Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 2 Offprint

Regression Biplots

First published: September 2010
ISBN: 978-84-923846-8-6

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

CHAPTER 2

Regression Biplots

Biplots rely on the decomposition of a target matrix into the product of two matrices. A common situation in statistics where we have such a decomposition is in regression analysis: $\hat{\mathbf{Y}} = \mathbf{XB}$, where $\mathbf{X}$ is a set of explanatory variables, $\mathbf{B}$ contains estimated regression coefficients and the values in $\hat{\mathbf{Y}}$ are the estimated values of one or more response variables. Thus $\hat{\mathbf{Y}}$ serves as the target matrix and $\mathbf{X}$ and $\mathbf{B}$ serve as the left and right matrices (actually $\mathbf{B}^{\mathsf{T}}$, since the right matrix of the decomposition is written in transposed form—see (1.4)). The coefficients in $\mathbf{B}$ are estimated to minimize the sum-of-squared errors between the original response variables in $\mathbf{Y}$ and the estimated values in $\hat{\mathbf{Y}}$. This context provides an excellent introduction to biplots as an approximation to higher-dimensional data.

Contents

Data set "bioenv"

Throughout this book we shall be using a small data set which serves as an excellent example of several biplot methods. The context is in marine biology and the data consist of two sets of variables observed at the same locations on the sea-bed: the first is a set of biological variables, the counts of five groups of species, and the second is a set of four environmental variables. The data set, called "bioenv", is shown in Exhibit 2.1. The species groups are abbreviated as "a" to "e". The environmental variables are "pollution", a composite index of pollution combining measurements of heavy metal concentrations and hydrocarbons; "depth", the depth in metres of the sea-bed where the sample was taken; "temperature", the temperature of the water at the sampling point; and "sediment", a classification

Exhibit 2.1: *Typical set of multivariate biological and environmental data: the species data are counts, while the environmental data are continuous measurements, each variable on a different scale; the last variable is a categorical variable classifying the substrate as mainly* C *(=clay/silt),* S *(=sand) or* G *(=gravel/stone)*

Site No.	Species counts					Environmental variables			
	a	*b*	*c*	*d*	*e*	*Pollution*	*Depth*	*Temperature*	*Sediment*
s1	0	2	9	14	2	4.8	72	3.5	S
s2	26	4	13	11	0	2.8	75	2.5	C
s3	0	10	9	8	0	5.4	59	2.7	C
s4	0	0	15	3	0	8.2	64	2.9	S
s5	13	5	3	10	7	3.9	61	3.1	C
s6	31	21	13	16	5	2.6	94	3.5	G
s7	9	6	0	11	2	4.6	53	2.9	S
s8	2	0	0	0	1	5.1	61	3.3	C
s9	17	7	10	14	6	3.9	68	3.4	C
s10	0	5	26	9	0	10.0	69	3.0	S
s11	0	8	8	6	7	6.5	57	3.3	C
s12	14	11	13	15	0	3.8	84	3.1	S
s13	0	0	19	0	6	9.4	53	3.0	S
s14	13	0	0	9	0	4.7	83	2.5	C
s15	4	0	10	12	0	6.7	100	2.8	C
s16	42	20	0	3	6	2.8	84	3.0	G
s17	4	0	0	0	0	6.4	96	3.1	C
s18	21	15	33	20	0	4.4	74	2.8	G
s19	2	5	12	16	3	3.1	79	3.6	S
s20	0	10	14	9	0	5.6	73	3.0	S
s21	8	0	0	4	6	4.3	59	3.4	C
s22	35	10	0	9	17	1.9	54	2.8	S
s23	6	7	1	17	10	2.4	95	2.9	G
s24	18	12	20	7	0	4.3	64	3.0	C
s25	32	26	0	23	0	2.0	97	3.0	G
s26	32	21	0	10	2	2.5	78	3.4	S
s27	24	17	0	25	6	2.1	85	3.0	G
s28	16	3	12	20	2	3.4	92	3.3	G
s29	11	0	7	8	0	6.0	51	3.0	S
s30	24	37	5	18	1	1.9	99	2.9	G

of the substrate of the sample into one of three sediment categories. Initially, we are going to consider the biological variables and only two of the environmental variables, "pollution" and "depth".

