Biplots in Practice

MICHAEL GREENACRE

Professor of Statistics at the Pompeu Fabra University

Chapter 3 Offprint

Generalized Linear Model Biplots

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

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

CHAPTER 3

Generalized Linear Model Biplots

Generalized linear models are a wide class of regression-type models for different types of data and analytical objectives. Linear regression is the simplest form of a generalized linear model, where the mean of the response variable, given values of the predictors, is a linear function of the predictors and the distribution of the data around the mean is normal. The regression coefficients are estimated by fitting lines or planes (or "hyperplanes" for more than two explanatory variables) by least squares to the data points. Then, as we saw in Chapter 2, estimated mean values of response variables can be obtained by projecting case points onto variable vectors. This idea is developed and extended in two ways to generalized linear models: first, the mean of the response can be transformed nonlinearly, and it is this transformed mean that is modelled as a linear function of the predictors; and second, the probability distribution of the response variable around the mean function can be different from the normal distribution. In this chapter we first discuss data transformations and then give two examples of generalized linear model biplots: Poisson regression biplots and logistic regression biplots.

Contents

Data transformations

As an intermediate step towards considering *generalized linear models* (GLM), let us suppose that we wished to transform the response variables and then perform the regression analysis. There could be many reasons for this, such as making the data more symmetrically distributed, or reducing the effect of outliers. Power transformations are commonly used on data such as the species counts in Exhibit 2.1 —these can be either a square root or double square root (i.e., fourth root) transformation, or the Box-Cox family of power transformations which includes the logarithmic transformation. For example, considering species *d* again, let us con-

Exhibit 3.1:
The regression coefficients for the five regressions where in each case the response variable has been fourth root transformed. Overall variance explained is 33.9%

	Std devn	*Constant*	y^*	x^*	R^2
$a^{1/4}$	0.905	1.492	–0.672	0.073	60.5%
$b^{1/4}$	0.845	1.301	–0.506	0.006	36.2%
$c^{1/4}$	0.907	1.211	0.387	0.086	15.9%
$d^{1/4}$	0.602	1.639	–0.288	0.060	27.6%
$e^{1/4}$	0.755	0.815	–0.375	–0.255	22.8%

sider the fourth root transformation $d_0 = d^{1/4}$. Fitting this transformed response to the two standardized predictors y^* ("pollution") and x^* ("depth") as before leads to the following equation for predicting d_0:

$$\hat{d}_0 = 1.642 - 0.288\,y^* + 0.060\,x^* \qquad R^2 = 0.276 \qquad (3.1)$$

Notice first that we have not centred the transformed data, hence the presence of the constant in the regression—the constant is equal to the mean of d_0 because the predictor variables are centred. Also, because the power transformation tends to homogenize the variances (i.e., make them more similar in value), we have not standardized d_0 either. If some transformed variables have more variance, then we would like to preserve this fact.[3]

The complete set of results for all of the transformed responses is given in Exhibit 3.1.

The constants give the predicted value for mean values of y and x, when $y^* = x^* = 0$; for example, the average of $d^{1/4}$ is 1.639, which transforms back to a value of d of $1.639^4 = 7.216$.

Biplot with nonlinear calibration

The regression coefficients can again be used as biplot vectors, shown in Exhibit 3.2. The positions of the sample points are identical to the previous Exhibits 2.3 and 2.4. The difference between this biplot and the previous one for untransformed data (Exhibit 2.4) is that the regression surfaces (in the third dimension, "above" the biplot space) indicated by the biplot vectors are linear planes for the transformed variables, and thus nonlinear in terms of the original ones. So the calibration of the biplot axes in terms of the original variables is more complicated because the intervals between scale units are not constant.

[3] In this example the standard deviations of the original frequencies varied from 3.96 for species *e* to 12.6 for species *a*, while in the double square root transforms the range is from 0.602 for species *d* to 0.907 for species *c*. Notice that the ordering of the standard deviations is not necessarily preserved by the power transformation of the data, even though the transformation is monotonic.

