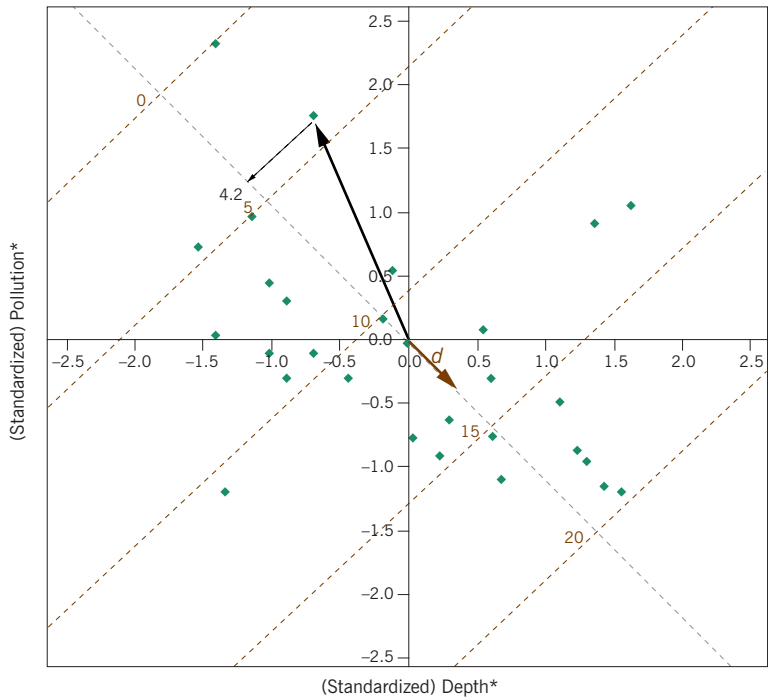


Exhibit 10.3: Regression plane shown as contour lines in the plane of the two explanatory variables, depth and pollution, both standardized. In (a) the contours are shown of the standardized response variable c^* , where the units are standard deviations (sd's) and the contour through the origin corresponds to mean 0 on the standardized scale, i.e. the mean on the original abundance scale. In (b) the contours are shown after unstandardizing to the original abundance scale of d . The sample shown in (b) corresponds to a height of 4.2 on the regression plane

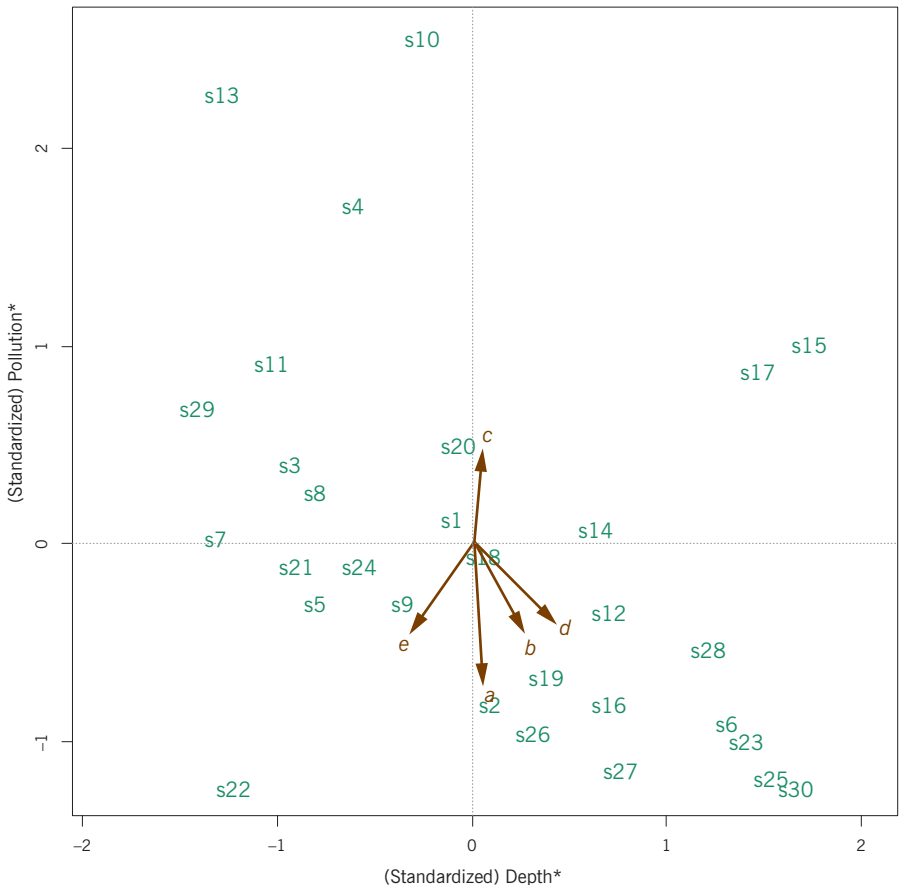
(b)