Simple linear regression on two explanatory variables

To start off let us consider species "d" as a response variable, denoted by d, being modelled as a linear function of two explanatory variables, "pollution" and "depth", denoting these two variables by y and x respectively. Simple linear regression leads to the following estimates of the response variable:

$$\hat{d} = 6.135 - 1.388\,y + 0.148\,x \qquad R^2 = 0.442 \qquad (2.1)$$

From a statistical inference point of view, both y and x are significant at the 5% level—their p-values based on the classical t-tests are 0.008 and 0.035 respectively.[1] The regression coefficients on the explanatory variables have the following interpretation: for every unit increase of pollution (variable y), abundance of species d decreases by 1.388 on average; while for every unit increase of depth (variable x), abundance of d increases by 0.148 on average. The amount of variance in d explained by the two variables is 44.2%, which means that the sum of squared errors across the 30 observations, $\sum_i (d_i - \hat{d}_i)^2$, which is minimized by the linear regression, is 55.8% of the total variance of d.

Standardized regression coefficients

The estimated regression coefficients in (2.1), i.e. the "slope" coefficients –1.388 and 0.148, have scales that depend on the scale of d and the scale of the two explanatory variables y and x, and so are difficult to compare with each other. To remove the effect of scale, all variables should be expressed in a comparable scale-free way. The most common way of doing this is to standardize all variables by centring them with respect to their respective means and dividing them by their respective standard deviations. We denote the standardized values of the three variables (their "z-scores") as d^*, y^* and x^* respectively, each having mean zero and variance 1 thanks to the standardization. The estimated regression relationship (2.1), including what are called the *standardized regression coefficients*, then becomes:[2]

$$\hat{d}^* = -0.446\,y^* + 0.347\,x^* \qquad R^2 = 0.442 \qquad (2.2)$$

Notice that there is no intercept, since all variables have mean zero. The regression coefficients now quantify the change in the standardized value of the response variable estimated from an increase of one standardized unit (i.e., one standard deviation) of each explanatory variable. The two coefficients can be compared and it seems that pollution has a bigger (negative) effect on species d than the (positive) effect of depth. Exhibit 2.2 shows schematically the difference between the regression plane for the unstandardized and standardized variables respectively.

Gradient of the regression plane

The standardized regression coefficients in (2.2) have an interesting geometric interpretation. They are the partial derivatives of $\hat{d}^*$ with respect to the two variables y^* and x^*, which are written mathematically as:

1. All calculations can be followed and repeated using the R code given in the Computational Appendix.
2. Since some regression programs do not automatically give the standardized regression coefficients, these can be easily calculated from the original ones as follows: standardized coefficient = original coefficient × (standard deviation of explanatory variable / standard deviation of response variable). See the Computational Appendix for examples of this calculation.

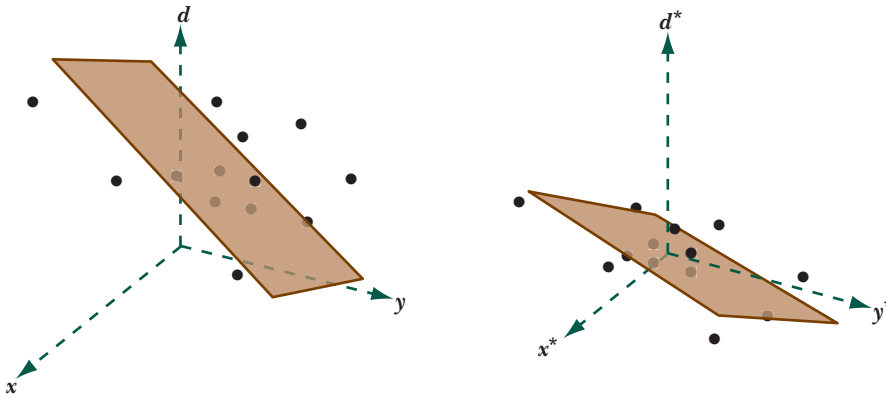


Exhibit 2.2: *Schematic geometric representation in three-dimensions of the regression plane for the original data* (on left) *and standardized data* (on right)*, where the response variable is the vertical dimension, and the two explanatory variables are on the "floor", as it were (imagine that we are looking down towards the corner of a room). The plane on the right goes through the origin of the three axes*

