Scientific Computing for Biologists

Introduction to Biplots

Paul M. Magwene

04 October 2010

Consider this data set

Exhibit 1.1: *Economic data for 12 European countries in 2008. X_1 = purchasing power per capita (expressed in euros), X_2 = gross domestic product (GDP) per capita (indexed at 100 for all 27 countries in the European Union for 2008) and X_3 = inflation rate (percentage)*

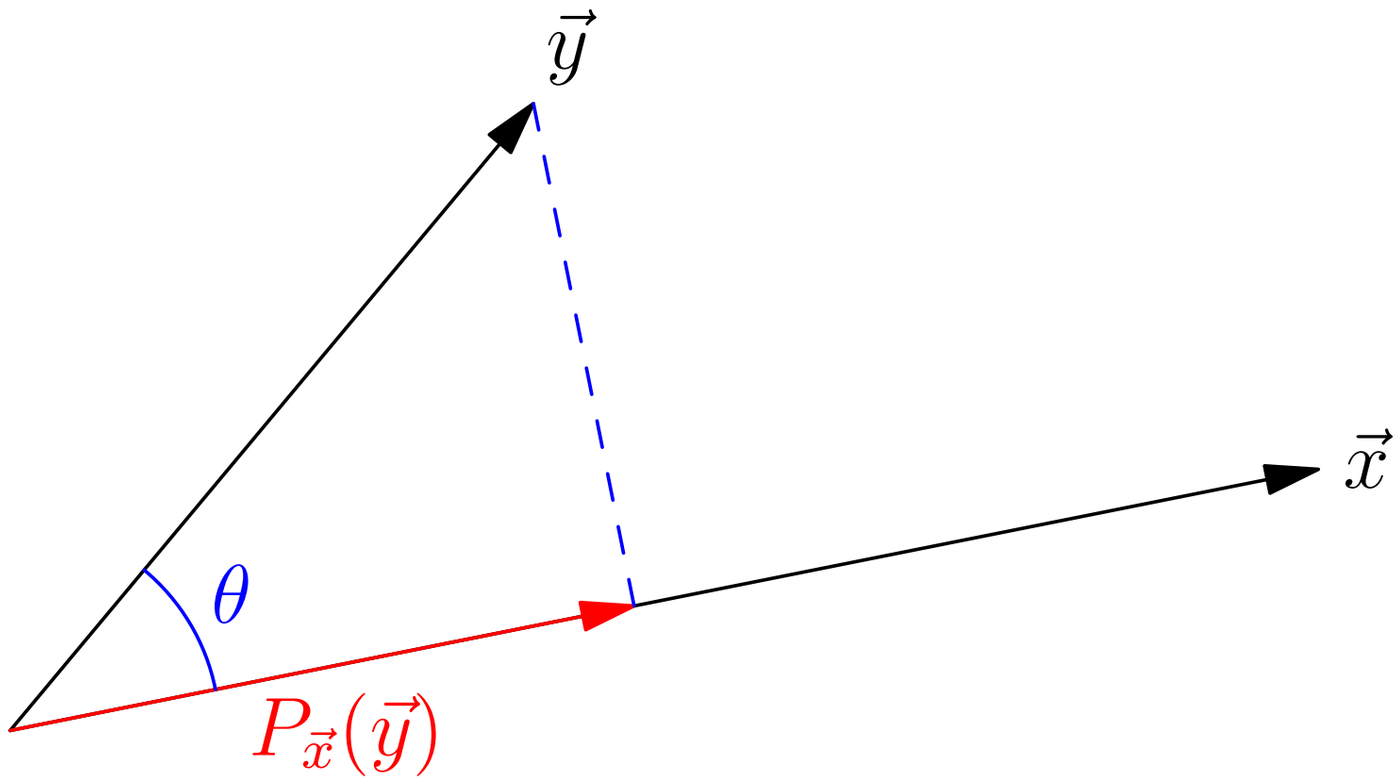
	Country	X_1	X_2	X_3
Be	Belgium	19,200	115.2	4.5
De	Denmark	20,400	120.1	3.6
Ge	Germany	19,500	115.6	2.8
Gr	Greece	18,800	94.3	4.2
Sp	Spain	17,600	102.6	4.1
Fr	France	19,600	108.0	3.2
Ir	Ireland	20,800	135.4	3.1
It	Italy	18,200	101.8	3.5
Lu	Luxembourg	28,800	276.4	4.1
Ne	Netherlands	20,400	134.0	2.2
Po	Portugal	15,000	76.0	2.7
UK	United Kingdom	22,600	116.2	3.6

from Greenacre (2010), *Biplots in practice.*

Do the following

1. Download `EU2008.txt` from the class website
2. Import the data set into R
3. Calculate the mean of each variable
4. Calculate a correlation matrix for the data set

