Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 7 Offprint

Log-ratio Biplots

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

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

CHAPTER 7

Log-ratio Biplots

All the techniques described in this book are variations of generalized principal component analysis defined in Chapter 5, and in Chapter 6 we demonstrated the simplest version of principal component analysis. As mentioned at the end of Chapter 6, when variables are on different scales they are usually standardized in some way to equalize the roles of the variables in the analysis: this can be thought of either as a pre-transformation of the data or equivalently as a reweighting of the variables. Many other pre-transformations of the data are possible: for example, to visualize multiplicative differences in the data the logarithmic transformation can be applied, in which case no further standardization is required. Data that are proportions are often transformed by the arcsine function (i.e., the inverse sine function) or by some power function such as the square root. In this chapter we treat the log-ratio transformation which is applicable to a common situation in practice: when data are all measured on the same units and strictly positive. The biplots that result have some special properties and this approach deserves a wider usage, hence a whole chapter is devoted to it.

Contents

Interval and ratio scales

What has not been stated or explained up to now is that PCA assumes the data are on *interval* scales. By this we mean that when we compare two values, we look at their (interval) differences. For example, if we compare a temperature of 3.6 degrees with 3.1 degrees, we say that the difference is 0.5 degrees and this difference is comparable to the difference between 21.9 and 21.4 degrees. On the oth-

er hand, many variables are measured on *ratio* scales where we would express the comparison as a multiplicative, or percentage, difference. For example, a factory worker obtains an increase in salary of 50 euros a month, but his salary before was only 1000 euros a month, so this is in fact a 5% increase; if he were earning 3000 euros before, the increase would be 1.67%. Here it is relevant to compare the ratios of the numbers being compared: 1050/1000 and 3050/3000, not their differences. One has to carefully consider whether the observed variables are on interval or ratio scales, as this affects the way we analyze them. In practice, the values of the variable may be so far away from the zero point of their scale that the distinction between interval and ratio scale is blurred: for example, a 50-unit increase on the much higher value of 100,000 is not much different, percentage-wise, from a 50-unit increase on the higher value of 110,000.

The logarithmic transformation

The classic way to treat ratio-scale variables is to apply the logarithmic transformation, so that multiplicative differences are converted to additive differences: $\log(x/y) = \log(x) - \log(y)$. If the variables are all ratio-scale in nature but in different measurement units, a blanket logarithmic transformation on all of them is an excellent alternative to variable standardization. For example, an economist might analyze the behaviour of several stock market indices such as Dow-Jones, Financial Times, Nikkei, CAC40, etc (the variables) over time (the rows). Each variable has its own inherent scale but differences between them are evaluated multiplicatively (i.e., percentage-wise). The logarithmic transformation will put them all on comparable interval scales, perfect for entering into a PCA, without any standardization necessary. In fact, standardization would be incorrect in this case, since we want to compare the natural variation of the indices on the logarithmic scale, and not equalize them with respect to one another. If we biplotted such data, then the variables would be calibrated non-linearly on a logarithmic scale, reminiscent of the biplots described in Chapter 3.

Log-ratios

Log-ratios are a bit more specialized than logarithms. Not only are values compared within each variable on a ratio scale, but also values within each case are compared across the variables. This means that all the variables must be measured on the same scale. This approach originated in compositional data analysis in fields such as chemistry and geology, where the variables are components of a sample and the data express proportions, or percentages, of the components in each sample (hence the values for each sample sum to 1, or 100%). It is equally applicable to data matrices of strictly positive numbers such as values all in dollars, measurements all in centimetres, or all counts. The R data set `USArrests` has the 50 US states as rows, and the columns are the numbers of arrests per 100,000 residents of three violent crimes: murder, assault and rape. The "ratio" in log-ratio analysis can refer either to ratios within a state or ratios within a crime. The first five rows of this data set are:

```
> USArrests[1:5,c(1,2,4)]
           Murder     Assault       Rape
Alabama      13.2         236       21.2
Alaska       10.0         263       44.5
Arizona       8.1         294       31.0
Arkansas      8.8         190       19.5
California    9.0         276       40.6
```