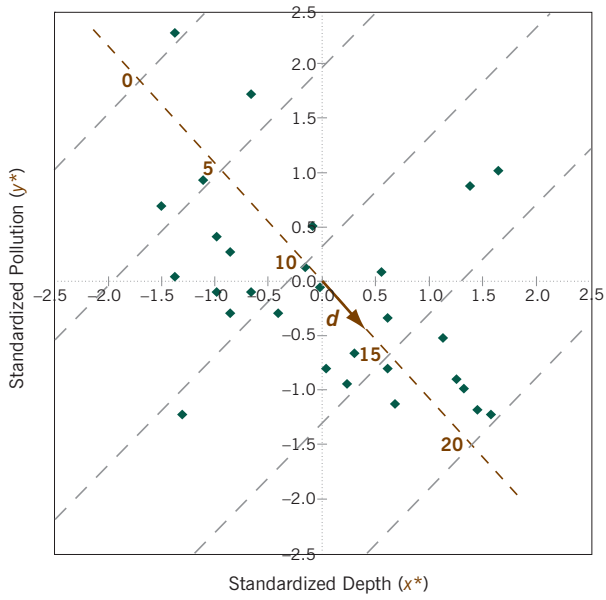
$$\frac{\partial \hat{d}^*}{\partial y^*} = -0.446 \qquad \frac{\partial \hat{d}^*}{\partial x^*} = 0.347$$

As a vector $[-0.446 \;\; 0.347]^T$ these two numbers indicate the *gradient* of the plane (2.2), that is the direction of steepest ascent up the plane in the right-hand image of Exhibit 2.2.

The complete geometry of the regression can then be drawn in the two-dimensional space of the explanatory variables, as shown in Exhibit 2.3. This is the scatterplot of the two variables y^* and x^*, whose values are given in the table alongside the figure. The brown arrow is the gradient vector, with coordinates -0.446 on the y^*-axis and 0.347 on the x^*-axis. So here we are looking at the plane on the right of Exhibit 2.2 from the top, down onto the y^*–x^* plane. The arrow indicates the gradient vector, pointing in the direction of steepest ascent of the plane, which is all we need to know to understand the orientation of the plane.

Contours of the regression plane

Now the contours of the regression plane, i.e., the lines of constant "height", by which we mean constant values of $\hat{d}^*$, form lines perpendicular to the gradient vector, just like in Exhibit 2.3. From the steepness of the plane (which we know from the gradient vector) it seems intuitively obvious that we can work out the heights of these contours on the standardized scale of d and then transform these back to d's original scale—in Exhibit 2.3 we show the contours for 0, 5, 10, 15 and 20. That is, we can calibrate the biplot axis for species d (we will explain exactly how to calibrate this axis below). Hence, to obtain the estimates of d for any given point y^* and x^* we simply need to see which contour line it is on, that is project it perpendicularly onto biplot axis d.



y*	x*
0.132	−0.156
−0.802	0.036
0.413	−0.988
1.720	−0.668
−0.288	−0.860
−0.895	1.253
0.039	−1.373
0.272	−0.860
−0.288	−0.412
2.561	−0.348
0.926	−1.116
−0.335	0.613
2.281	−1.373
0.086	0.549
1.020	1.637
−0.802	0.613
0.880	1.381
−0.054	−0.028
−0.662	0.292
0.506	−0.092
−0.101	−0.988
−1.222	−1.309
−0.989	1.317
−0.101	−0.668
−1.175	1.445
−0.942	0.228
−1.129	0.677
−0.522	1.125
0.693	−1.501
−1.222	1.573

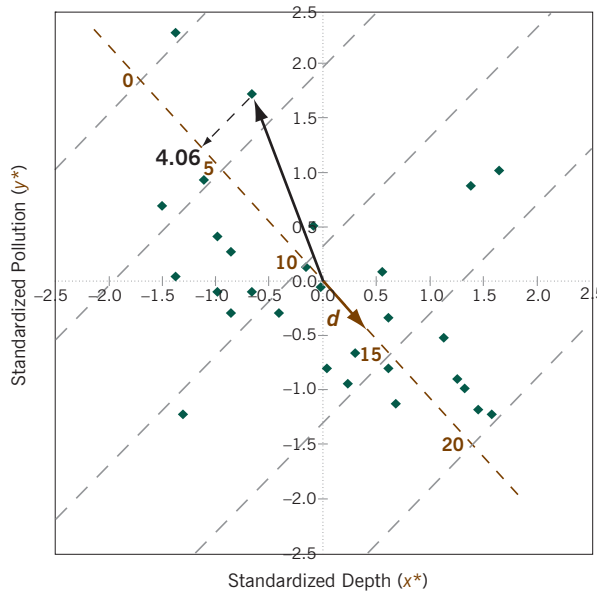
Exhibit 2.3:
The regression plane for species ***d*** *is shown by its gradient vector in the* x^*–y^* *space of the explanatory variables. Contour lines (or isolines) are drawn at selected heights*