Regression biplot

We have given a different geometric view of multiple regression, for the case of two predictor variables, reducing the regression model to the gradient vector of regression coefficients in the plane of the predictors (we will come to the case of more predictor variables later). The contours of the plane are perpendicular to the gradient vector. We can now perform the regressions of the other four species with the same two predictors. Each species has a different pair of regression coefficients defining its gradient vector and all five of these are plotted together in Exhibit 10.4. The fact that *b* and *d* point in similar directions means that they have similar regression relationships with the two predictors, and the samples will have similar projections onto the two biplot axes through *b* and *d*. Species *a* and *c* point in opposite directions and thus have opposite relationships, *a* negative with pollution and *c* positive. While samples high in the vertical direction such as *s10* and *s13* have high (modelled) abundances of *c*, they also have the lowest abundances of *a*.

Exhibit 10.4:
Regression biplot of the five species with respect to the predictors depth and pollution



Each regression has an associated R^2 value: for the five species these are (as percentages) (a) 52.9%, (b) 39.8%, (c) 21.8%, (d) 44.2%, (e) 23.5%. An overall measure of variance explained for all five regressions in the biplot is the ratio of the sum of the explained variances in each and the sum of the total variances, which gives a value of 41.5%. As far as statistical significance is concerned, all species have significant linear relationship with pollution, but only *d* and *e* are significantly related to depth as well (these have the highest standardized regression coefficients on the horizontal axis in Exhibit 10.4).

In Chapter 9 we have already shown how the environmental variable sediment (Exhibit 1.1), which is categorical, can be added to a MDS display. Each category is placed at the average of the samples in which it is contained – we call these *supplementary points*. Similarly, we situated species in an MDS map as supplementary points by positioning them at their weighted averages of the sample points, with weights equal to the relative abundances. This approach can be used here as well, but their positions do not reflect any formal relationship between the species and the predictors. *Logistic regression* can be used in this case to give gradient vectors to represent these categories.

Generalized linear model
biplots with categorical
variables

Logistic regression models the logarithm of the odds of being in a given category, in this case a particular sediment category. Modelling the log-odds (i.e., the *logit*) for each sediment category as a function of (standardized) depth and pollution using logistic regression leads to three sets of regression coefficients in the linear part of the model. For example, for gravel (G), the model is:

$$\text{logit}(p_G) = \log\left(\frac{p_G}{1-p_G}\right) = -3.322 + 2.672 \times \text{depth}^* - 2.811 \times \text{pollution}^*$$