A bit of review



$$\cos\theta = ?$$

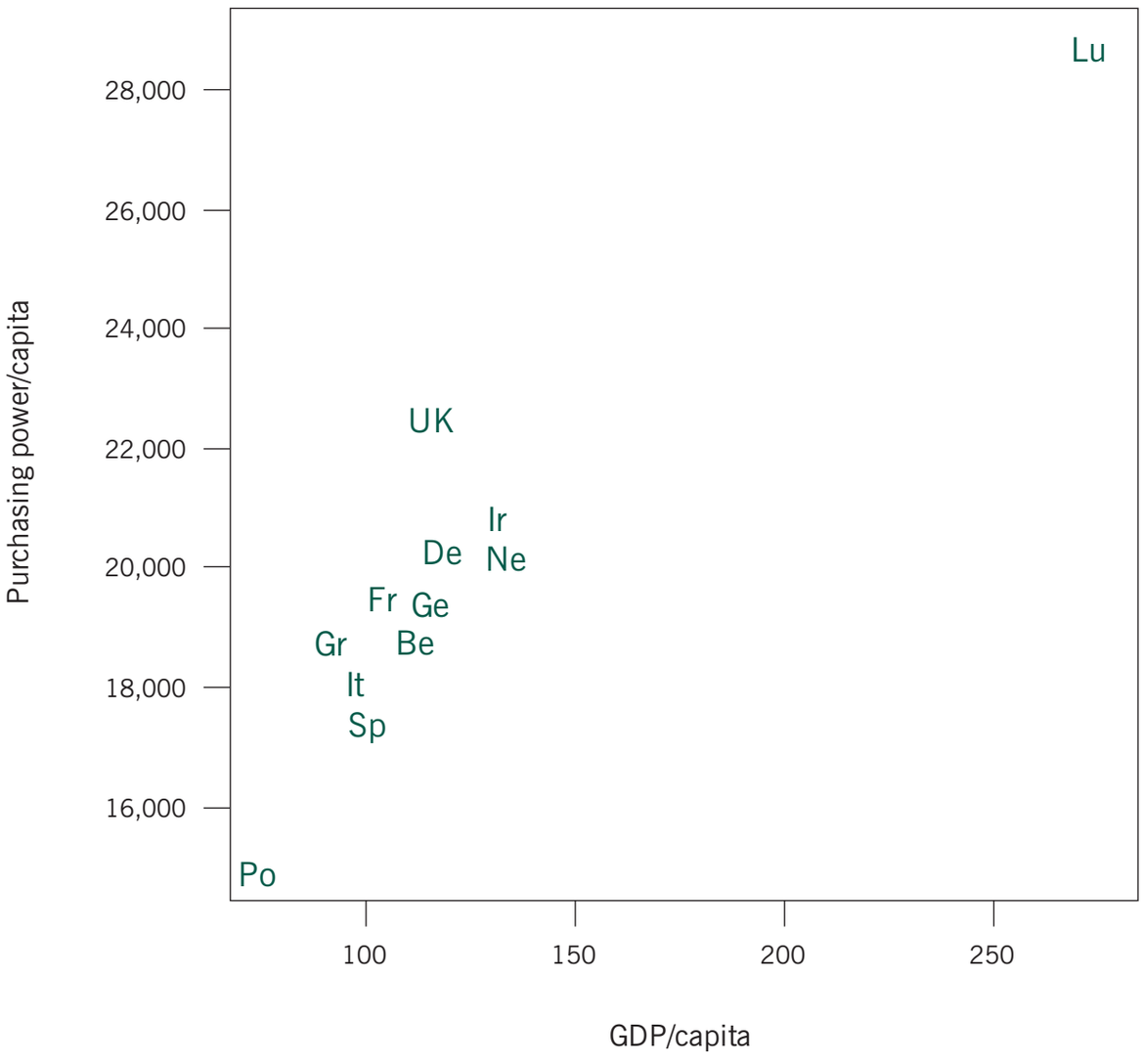
$$P_{\vec{x}}(\vec{y}) = ? \qquad C_{\vec{x}}(\vec{y}) = ?$$

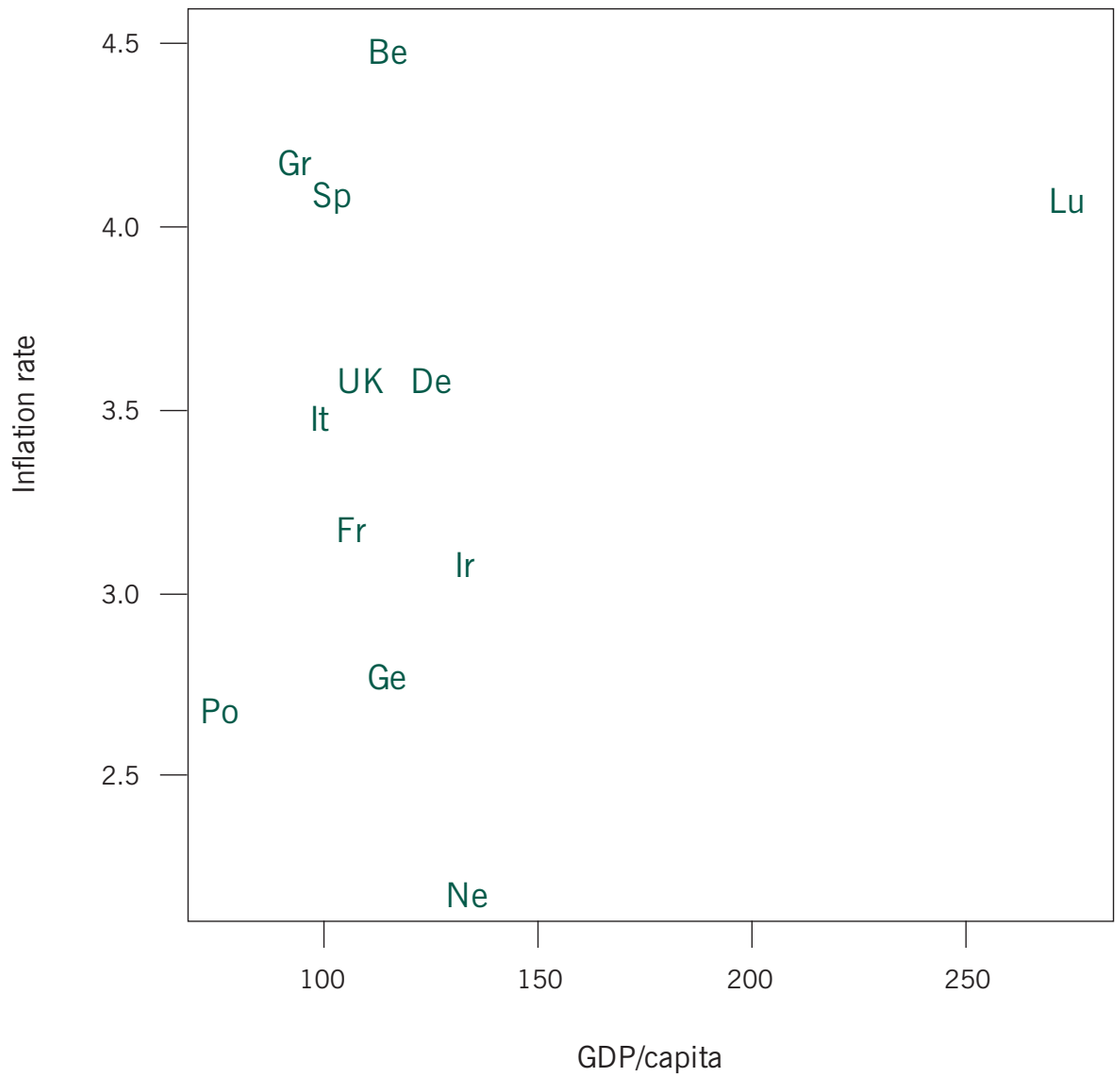
Do the following

Recall that we can represent variables as vectors. For each pair of variables:

1. Calculate the angle, in radians and degrees, between their corresponding vectors
2. Draw a sketch showing their geometric relationships

Recreate these plots





from Greenacre (2010), *Biplots in practice.*

Here's how I did it

```
> data <- read.table('EU2008.txt')
> names(data) <- c("Purch.Power", "GDP",
    "Infl.")
> plot(data$GDP, data$Purch.Power, type=
    "n", xlab="GDP/capita", ylab="
    Purchasing Power/capita")
> text(data$GDP, data$Purch.Power,
    rownames(data))
```

Create a 3D plot using scatterplot3d

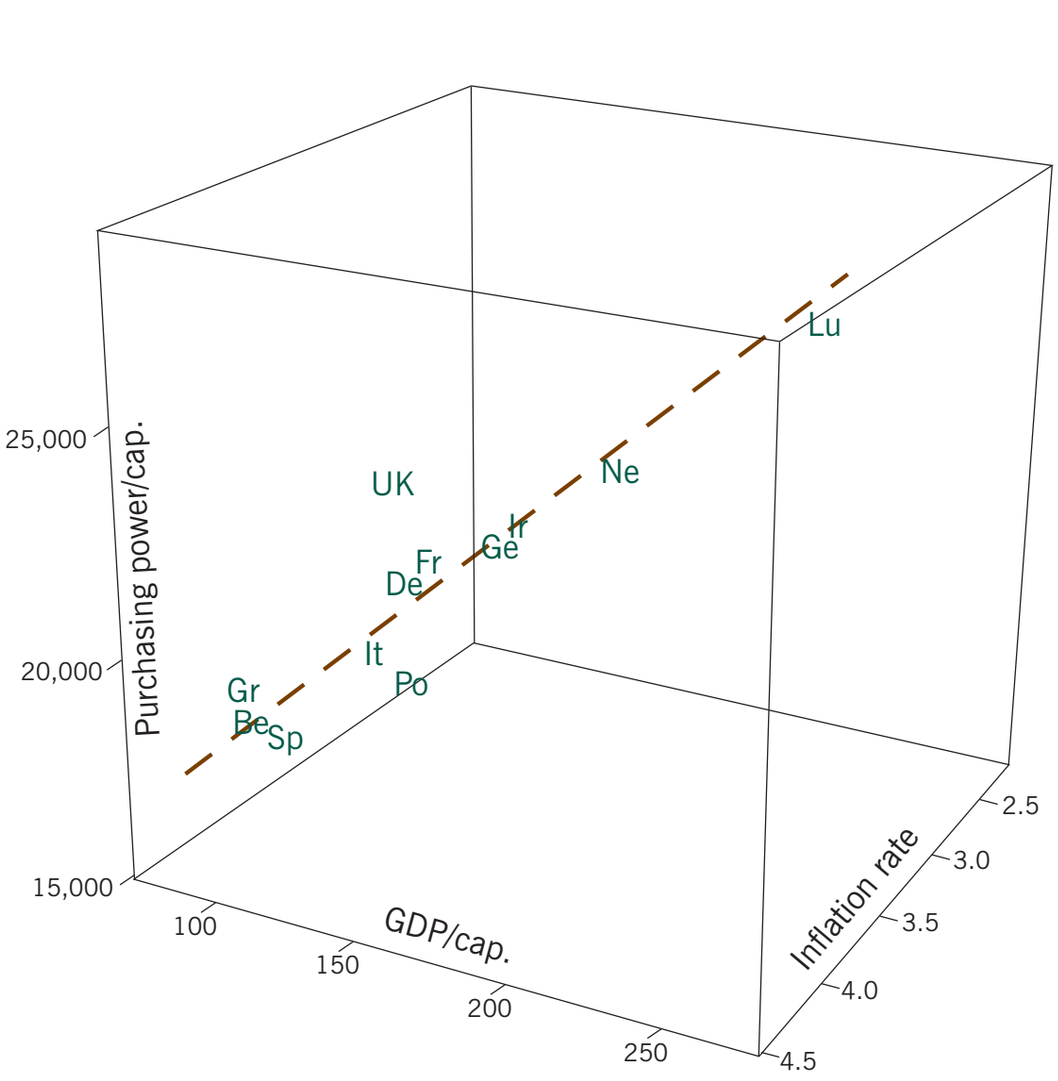
```
> library(scatterplot3d)
> ?scatterplot3d # read the help!
> scatterplot3d(data$GDP, data$Purch.Power,
    data$Infl)
```