By ratios within a row (state) we mean the three unique ratios Murder/Assault, Murder/Rape and Assault/Rape, which for Alabama are (to four significant figures) 0.05593, 0.6226 and 11.13 respectively. By ratios within a column (crime) we mean the $50 \times 49/2 = 1225$ pairwise comparisons between states, of which the first four for the column Murder are Alabama/Alaska: 1.320, Alabama/Arizona: 1.630, Alabama/Arkansas: 1.500, Alabama/California: 1.467. The basic idea in log-ratio analysis is to analyze all these ratios on a logarithmic scale, which are interval differences between the logarithms of the data. In general, for a $I \times J$ matrix there are $\frac{1}{2}I(I-1)$ unique ratios between rows and $\frac{1}{2}J(J-1)$ unique ratios between columns. Fortunately, we do not have to calculate all the above ratios—there is a neat matrix result that allows us to work on the original $I \times J$ matrix and effectively obtain all the log-ratios in the resulting map.

Log-ratio analysis

The algorithm for performing *log-ratio analysis* (LRA) relies on a double-centring of the log-transformed matrix and a weighting of the rows and columns proportional to the margins of the data matrix $\mathbf{N}$:

– Let the row and column sums of $\mathbf{N}$, relative to its grand total $n = \sum_i \sum_j n_{ij}$ be denoted by $\mathbf{r}$ and $\mathbf{c}$ respectively:

$$\mathbf{r} = (1/n)\mathbf{N1},\ \mathbf{c} = (1/n)\mathbf{N}^{\mathsf{T}}\mathbf{1} \tag{7.1}$$

– Logarithmic transformation of elements of $\mathbf{N}$: $\mathbf{L} = \log(\mathbf{N})$ (7.2)

– Weighted double-centring of $\mathbf{L}$: $\mathbf{Y} = (\mathbf{I} - \mathbf{1r}^{\mathsf{T}})\mathbf{L}(\mathbf{I} - \mathbf{1c}^{\mathsf{T}})^{\mathsf{T}}$ (7.3)

– Weighted SVD of $\mathbf{Y}$: $\mathbf{S} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{Y}\,\mathbf{D}_c^{\frac{1}{2}} = \mathbf{U}\mathbf{D}_\varphi\mathbf{V}^{\mathsf{T}}$ (7.4)

– Calculation of coordinates:

Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\varphi$, of columns: $\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\varphi$ (7.5)

Standard coordinates of rows: $\mathbf{\Phi} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}$, of columns: $\mathbf{\Gamma} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$ (7.6)

(The above analysis is the weighted form of LRA, which is usually preferred above the unweighted form, which has equal weights on the rows and columns;

that is, unweighted LRA uses the same steps (7.1) to (7.6) but with $\mathbf{r} = (1/I)\mathbf{1}$, $\mathbf{c} = (1/J)\mathbf{1}$.)

As before, two biplots are possible, but in this case they have completely symmetric interpretations—in our terminology of Chapter 6, they are actually both form biplots. The row-metric preserving biplot of $\mathbf{F}$ and $\boldsymbol{\Gamma}$ plots the rows according to their log-ratios across columns, while the column-metric preserving biplot of $\mathbf{G}$ and $\boldsymbol{\Phi}$ plots the columns according to their log-ratios across rows. The points displayed in standard coordinates represent all the log-ratios by vectors between pairs of points, called *link* vectors. It is the double-centring in (7.3) which gives this special result that analyzing the matrix of I rows and J columns yields the representation of all the pairwise log-ratios. To demonstrate these geometric features with a simple example, we show the row-principal ($\mathbf{F}$, $\boldsymbol{\Gamma}$) LRA biplot of the USArrests data in Exhibit 7.1.