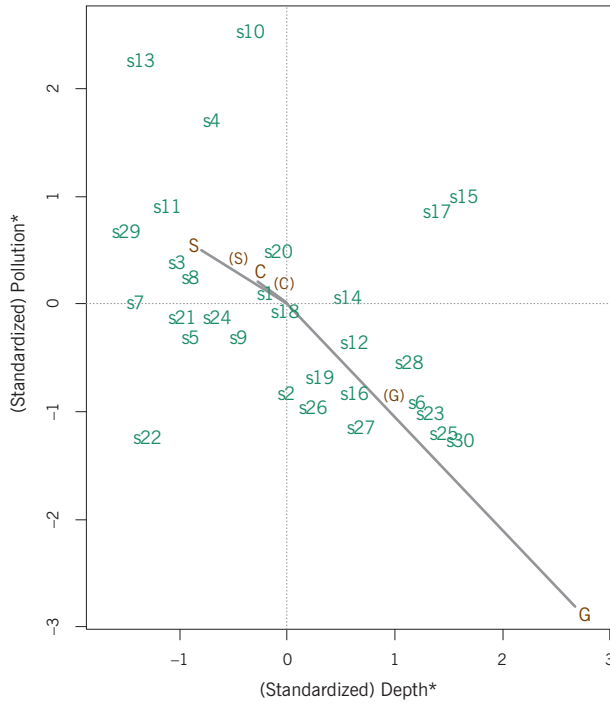
The three sediment categories are shown according to their logistic regression coefficients in Exhibit 10.5(a), connected to the origin. In fact, the above logistic regression is the only one that is statistically significant, those for clay (C) and sand (S) are not. The categories as supplementary points are also shown in Exhibit 10.5(a) by smaller labels in parentheses.

Rather than use linear regression to display the species in a regression biplot, as in Exhibit 10.4, there are two other alternatives: Poisson regression, which may be considered more appropriate because it applies to count response data, or fuzzy coding. The Poisson regressions lead to the coefficients displayed as vectors in Exhibit 10.5(b). Significance with respect to the two predictors is the same as for the linear regression (see above), with in addition species *b* being significantly related to depth. Both Poisson and logistic regression are treated in more detail in Chapter 18.

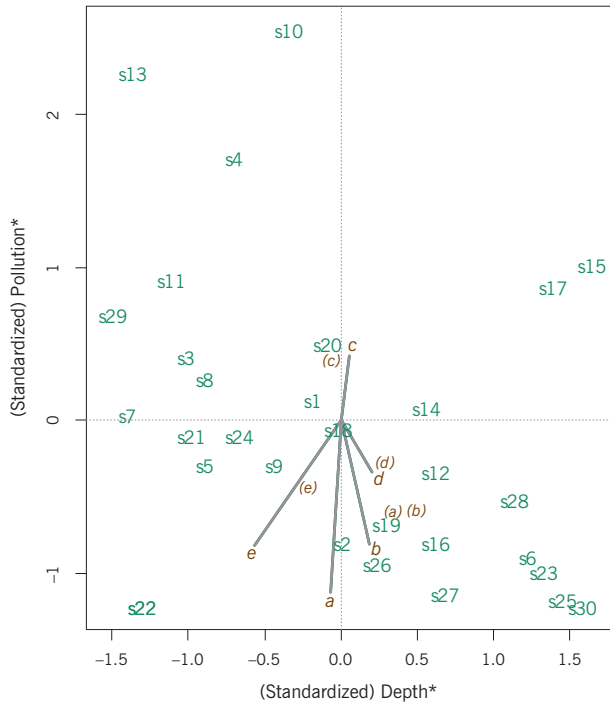
Exhibit 10.5:

(a) Logistic regression biplot of the three sediment categories and (b) Poisson regression biplot of the five species as predicted by depth and pollution. In each biplot the gradient vectors are shown connected to the origin. In addition, the positions of the sediment categories and the species as supplementary points are given in their respective biplots by their labels in parentheses

(a)



(b)



For the fuzzy coding, because there are several zeros in the abundance data, we can set up fuzzy codes for the species with a “crisp” code just for zeros and then three fuzzy categories for the nonzero values. Thus, for species *a*, for example, the code *a0* refers to zero abundance and *a1*, *a2* and *a3* refer to low, medium and high positive abundances. Exhibit 10.6 shows the two ways of representing the fuzzy categories, first in terms of their (linear) regressions on depth and pollution, and second, in terms of their supplementary point positions as weighted averages of the samples. Overall, 10.6(a) and (b) tell the same story: most of the variation is in a vertical direction, along the pollution direction, with high values of species *e* (i.e., category *e3*) ending up bottom left, while the corresponding categories for species *b* and *d* end up bottom right. The trajectories of each species contain features that are not possible to see in the previous biplots. For example, species *d* has an interesting nonlinear trajectory, with low positive values (*d1*) pulled out towards the shallowest depths. Since the sample size is small, this feature may not be statistically significant – we shall return to this aspect later, the point we are making here is that fuzzy coding can reveal more information in the relationships than linear models.