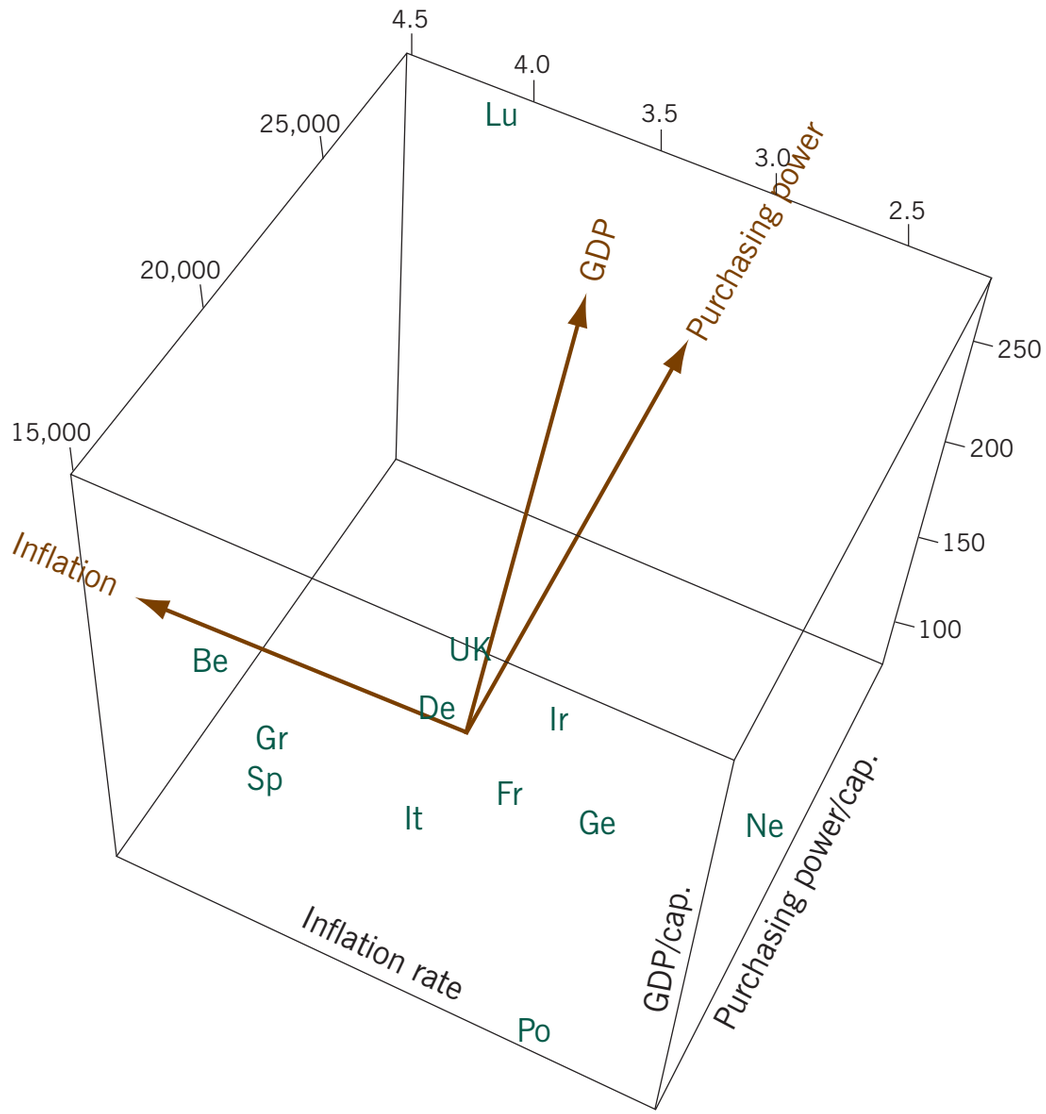
Create a 3D plot using plot3d from the rgl library

```
> library(rgl)
> ?plot3d
> plot3d(data$GDP, data$Purch.Power,
    data$Infl)
```

Notice that you can interact with the plot!

3D projections of the data set





from Greenacre (2010), *Biplots in practice.*

Matrix decomposition

A matrix decomposition is a factoring of a matrix into simpler parts.

Some familiar factorizations

$$12 = 3 \times 4 = 2 \times 6$$

$$x^2 - 4 = (x - 2)(x + 2)$$

Matrix decomposition is the same idea

Matrix decomposition

$$\textit{target matrix} = \textit{left matrix} \cdot \textit{right matrix}$$

$$\mathbf{S} = \mathbf{X}\mathbf{Y}^{\mathsf{T}} = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \mathbf{x}_3^{\mathsf{T}} \\ \mathbf{x}_4^{\mathsf{T}} \\ \mathbf{x}_5^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 & \mathbf{y}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}}\mathbf{y}_1 & \mathbf{x}_1^{\mathsf{T}}\mathbf{y}_2 & \mathbf{x}_1^{\mathsf{T}}\mathbf{y}_3 & \mathbf{x}_1^{\mathsf{T}}\mathbf{y}_4 \\ \mathbf{x}_2^{\mathsf{T}}\mathbf{y}_1 & \mathbf{x}_2^{\mathsf{T}}\mathbf{y}_2 & \mathbf{x}_2^{\mathsf{T}}\mathbf{y}_3 & \mathbf{x}_2^{\mathsf{T}}\mathbf{y}_4 \\ \mathbf{x}_3^{\mathsf{T}}\mathbf{y}_1 & \mathbf{x}_3^{\mathsf{T}}\mathbf{y}_2 & \mathbf{x}_3^{\mathsf{T}}\mathbf{y}_3 & \mathbf{x}_3^{\mathsf{T}}\mathbf{y}_4 \\ \mathbf{x}_4^{\mathsf{T}}\mathbf{y}_1 & \mathbf{x}_4^{\mathsf{T}}\mathbf{y}_2 & \mathbf{x}_4^{\mathsf{T}}\mathbf{y}_3 & \mathbf{x}_4^{\mathsf{T}}\mathbf{y}_4 \\ \mathbf{x}_5^{\mathsf{T}}\mathbf{y}_1 & \mathbf{x}_5^{\mathsf{T}}\mathbf{y}_2 & \mathbf{x}_5^{\mathsf{T}}\mathbf{y}_3 & \mathbf{x}_5^{\mathsf{T}}\mathbf{y}_4 \end{pmatrix}$$