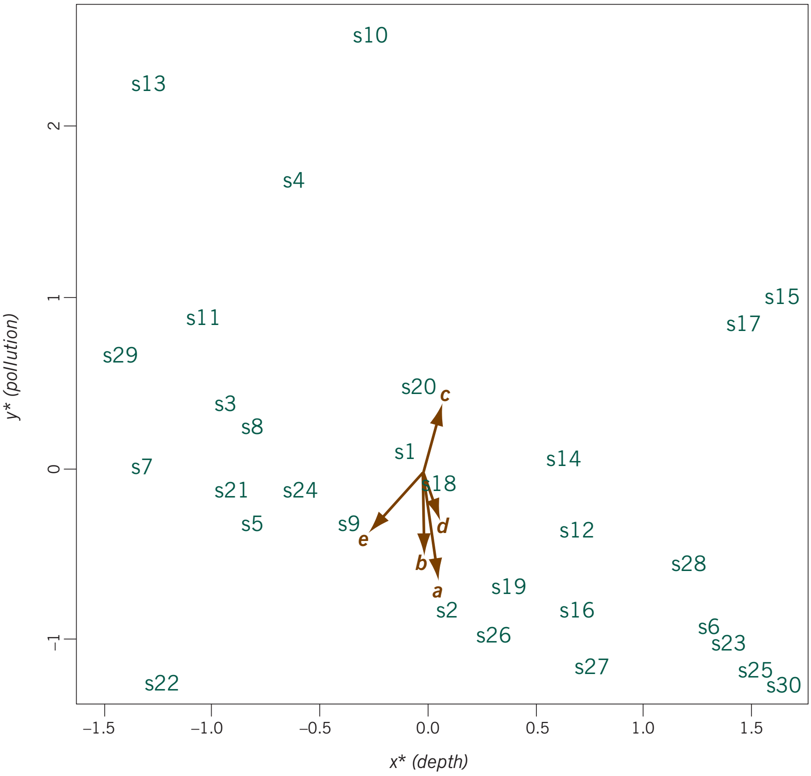


Exhibit 3.2:

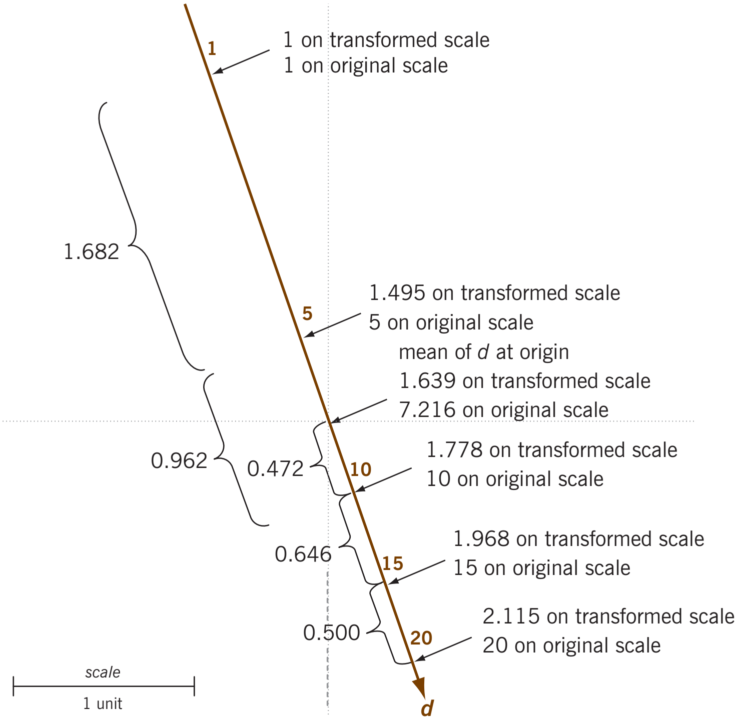
Biplot of the fourth root transformed species data, showing biplot vectors given by regression coefficients in Exhibit 3.1, i.e., the directions of planes corresponding to regressions of the transformed species variables on standardized "pollution" (y) and standardized "depth" (x*)*

For example, let us consider variable d again, and its transformed version $d_0 = d^{1/4}$. Exhibit 3.3 illustrates the calculations which we describe now. The centre of the map corresponds to the mean of the transformed variable, 1.639, which we already calculated above to correspond to 7.216 of the original variable, so we know that the value of d at the origin is 7.216 and that it is increasing in the direction of the arrow, i.e., downwards in the biplot.