The two predictor variables depth and pollution form what is called the *support* of the biplot. With the aid of three-dimensional graphics we could have a third variable, in which case the gradient vectors would be three-dimensional. But if we only have a two-dimensional “palette” on which to explore the relationships, multivariate analysis can provide the solution, at the expense of losing some information. As in the case of MDS, however, we are assured that a minimum amount of information is lost.

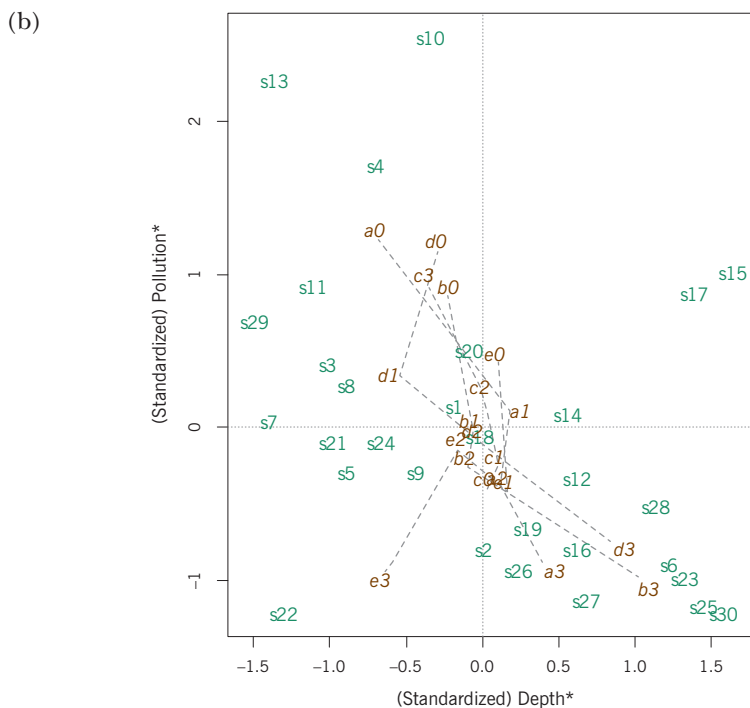
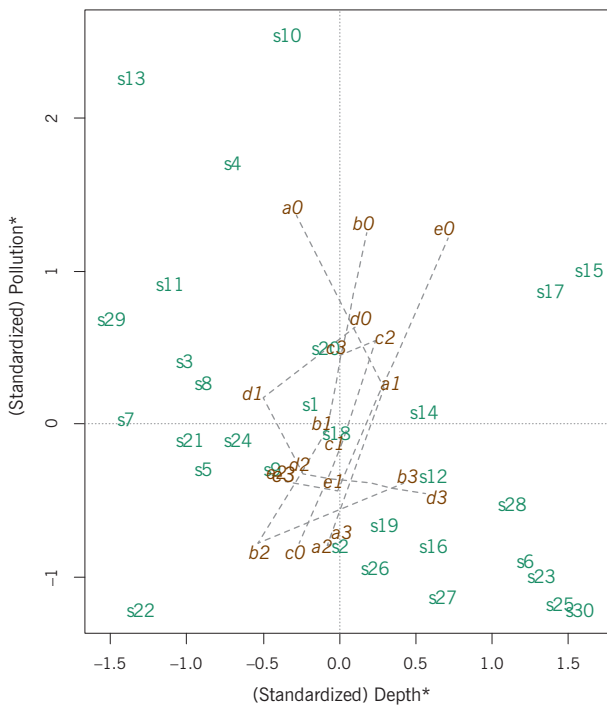
Without entering into all the details of a multivariate method called *canonical correlation analysis*, which is a form of linear regression analysis between two sets of variables, we simply show its results in Exhibit 10.7, which visualizes all the (linear) relationships between the five species and the three predictor variables depth, pollution and temperature. The configuration of the samples looks very similar to a 90 degree counter-clockwise rotation of the scatterplots in previous biplots. However, the support dimensions are no longer identified with single predictors but are rather linear combinations of predictors. The two canonical axes are defined as follows:

canonical dimension 1 = $0.203 \times \text{depth}^* - 0.906 \times \text{pollution}^* - 0.009 \times \text{temperature}^*$

canonical dimension 2 = $1.057 \times \text{depth}^* + 0.607 \times \text{pollution}^* - 0.102 \times \text{temperature}^*$