Seeing which contour line corresponds to any given sample point is thus equivalent to projecting the point perpendicularly onto the biplot axis and reading off the value. Exhibit 2.4 shows an example of estimating the value of d for the fourth sample, giving a value of 4.06. This is not equal to the observed value of 3 (see original data on the left of Exhibit 2.4), but we do not expect it to be, since the variance explained by the regression of d on y and x is 44.2%. If we projected all the sample points onto the d-axis, and recovered the original values exactly, this would mean the regression plane passes exactly through the data points and the explained variance would be 100%. We are not in this situation, unfortunately, since our estimated values explain only 44.2% of the variance of d, in other words 55.8% of the variance is due to differences between the estimates and the observed values.

All this sounds exactly like what we described in Chapter 1, particularly concerning Exhibit 1.3, and indeed it is, because the regression equation (2.2) is nothing else but the scalar product between the gradient vector (indicating the biplot axis) and a general point in the x^*–y^* plane. Hence, this equation for estimating the response, given values of the two predictors as a biplot point, can be converted into the scalar product between this biplot point and the biplot gradient vector (the regression coefficients). This is equivalent to projecting the biplot point onto the biplot axis defined by the gradient vector, which is calibrated in units of the response variable.

Exhibit 2.4:
Projection of sample 4 onto the biplot axis, showing sample 4's original values in the table on the left and standardized values of the predictors on the right. The predicted value is 4.06, compared to the observed value of 3, hence an error of 1.06. The sum of squared errors for the 30 samples accounts for 55.8% of the variance of ***d****, while the explained variance* (R^2) *is 44.2%*

d	*y*	*x*
14	4.8	72
11	2.8	75
8	5.4	59
3	8.2	64
10	3.9	61
16	2.6	94
11	4.6	53
0	5.1	61
14	3.9	68
9	10.0	69
6	6.5	57
15	3.8	84
0	9.4	53
9	4.7	83
12	6.7	100
3	2.8	84
0	6.4	96
20	4.4	74
16	3.1	79
9	5.6	73
4	4.3	59
9	1.9	54
17	2.4	95
7	4.3	64
23	2.0	97
10	2.5	78
25	2.1	85
20	3.4	92
8	6.0	51
18	1.9	99



*y**	*x**
0.132	−0.156
−0.802	0.036
0.413	−0.988
1.720	−0.668
−0.288	−0.860
−0.895	1.253
0.039	−1.373
0.272	−0.860
−0.288	−0.412
2.561	−0.348
0.926	−1.116
−0.335	0.613
2.281	−1.373
0.086	0.549
1.020	1.637
−0.802	0.613
0.880	1.381
−0.054	−0.028
−0.662	0.292
0.506	−0.092
−0.101	−0.988
−1.222	−1.309
−0.989	1.317
−0.101	−0.668
−1.175	1.445
−0.942	0.228
−1.129	0.677
−0.522	1.125
0.693	−1.501
−1.222	1.573

Calibrating a regression biplot axis

In Chapter 1 we showed how to calibrate a biplot axis—one unit is inversely proportional to the length of the corresponding biplot vector (see equation (1.7)). In regression biplots the situation is the same, except the unit is a standardized unit and we prefer to calibrate according to the original scale of the variable. To express the regression coefficients on the original scale of d in the present case, we would simply multiply them by the standard deviation of d, which is 6.67, making them 6.67×-0.446 and 6.67×0.347 respectively. Then the calculation is as before, using the rescaled regression coefficients:

$$\begin{aligned}\text{one unit of } d &= 1 / \text{ length of biplot vector}\\ &= 1 / \sqrt{(6.67\times -0.446)^2 + (6.67\times 0.347)^2}\\ &= 1 / \left(6.67\times \sqrt{(-0.446)^2 + (0.347)^2}\right)\\ &= 1 / \left(6.67\times 0.565\right)\\ &= 0.265\end{aligned}$$