The double-centring removes one dimension from the data, hence the dimensionality of this 3-column matrix is equal to 2 and Exhibit 7.1 displays 100% of the variance, equal to 0.01790. This can be explained alternatively by the fact that any of the three log-ratios is linearly dependent on the other two, hence the rank of the matrix of log-ratios is 2. In this LRA biplot it is not the positions of the three columns that are of interest but the link vectors joining them, which represent the pairwise log-ratios. For example, the link from Rape to Assault represents the

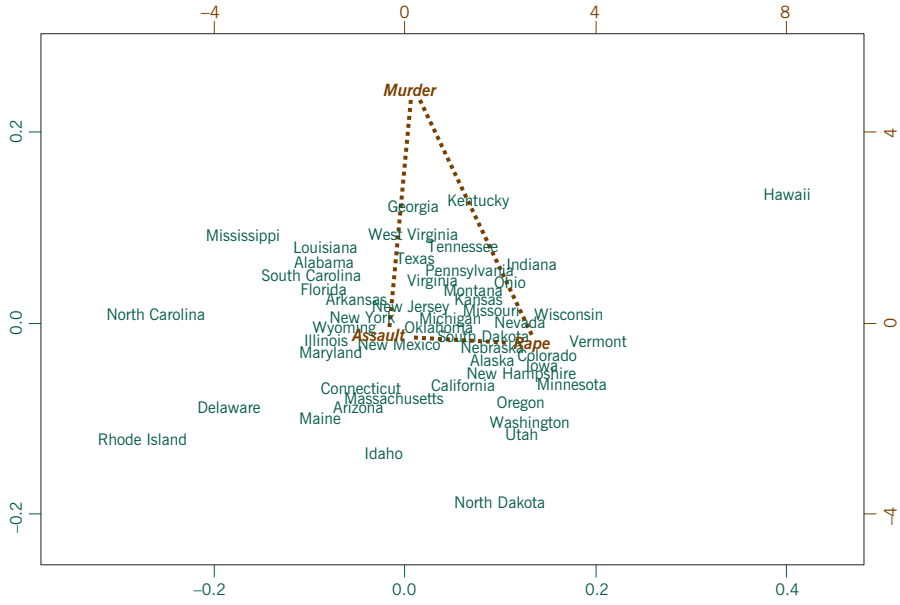


Exhibit 7.1: *Log-ratio biplot of the* `USArrests` *data set from the* R *package, with rows in principal and columns in standard coordinates. The columns are connected by links which represent the pairwise log-ratios. 100% of the log-ratio variance is displayed. Notice the different scales for the two sets of points*

logarithm of Rape/Assault, and the link in the opposite direction represents the negative of that log-ratio, which is the logarithm of Assault/Rape. A biplot axis can be drawn through this link vector and the states projected onto it. The position of Hawaii separates out at the extreme right of this Rape/Assault axis—in fact, Hawaii has an average rate of Rape, but a very low rate of Assault, so the Rape/Assault ratio is very high. Likewise, to interpret Murder/Assault log-ratios, project the states onto this vector which is almost vertical. Hawaii projects high on this axis as well, but it projects more or less at the average of the Murder/Rape link. To know what "average" is, one has to imagine the link translated to the origin of the biplot, which represents the average of all the log-ratios; that is, draw an axis through the origin parallel to the Murder/Rape link—the projection of Hawaii is then close to the origin. Alternatively, project the origin perpendicularly onto the links and this indicates the average point on the corresponding log-ratios.