These dimensions are established to maximize the correlation between the species and the environmental variables. The first dimension is principally pollution

Exhibit 10.6: (a)
 Fuzzy coding of the species, showing for the fuzzy categories (a) their regressions on (standardized) depth and pollution, and (b) their weighted average positions with respect to the samples (i.e., supplementary points)



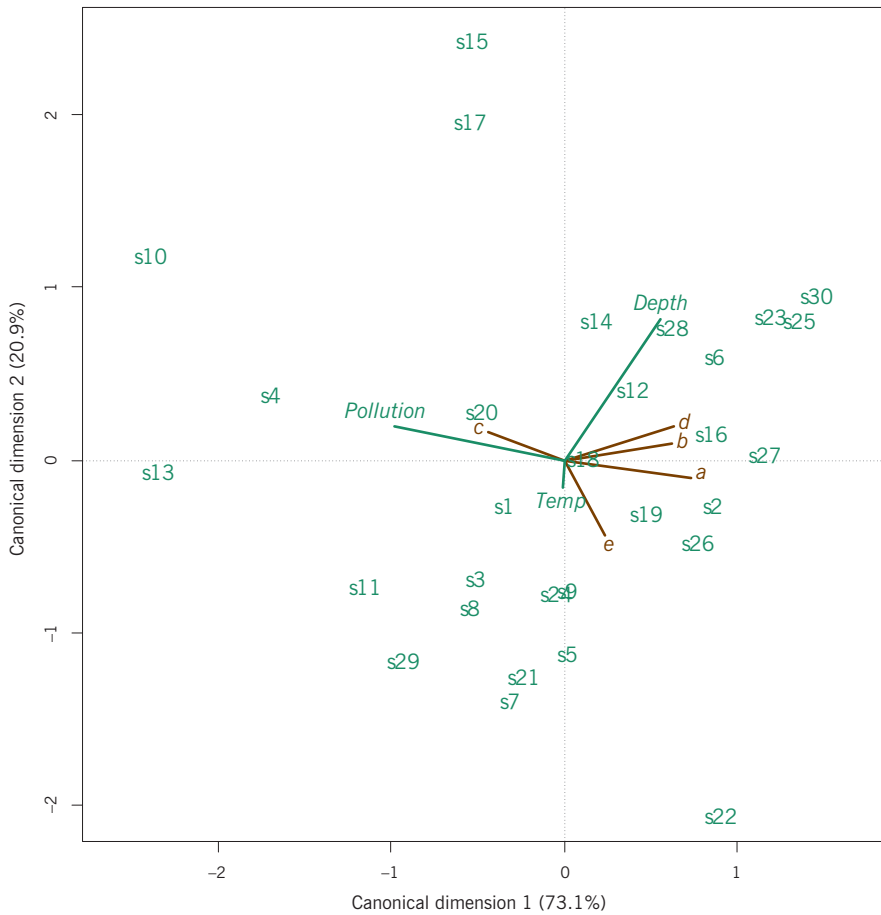


Exhibit 10.7:
Canonical correlation biplot of the five species with respect to the predictors depth, pollution and temperature