In general, the calculation is:

$$\text{one unit of variable} = 1 \Big/ \left(\begin{array}{c}\text{standard deviation}\\ \text{of variable}\end{array} \times \begin{array}{c}\text{length of biplot}\\ \text{vector}\end{array}\right) \qquad (2.3)$$

that is, the unit length of the standardized variable divided by the (unstandardized) variable's standard deviation. As far as the centre of the biplot is concerned, the variable's average is at the origin—in the present example the origin should be at the value 10.9, the average of *d*. We know that values are increasing in the direction of the biplot vector (towards bottom right in Exhibits 2.3 and 2.4), and also have computed the length of one unit on the biplot axis, so we have all we need to calibrate the axis. In the exhibits we calibrated at every 5 units, so the distance interval along the axis between consecutive values is $5 \times 0.265 = 1.325$.

Regression biplots with several responses

Each of the five species in Exhibit 2.1 can be linearly regressed on the two predictors "pollution" (*y*) and "depth" (*x*), using standardized scales for all variables, to obtain standardized regression coefficients that can be used as biplot vectors. Exhibit 2.5 shows the five regression analyses in one biplot. Each biplot vector points in the direction of steepest ascent of the regression plane. The larger the regression coefficients, the longer are the arrows and thus the steeper is the regression plane. If two biplot vectors are pointing in the same direction (for example, *b* and *d*) their relationships with the explanatory variables are similar. Species *c* clearly has an opposite relationship to the others, in that its regression

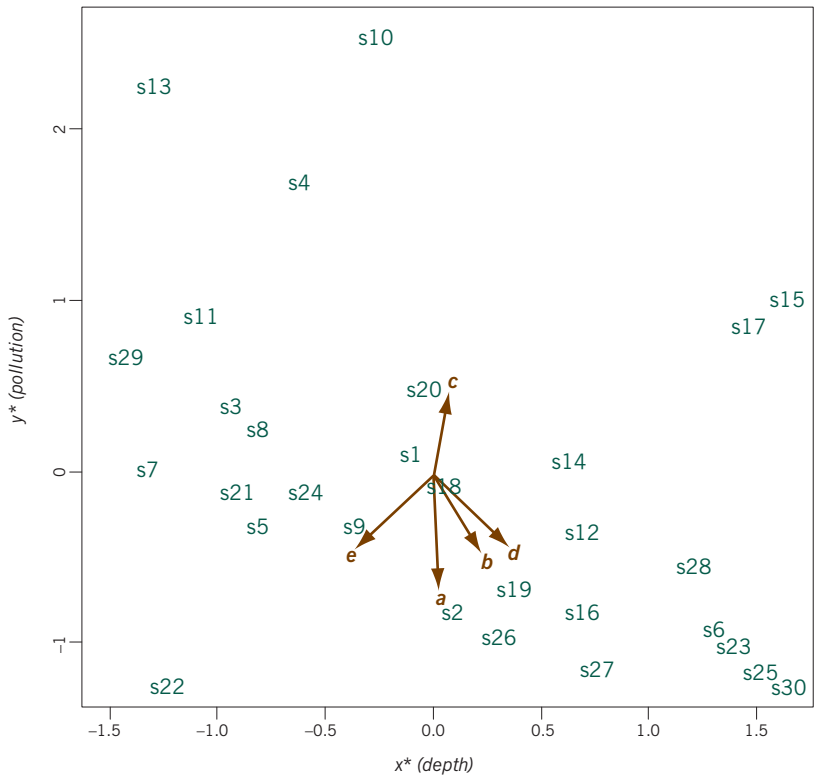


Exhibit 2.5:
Regression biplot of five response variables, species ***a*** *to* ***e****, in the space of the two standardized explanatory variables. The overall explained variance for the five regressions is 41.5%, which is the measure of fit of the biplot*

coefficient with pollution is positive while all the others are negative. The biplot axes through the biplot vectors can each be calibrated in the same way as explained above for response variable *d*, and the projections of the 30 samples (numbered in Exhibit 2.5) onto a particular biplot axis give the estimated values for that response. How close these estimated values are to the observed ones is measured by the R^2 for each regression: the percentages of explained variance, respectively for *a* to *e*, are 52.9%, 39.1%, 21.8%, 44.2% and 23.5%, with an overall R^2 of 41.5%. This overall value measures the quality of the regression biplot to explain all five response variables.