Do the following

1. Create this matrix in R

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix}$$

2. Calculate a correlation matrix for the data set

Consider this decomposition

$$\textit{target matrix} = \textit{left matrix} \cdot \textit{right matrix}$$

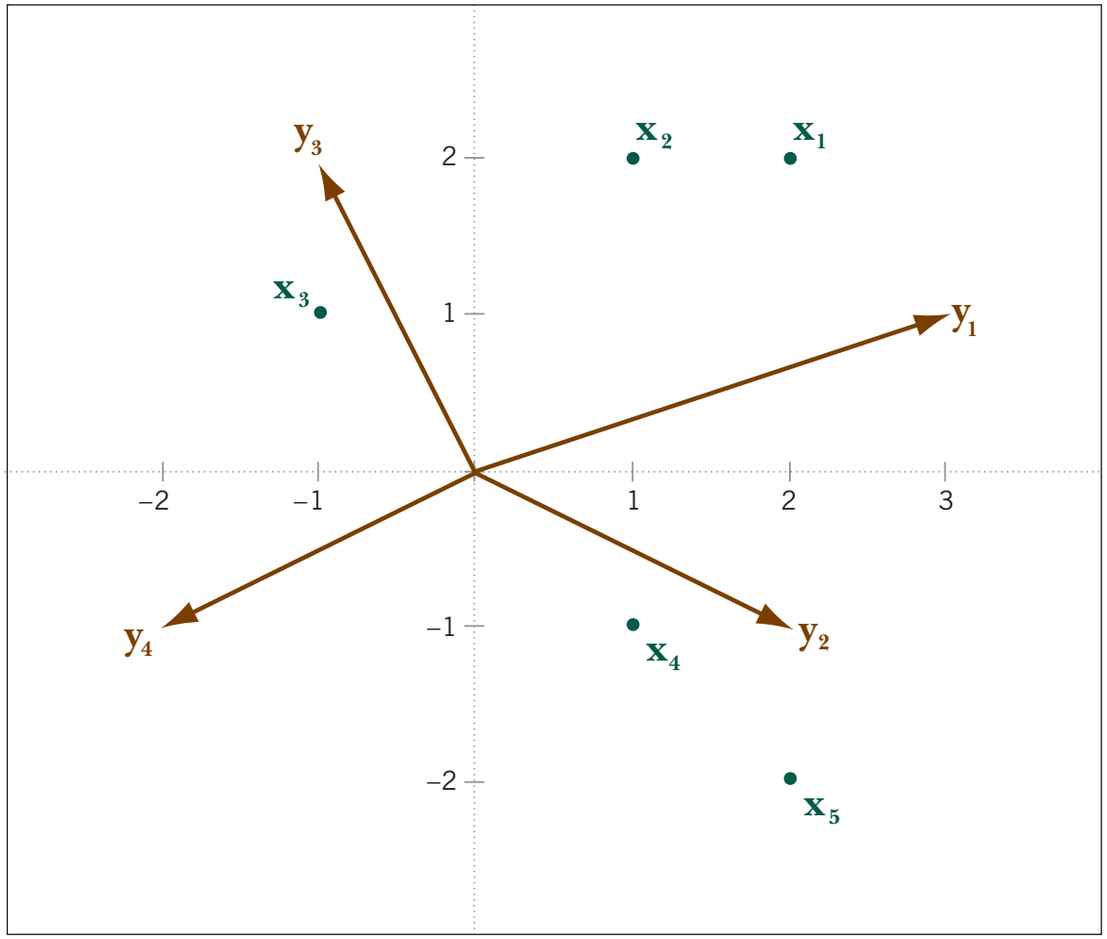
$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ -1 & 1 \\ 1 & -1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 3 & 2 & -1 & -2 \\ 1 & -1 & 2 & -1 \end{pmatrix}$$

In R, confirm that this is a valid decomposition.

Geometry of the decomposition

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ -1 & 1 \\ 1 & -1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 3 & 2 & -1 & -2 \\ 1 & -1 & 2 & -1 \end{pmatrix}$$

Exhibit 1.4:
The five points $\mathbf{x}_i$ of the left matrix and four points $\mathbf{y}_j$ of the right matrix in decomposition (1.1) (the latter points are shown as vectors connected to the origin). The scalar product between the i-th row point and the j-th column point gives the (i,j)-th value s_{ij} of the target matrix in (1.1)



$\mathbf{x}_1 = [\ 2 \quad 2\ ]^T$

$\mathbf{x}_2 = [\ 1 \quad 2\ ]^T$

$\mathbf{x}_3 = [-1 \quad 1\ ]^T$

$\mathbf{x}_4 = [\ 1 \quad -1\ ]^T$

$\mathbf{x}_5 = [\ 2 \quad -2\ ]^T$

$\mathbf{y}_1 = [\ 3 \quad 1\ ]^T$

$\mathbf{y}_2 = [\ 2 \quad -1\ ]^T$

$\mathbf{y}_3 = [-1 \quad 2\ ]^T$

$\mathbf{y}_4 = [-2 \quad -1\ ]^T$

from Greenacre (2010), *Biplots in practice.*

Calibrating the biplot

$$\text{target value in row } i = \begin{pmatrix}\text{length of projection}\\ \text{of } i\text{-th biplot point}\end{pmatrix} \times \begin{pmatrix}\text{length of biplot}\\ \text{vector}\end{pmatrix}$$

$$\text{length of projection of one unit} = 1 \,/\, \text{length of biplot vector}$$