but to a lesser extent depth, while the second is mainly depth but to a lesser extent pollution. The first dimension is the most important, accounting for 73.1% of the correlation between the two sets, while the second accounts for 20.9%, that is 94.0% for the two-dimensional solution. Temperature plays a very minor role in the definition of these dimensions. The canonical dimensions are standardized, hence the support has all the properties of previous biplots except that the dimensions are combinations of the variables, chiefly depth and pollution. In addition, the canonical axes have zero correlation, unlike the depth–pollution support where the two variables had a correlation of -0.396 . Now that the support is defined, the species can be regressed on these two dimensions, as before, to show their regressions in the form of gradient vectors. The three environmental variables can be regressed on the dimensions as well, and their relationship shown using their gradient vectors, as in Exhibit 10.7. Notice that the angle between pol-

lution and depth suggests the negative correlation between them – see Chapter 6. In fact, the cosine of the angle between these two vectors is -0.391 , very close to the actual sample value of -0.396 . Notice as well the absence of relationship of temperature with the canonical dimensions.

Exhibit 10.7 is two biplots in one, often called a *triplot*. We shall return to the subject of triplots in later chapters – they are one of the most powerful tools that we have in multivariate analysis of ecological data because they combine the samples, responses (e.g., the species) and the predictors (e.g., the environmental variables) in a single graphical display, always optimizing some measure of variance explained.

SUMMARY:
Regression biplots

1. When there are two predictor variables and a single response variable in a multiple regression, the modelled regression plane can be visualized by its contours in the plane of the predictors (usually standardized). The contours, which are parallel straight lines, show the predicted values on the regression plane.
2. A regression biplot is built on a scatterplot of the samples in terms of the two predictors, called the *support* of the biplot. The gradient of the response variable is the vector of its regression coefficients, indicating the direction of steepest ascent on the regression plane. The gradient is perpendicular to the contour lines.
3. Several (continuous) response variables can be depicted by their gradient vectors in the support space, giving a *biplot axis* for each variable, and sample points can be projected perpendicularly onto the biplot axes to obtain predicted values according to the respective regression models.
4. When response variables are categorical, their gradient vectors can be obtained by performing a logistic regression on the predictors. Alternatively, the categories can be displayed at the averages of the sample points that contain them.
5. When response variables are counts (e.g., abundances), their gradient vectors can be obtained by Poisson regression. Again, there is the alternative of displaying them at the weighted averages of the sample points that contain them, where weights are the relative abundances of each variable across the samples.
6. For more than two predictor variables the support space of low dimensionality can be obtained by a dimension-reducing method such as canonical correlation analysis. Dimension reduction is the main topic of the rest of this book.

LIST OF EXHIBITS

Exhibit 10.1: Regression plane defined by Equation (10.4) for standardized response d^* and standardized explanatory variables pollution* and depth*. The view is from above the plane	129
Exhibit 10.2: Another view of the regression plane, showing lines of equal height (dashed white lines in the plane) and their projection onto the depth–pollution plane (red dashed lines in the darker “shadow” of the plane). The view is now from below the regression plane but above the depth–pollution plane. The short solid white line in the regression plane shows the direction of steepest ascent, and its projection down onto the depth–pollution plane is the gradient vector	130
Exhibit 10.3: Regression plane shown as contour lines in the plane of the two explanatory variables, depth and pollution, both standardized. In (a) the contours are shown of the standardized response variable d^* , where the units are standard deviations (sd’s) and the contour through the origin corresponds to mean 0 on the standardized scale, i.e. the mean on the original abundance scale. In (b) the contours are shown after unstandardizing to the original abundance scale of d . The sample shown in (b) corresponds to a height of 4.2 on the regression plane	131
Exhibit 10.4: Regression biplot of the five species with respect to the predictors depth and pollution	132
Exhibit 10.5: (a) Logistic regression biplot of the three sediment categories and (b) Poisson regression biplot of the five species as predicted by depth and pollution. In each biplot the gradient vectors are shown connected to the origin. In addition, the positions of the sediment categories and the species as supplementary points are given in their respective biplots by their labels in parentheses	134
Exhibit 10.6: Fuzzy coding of the species, showing for the fuzzy categories (a) their regressions on (standardized) depth and pollution, and (b) their weighted average positions with respect to the samples (i.e., supplementary points)	136
Exhibit 10.7: Canonical correlation biplot of the five species with respect to the predictors depth, pollution and temperature.....	137