Finally, to show how regression biplots fit the general definition of the biplot given in Chapter 1, we write the estimation equation (2.4) for all five response variables in a matrix decomposition formulation as follows:

$$\left[\begin{array}{ccccc}\hat{\mathbf{a}}^* & \hat{\mathbf{b}}^* & \hat{\mathbf{c}}^* & \hat{\mathbf{d}}^* & \hat{\mathbf{e}}^*\end{array}\right] = \left[\begin{array}{cc}\mathbf{y}^* & \mathbf{x}^*\end{array}\right] \left[\begin{array}{ccccc}-0.717 & -0.499 & 0.491 & -0.446 & -0.475 \\ 0.025 & 0.229 & 0.074 & 0.347 & -0.400\end{array}\right] \quad (2.4)$$

that is, $$\hat{\mathbf{S}} = \mathbf{U}\mathbf{B}^{\mathsf{T}}$$

where the target matrix is the 30 × 5 matrix of estimated response values (standardized), the left matrix of the decomposition is the 30 × 2 matrix of standardized explanatory variables and the right matrix contains the standardized regression coefficients. The target matrix is an estimation $\hat{\mathbf{S}}$ of the observed (standardized) responses $\mathbf{S} = [\mathbf{a}^* \ \mathbf{b}^* \ \mathbf{c}^* \ \mathbf{d}^* \ \mathbf{e}^*]$, which can be written as: $\mathbf{S} \approx \hat{\mathbf{S}}$, which reads "**S** is approximated by $\hat{\mathbf{S}}$". In this case the sense of the approximation is that of least-squares regression, where $\mathbf{U} = [\mathbf{y}^* \ \mathbf{x}^*]$ is the fixed matrix of explanatory variables and the regression coefficients $\mathbf{B}^{\mathsf{T}}$ are calculated in the usual way by least squares as follows:

$$\mathbf{B}^{\mathsf{T}} = (\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{S} \quad (2.5)$$

The complete process of the regression biplot can thus be summarized theoretically as follows:

$$\mathbf{S} \approx \hat{\mathbf{S}} = \mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{S} \quad (2.6)$$

The matrix $\mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}$ is called the *projection matrix* of **S** onto the explanatory variables in **U**. In fact, we can write **S** as the following sum:

$$\begin{aligned}\mathbf{S} &= \hat{\mathbf{S}} + (\mathbf{S} - \hat{\mathbf{S}}) \\ \mathbf{S} &= (\mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}})\mathbf{S} + (\mathbf{I} - \mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}})\mathbf{S} \quad (2.7)\end{aligned}$$

where the first term is the projection of **S** onto the space of the explanatory variables, and the second term is the projection of **S** onto the space orthogonal to (uncorrelated with) the explanatory variables. (**I** denotes the identity matrix, a diagonal matrix with 1's down the diagonal.) The part of **S** that is explicable by the explanatory variables **U** can be biplotted according to (2.6), as we have done in Exhibit 2.5, using **U** as the left matrix and the standardized regression coefficients in $(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{S}$ as the right matrix.

Notice that in these initial chapters we consider the case of only two explanatory variables, which conveniently gives a biplot in two dimensions. If we used the three variables "pollution", "depth" and "temperature" we would need three dimensions to show the variables and the regression relationships. We should also stress again that the "bi" in biplot does not refer to the bidimensional nature of the figures, but the fact that we depict rows and columns together. The case of three or more explanatory variables will be dealt with in later chapters (from Chapter 5 onwards).

SUMMARY: Regression Biplots

1. A *regression biplot* shows the cases (usually rows) and a set of response variables (usually columns) of a data matrix in the same joint representation, which is constructed using a set of explanatory variables, or predictors. Cases are shown as biplot points with respect to standardized values of the predictors and variables are shown as biplot vectors, each according to the standardized regression coefficients of its regression on the predictors.
2. The biplot vectors represent the separate linear regressions and define biplot axes onto which the case points can be projected. The axes can be calibrated so that predicted values from the regressions can be read off.
3. The quality of the regression biplot is measured by the percentages of variance explained by the individual regressions that build the biplot.