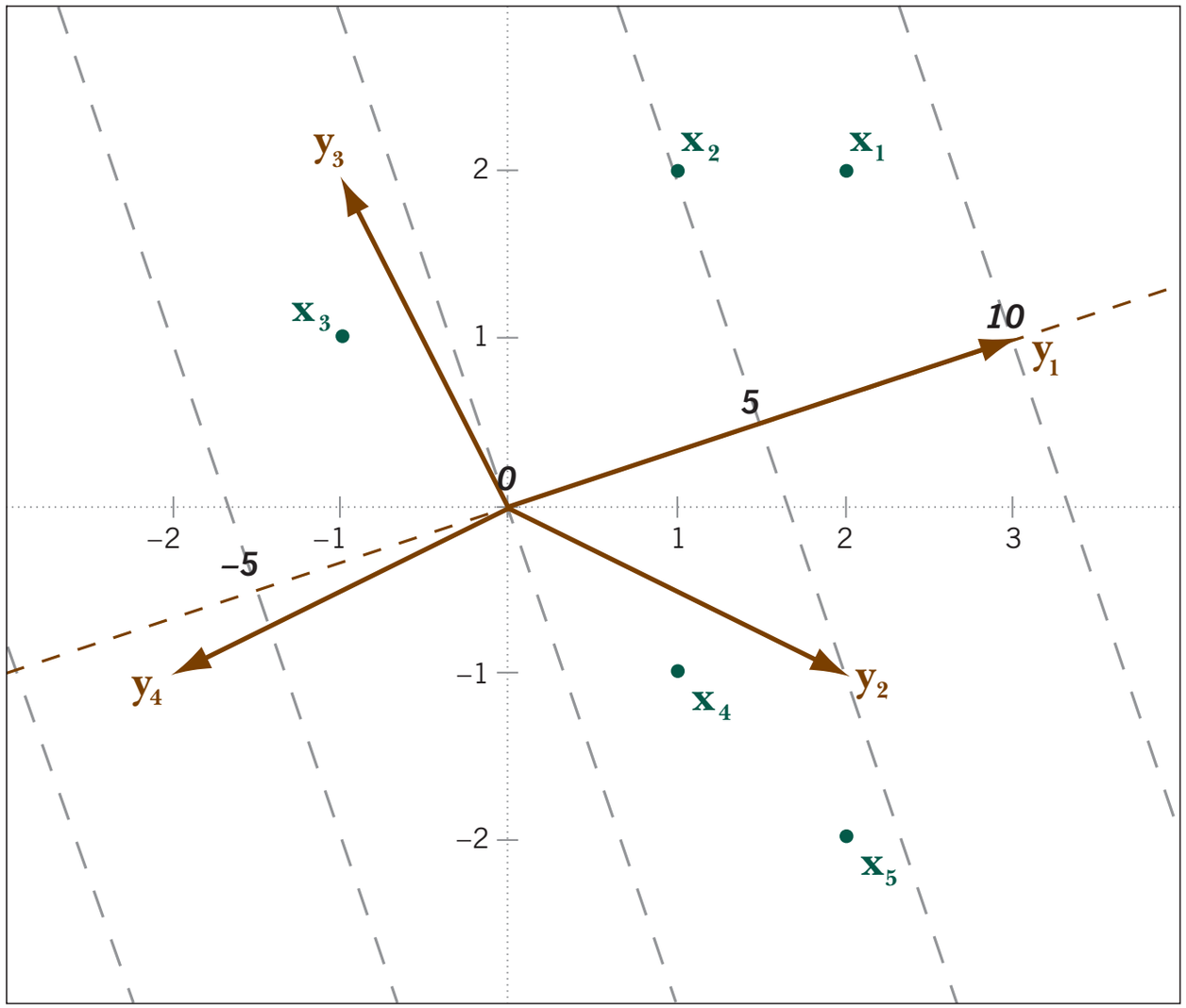


Exhibit 1.6:
Calibrating a biplot axis through vector $\mathbf{y}_1$*, shown as dashed line. The distance between units on this axis is the inverse of the length of* $\mathbf{y}_1$*, 0.3162, and allows placing values on the axis (shown in black). Points projected perpendicularly onto the biplot axis give values on the calibrated scale equal to the values in the first column of the target matrix (corresponding to* $\mathbf{y}_1$*, the first biplot vector). Thus we can read off the target values of 8, 5, –2, 2 and 4 for points* $\mathbf{x}_1$*, . . .,* $\mathbf{x}_5$*, respectively —see first column of target matrix in (1.1)*

from Greenacre (2010), *Biplots in practice.*

1. Biplots are defined as the decomposition of a *target matrix* into the product of two matrices, called *left and right matrices*: $\mathbf{S} = \mathbf{X}\mathbf{Y}^T$

2. Elements in the target matrix $\mathbf{S}$ are equal to *dot products* between corresponding pairs of vectors in the rows of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

3. Geometric interpretation – the vectors in the left and right matrices provide two sets of points, one of which can be considered as a set of *biplot vectors* defining *biplot axes*, and the other as a set of *biplot points*. Points can be projected perpendicularly onto biplot axes to recover the values in the target matrix.

4. The "bi" in biplot refers to the fact that two sets of points (i.e., the rows and columns of the target matrix) are visualized by dot products, not the fact that the display is usually two-dimensional.

5. The biplot and its geometry hold for spaces of any dimensionality, but we need need dimension-reducing techniques when data matrices have high inherent dimensionality and we want a low-dimensional representation

from Greenacre (2010), *Biplots in practice.*

Regression biplots

Consider the matrix equation of multiple regression:

$$\hat{Y} = XB$$

Look familiar?

$$\textit{target matrix} = \textit{left matrix} \cdot \textit{right matrix}$$

Consider this data set

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

from Greenacre (2010), *Biplots in practice.*

Do the following

1. Download `bioenv.txt` from the class website
2. Import the data set into R
3. Create a data subset that includes just the numerical variables
4. Calculate the means of the numerical variables
5. Calculate a correlation matrix for the numerical variables