Log-ratio distance and variance

The positions of the states in Exhibit 7.1 are a representation of *log-ratio distances* between the rows. This distance is a weighted Euclidean distance between the log-ratios within each row, for example the squared distance between rows i and i':

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{ij'}} - \log \frac{n_{i'j}}{n_{i'j'}} \right)^2 \tag{7.7}$$

This can be written equivalently as:

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} - \log \frac{n_{ij'}}{n_{i'j'}} \right)^2 \tag{7.8}$$

showing that log-ratios can be considered between the pair of values in corresponding columns. Both (7.7) and (7.8) can be written equivalently in terms of the logarithms of odds-ratios for the four cells defined by row indices i,i' and column indices j,j':

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} \frac{n_{i'j'}}{n_{ij'}} \right)^2 \tag{7.9}$$

The log-ratio distances $d^2_{jj'}$ between columns in the alternative column-principal biplot are the symmetric counterparts of (7.7), (7.8) or (7.9), with index i substituting j in the summations, and r_i substituting c_j.

The total variance of the data is measured by the sum of squares of (7.4), which can be evaluated as a weighted sum of squared distances $\sum\sum_{i<i'} r_i r_{i'} d^2_{ii'}$, for example using the definition (7.9) of squared distance in terms of the odds-ratios:

$$\sum\sum_{i<i'}\sum\sum_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} \frac{n_{ij'}}{n_{ij'}} \right)^2 \tag{7.10}$$

In the above example the total variance is equal to 0.01790.

Data set "morphology"

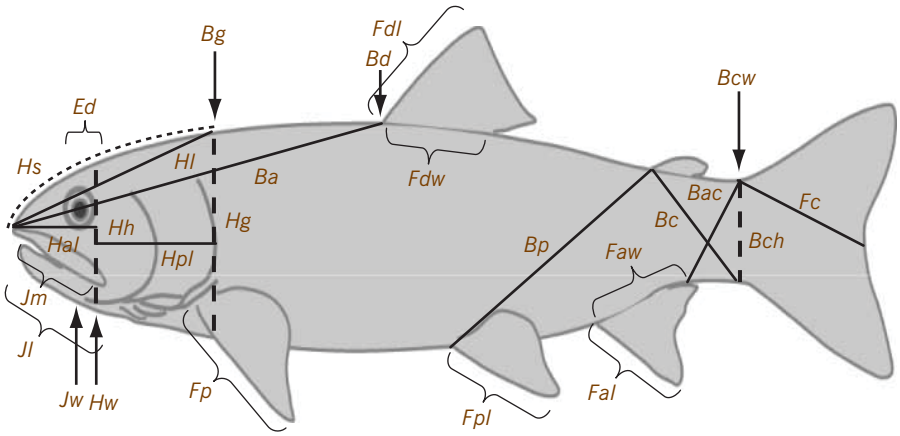
The LRA biplot works well for any strictly positive data that are all measured on the same scale, and for which multiplicative comparisons of data elements, row- or column-wise, make more sense than additive (interval) comparisons. Morphometric data in biology are an excellent candidate for this approach, so we show an application to a data set of 26 measurements on 75 *Arctic charr* fish. The data come from a study of the diet and habitat of the fish and their relationships to their body form and head structure.[5] Exhibit 7.2 shows the abbreviated names of the measurements.

The total variance in these data is 0.001961, much lower than the previous example, indicating that the fish are quite similar to one another in an absolute sense, which is not surprising since they are all of the same species. Nevertheless there are interesting differences between them which may be related to their environment and diets. In Exhibit 7.3 shows the row-principal LRA biplot, where the scale of the low-variance fish points has been enlarged 50 times to show them more legibly. The fish have been labelled according to their sex (f = female, m = male) and habitat where they were caught (L = littoral near shore, P = pelagic in open sea). There does not seem to be any apparent connection with the distribution of the fish and these labels—this can be tested formally using a permutation test, described in the Computational Appendix, while in Chapter 11 the topic of direct comparison of groups of cases is treated (as opposed to comparisons between individual cases, which is what is being analyzed here).