The first "tic mark" we would put on the biplot axis through d is at the value 10, which on the transformed scale is $10^{1/4} = 1.778$. The difference between this value and the mean value of 1.639 is 0.139. A unit length on the axis is, as before, 1 divided by the length of the biplot vector: $1/\sqrt{0.288^2 + 0.060^2} = 3.399$, as shown in Exhibit 3.3. Hence, the distance along the biplot axis to put the tic mark is $0.139 \times 3.399 = 0.472$. The next tic mark at $d = 15$ corresponds to $15^{1/4} = 1.968$, a difference of $1.968 - 1.778 = 0.190$ from the position of $d = 10$, hence at a distance of $0.190 \times 3.399 = 0.646$ from the tic mark for 10. Going in the other direction, to put a tic mark for $d = 5$, the transformed value is $5^{1/4} = 1.495$, a difference relative to the position of $d = 10$ of $1.778 - 1.495 = 0.283$, or $0.283 \times 3.399 = 0.962$ units away from the tic mark for 10 in the negative direction for d (i.e., upwards

Exhibit 3.3:

*Nonlinear calibration of biplot axis through reponse variable **d**. Because **d** has been fourth root transformed, the calibrations are not at regular intervals*



in Exhibit 3.2). The tic mark for 1 (same on original and transformed scales) is 0.495 transformed units away from 5, and so 0.495 × 3.399 = 1.682 units away from 5 on the biplot axis. The nonlinearity of the calibration on the biplot axis is clear in Exhibit 3.3, with the tic marks getting closer together as *d* increases. The contours are still perpendicular to the biplot axis, so the interpretation is still based on projecting the biplot points onto the biplot axes, bearing in mind the contracting scale due to the transformation.

The fourth root transformation of the response variable is a *monotonically increasing* function, hence the calibration of the biplot axis shows values increasing in the direction of the biplot vector, albeit increasing in a nonlinear fashion. Although seldom done, a non-monotonic transformation, for example a quadratic transformation which rises and then falls, could also be applied to the response variable. The effect would be that the calibrations on the biplot axis would increase and then decrease again.

Poisson regression biplots

The regression biplots in Chapter 2 and those described above with transformed responses use regression coefficients for the biplot vectors that have been obtained using least-squares fitting of the response variable, with or with-

	Constant	y^*	x^*	*Error*
$\log(\bar{a})$	2.179	–1.125	–0.067	0.388
$\log(\bar{b})$	1.853	–0.812	0.183	0.540
$\log(\bar{c})$	2.041	0.417	0.053	0.831
$\log(\bar{d})$	2.296	–0.337	0.199	0.614
$\log(\bar{e})$	0.828	–0.823	–0.568	0.714

Exhibit 3.4: *The regression coefficients for the five Poisson regressions of the species responses on the predictors "pollution" y^* and "depth" x^*. Rather than variance explained, the "error" of the model fit is reported as the deviance of the solution relative to the null deviance when there no predictors (so low values mean good fit)*

out transformation, to the explanatory variables. This idea originates in the assumption that, conditional on the explanatory variables, the distribution of the response variable in the population is normal, with the conditional mean of the response equal to a linear function of the explanatory variables. In generalized linear modelling this idea is extended to different distributions, but in every case some transformation of the mean of the response, called the *link function*, is modelled as a linear function of the explanatory variables. Linear regression is the simplest example of a *generalized linear model* (GLM) where there is no transformation of the mean (i.e., the link function is the identity function) and the conditional distribution is normal. The coefficients of a GLM are estimated using the principle of maximum likelihood, which has as a special case the least-squares procedure when the assumed distribution is normal.