Do the following

1. Convert the numerical variables to centered and standardized variables (z-scores) – see R function `scale()`

2. With the standardized scores, calculate the multiple regression of variable ‘d’ on Pollution and Depth

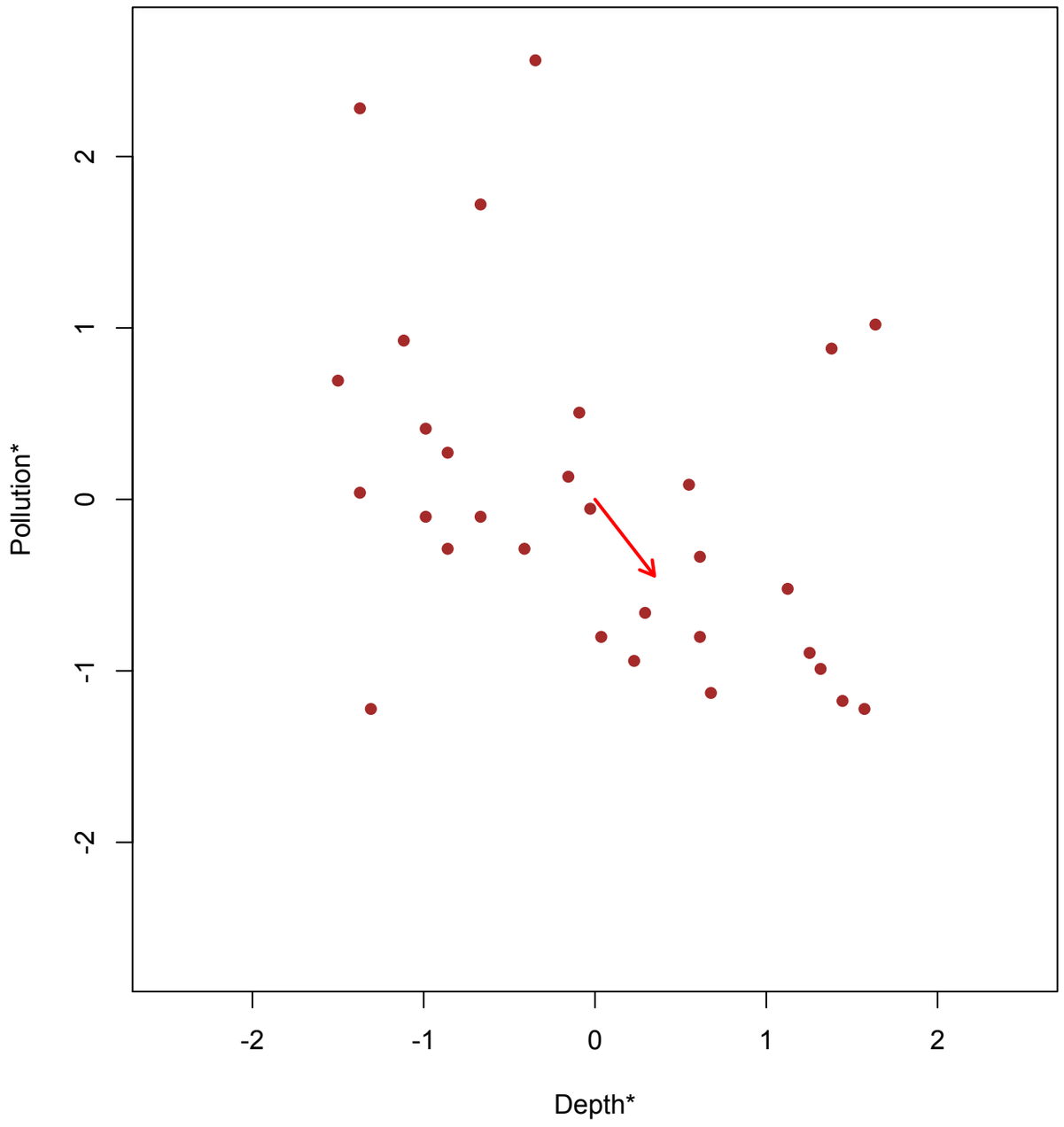
3. What regression model do you get?

4. What is the fraction of variance explained by the model?

Do the following

1. Create a scatter plot with standardized Depth on the x-axis and standardized Pollution on the y-axis
    - Set the limits for the x- and y-axes to go from -2.5 to 2.5
    - Set the 'aspect' of the plot to 1 using the asp argument
    - Do `?plot.default` for info about the above
2. Draw a red arrow from the point (0,0) to the point (0.347,-0.446)
    - See the `arrows()` function in R.

Here's what I got



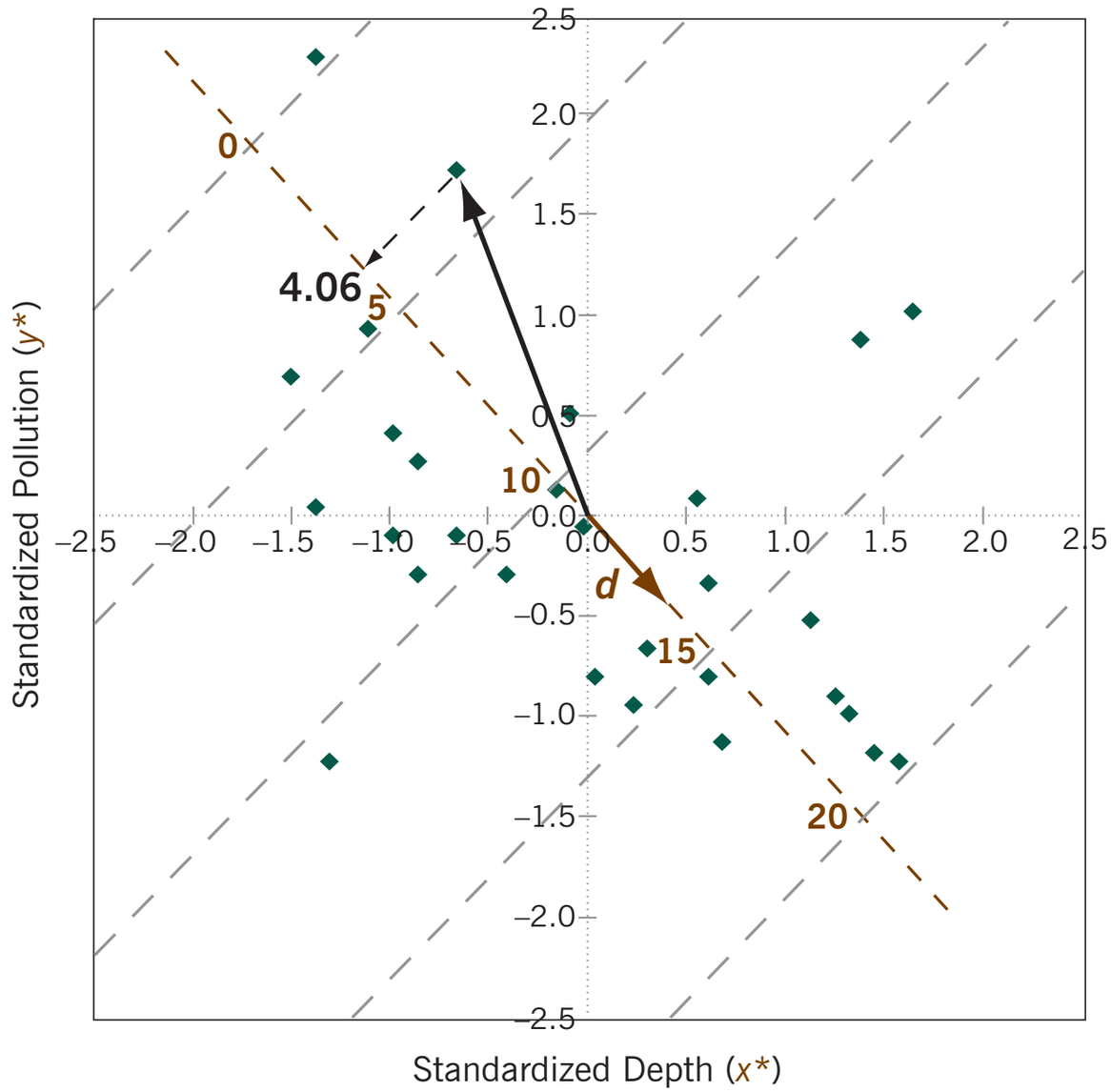

```
> plot(s.sub$Depth, s.sub$Pollution, xlim=c(-2.5,2.5),
    ylim=c(-2.5,2.5),asp=1, xlab='Depth*',ylab='
    Pollution*',pch=16,col='Brown')
> arrows(0,0, 0.347, -0.446, col='red',lwd=2,length
    =0.1)
```