Diagnosing equilibrium relationships

The variable points have no relevance *per se*, rather it is the links between all pairs of variables that approximate the log-ratios—in fact, one could imagine all these links transferred to the origin as vectors representing the pairwise log-ratios. Thus the logarithm of the ratio *Bc/Hpl* (caudal peduncle length/posterior head length) has one of the highest variances—its calibrations, proportional to the inverse of

5. Data provided by Prof. Rune Knudsen and the freshwater biology group of the Department of Arctic and Marine Biology at the University of Tromsø, Norway.



Jl: lower jaw length; *Jw*: lower jaw width; *Jm*: upper jaw length; *Ed*: eye diameter; *Hw*: head width; *Hh*: head height; *Hg*: head height behind gills; *HpI*: posterior head length; *HI*: head length; *HaI*: anterior head length; *Hs*: snout length, head curvature from the snout to back of the gills; *Ba*: anterior body length from the snout to the dorsal fin; *Bp*: posterior body length from the anal fin to the adipose fin; *Bch*: caudal height; *Bc*: caudal peduncle length, from the adipose fin to ventral caudal height; *Bac*: caudal peduncle length, anal fin to dorsal caudal height; *Bg*: body width at gills; *Bd*: body width at dorsal fin; *Bcw*: caudal body width; *Fc*: caudal fin length; *Fp*: pectoral fin length; *FdI*: dorsal fin length; *Fdw*: dorsal fin width; *FpI*: pectoral fin length; *FaI*: anal fin length; *Faw*: anal fin width.

Exhibit 7.2:
Morphological characteristics from the left side of Arctic charr fish. Dashed lines indicate heights, arrows indicate widths

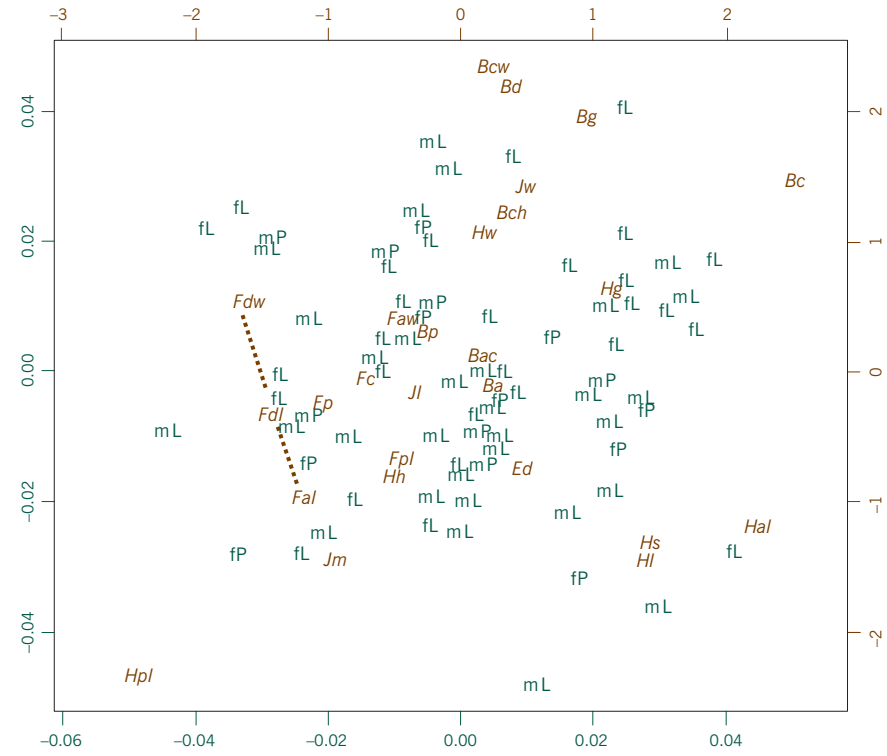


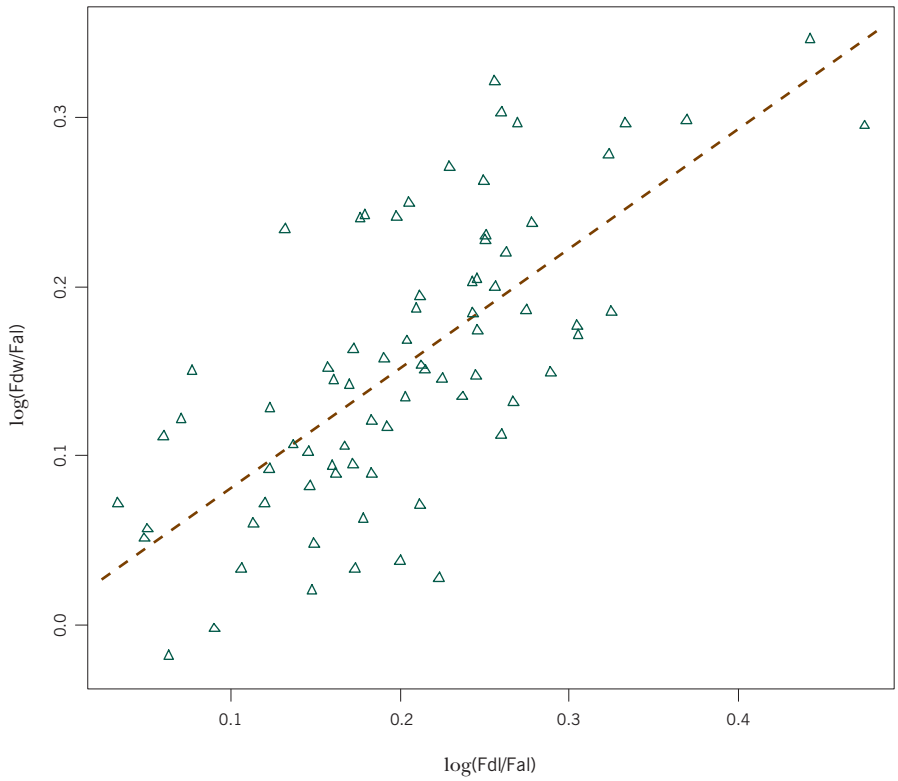
Exhibit 7.3:
Log-ratio biplot of the "morphology" data set, with rows in principal and column in standard coordinates. Labels for the fish are: fL = *female littoral;* mL = *male littoral;* fP = *female pelagic;* mP = *male pelagic. 34.5% of the total variance is explained in this map*

the vector length would be very close together, and thus the projections of the fish onto this direction would vary greatly in value.

A further interesting property of log-ratio analysis is that a certain class of models can be diagnosed between subsets of variables if they line up in straight lines in the biplot. In Exhibit 7.3 there is a lining up of the three variables Fdw (dorsal fin width), Fdl (dorsal fin length) and Fal (anal fin length), indicated by the dotted line. This means that the log-ratios formed from Fdw, Fdl and Fal could be linearly related—we say "could be" because only 34.5% of the variance is explained by the map and this lining up in the low-dimensional projection of the biplot does not necessarily mean that the points line up in the full space (however, not lining up in the biplot means they definitely do not line up in the full space). To investigate this possible relationship, Exhibit 7.4 shows a scatterplot of two log-ratios, and indeed there is a high positive correlation of 0.700 ($R^2 = 0.490$).

To quantify the relationship we find the best-fitting straight line through the points (this is the first principal axis of the points, not the regression line), and

Exhibit 7.4:
Plot of two log-ratios diagnosed from Exhibit 7.3 to be possibly in a linear relationship (the correlation is 0.70). The best-fitting line through the scatterplot has slope equal to 0.707 and intersection 0.0107



this line turns out to have slope 0.707, and intersection with the vertical axis at 0.0107. So the relationship simplifies to:

$$\log(\mathsf{Fdw}/\mathsf{Fal}) = 0.707 \log(\mathsf{Fdl}/\mathsf{Fal}) + 0.0107$$

Exponentiating:

$$\mathsf{Fdw}/\mathsf{Fal} = 1.0108 \times (\mathsf{Fdl}/\mathsf{Fal})^{0.707}$$

Simplifying:

$$\mathsf{Fdw} = 1.0108 \times \mathsf{Fdl}^{0.707} \times \mathsf{Fal}^{0.293} \qquad (7.11)$$

Then, calculating the predicted values of Fdw, the dorsal fin width, as a function of Fdl (dorsal fin length) and Fal (anal fin length), we get a good fit (correlation of 0.750, $R^2 = 0.562$) between the predicted and observed values (Exhibit 7.5).