What's the biplot telling us?

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

from Greenacre (2010), *Biplots in practice.*

Do the following

1. Carry out similar regressions for each of the other species group variables (`a`, `b`, `c`, `e`) and plot their biplot axes in the same plot space as `d`. Recreate the following figure:

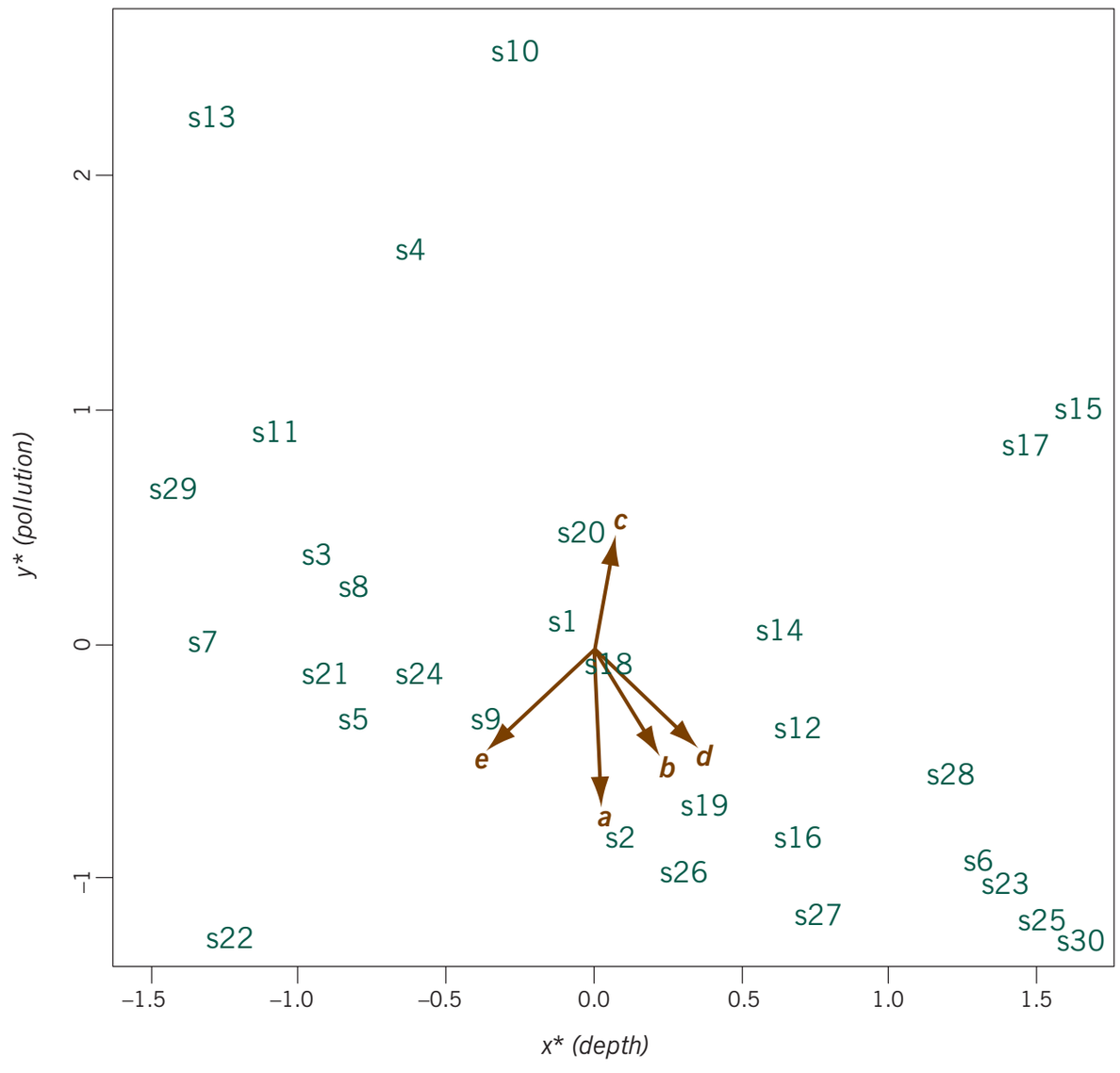


Exhibit 2.5:
*Regression biplot of five response variables, species **a** to **e**, in the space of the two standardized explanatory variables. The overall explained variance for the five regressions is 41.5%, which is the measure of fit of the biplot*

from Greenacre (2010), *Biplots in practice*.

Before you leave:

1. Show me your final plot
2. Clarify any issues or details that you are confused abouts