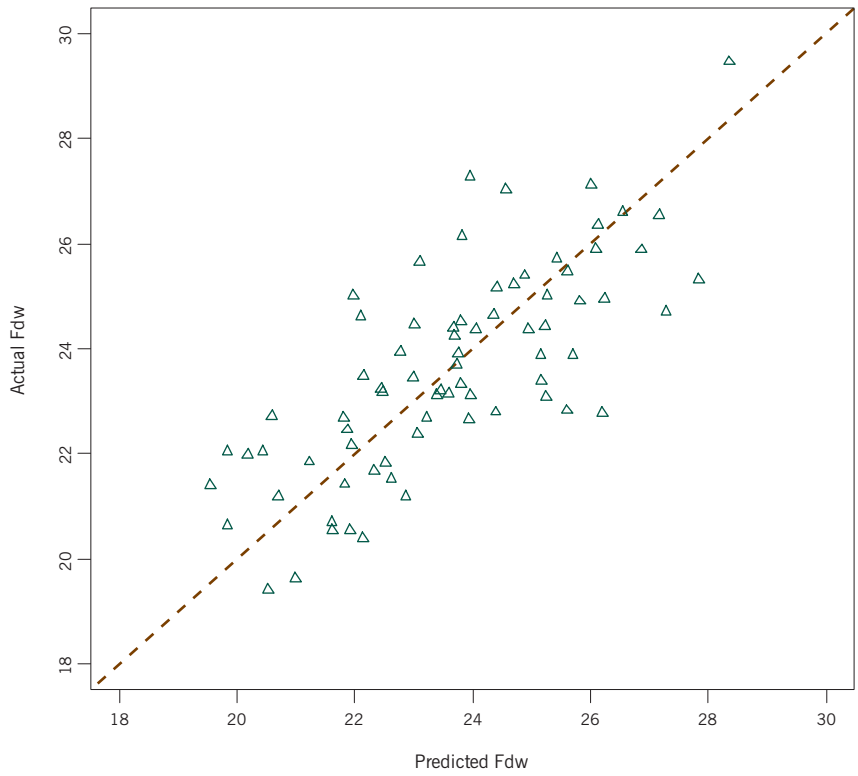


Exhibit 7.5:
Predicted versus actual values of Fdw *(dorsal fin width) based on the model of (7.11)*

Notice that the model of (7.11) really only has two parameters, the multiplicative constant and a single parameter for the two "predictor" variables, because their power coefficients sum to 1. It is this type of restricted parameter model that is called an "equilibrium model" in certain areas of research such as genetics, chemistry and geology.

SUMMARY: Log-ratio Biplots

1. Log-ratio analysis applies to any table of strictly positive data, where all data entries are measured on the same scale.
2. The original $I \times J$ table is logarithmically transformed and then double-centred, where the rows and columns are weighted proportionally to their marginal sums, followed by a SVD decomposition. The form biplot, where singular values are assigned to the left vectors corresponding to the cases, displays approximate Euclidean distances between the cases based on all the pairwise log-ratios of the variables.
3. Log-ratio biplots represent the pairwise log-ratios between all the columns, or between all the rows, as the case may be. These are the vectors that connect the pairs of columns or pairs of rows.
4. If a subset of columns, for example, line up in straight lines, this diagnoses possible equilibrium relationships in that subset, in the form of a multiplicative power model relating the columns.