The first example of a "non-normal" GLM that we consider is *Poisson regression*. Since the species variables *a* to *e* are counts, a more appropriate distribution would be the Poisson distribution. In Poisson regression the link function is logarithmic, so the model postulates that the logarithm of the response mean is a linear function of the explanatory variables, and the assumed conditional distribution of the response is Poisson. Fitting this model for each of the five responses is just as easy in R as fitting regular regression, using the `glm` function (see the Computational Appendix) and the estimated coefficients are given in Exhibit 3.4. Notice that the way the success of the model fit is measured is the opposite here, in the sense that for good fit the "error" should be low. In the simple regression case, subtracting the "error" from 1 would give R^2.

Notice the difference between the GLM and what we did before: we have not log-transformed the original data, which would have been a problem since there are zeros in the data, but have rather modelled the logarithm of the (conditional) mean as a linear function of the explanatory variables. For example, in the case of species *d* the model estimates are given by:

$$\log(\bar{d}) = 2.296 - 0.337y^* + 0.199x^* \tag{3.2}$$

Exponentiating both sides, this gives the equation:

$$\bar{d} = \exp(2.296) \cdot \exp(-0.337y^*) \cdot \exp(0.199x^*) \tag{3.3}$$

so that the exponentials of the coefficients –0.337 and 0.199 model the multiplicative effect on the estimated mean response: a one (standard deviation) unit increase in y^* multiplies $\bar{d}$ by $\exp(-0.337) = 0.714$ (a 28.6% decrease), while a one unit increase in x^* multiplies $\bar{d}$ by $\exp(0.199) = 1.221$ (a 22.1% decrease). Again, the coefficients define a biplot vector in the space of the explanatory variables. To calibrate the vector, the value at the origin corresponds to the value of the mean response at the means of the explanatory variables $y^* = x^* = 0$, that is $\bar{d} = \exp(2.296) = 9.934$. To place tic marks on the biplot axis we would again calculate what a unit length on the axis is: $1/\sqrt{0.337^2 + 0.199^2} = 2.556$, which corresponds to a unit on the logarithmic scale. Using this we can work out where tic marks should be placed for values of $\bar{d}$ such as 0, 5, 10, 15, etc.—this will be a logarithmic scale on the biplot axis, with intervals between tic marks contracting as the response variable increases. We do not show the Poisson biplot here, but it can be computed using the script in the Computational Appendix.

Logistic regression biplots

Let us suppose, as is indeed common in ecological research, that we are interested more in the presence/absence of species than their actual abundances; that is, we replace all positive counts by 1 and leave the zeros as 0. The mean of 0/1 data is the probability p of a presence (i.e., a 1), so we write p_a, p_b, …, p_e, for the probabilities of the five species presence/absence variables. Logistic regression can be used to predict the dichotomous presence/absence response variables, given the explanatory variables. This is another GLM where the assumed distribution of the 0/1 data is binomial and the link function is the *logit*, or *log-odds*, function. The logit function is $\log(p/(1-p))$, abbreviated as $\text{logit}(p)$. Again, the fitting of this GLM is a simple option of the R `glm` function (see the Computational Appendix) and the estimated coefficients are listed in Exhibit 3.5.

Using species *d* once more as an example, the estimating equation is:

$$\text{logit}(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = 2.712 - 1.177y^* - 0.137x^* \tag{3.4}$$

and the coefficients –1.177 and –0.137 estimate the changes in the log-odds of the probability of species *d*. Using the coefficients we can again make a biplot of the five species in the space of y^* and x^*, shown in Exhibit 3.6. This could be calibrated in units of odds, $p_d/(1 - p_d)$, or transformed back to units of p_d as follows, thanks to the inverse transformation:

Exhibit 3.5:
The regression coefficients for the five logistic regressions of the species responses on the predictors "pollution" y^* and "depth" x^*, showing their error deviances

	Constant	y^*	x^*	*Error*
logit(p_a)	2.384	–2.889	0.863	0.464
logit(p_b)	1.273	–1.418	–0.143	0.756
logit(p_c)	0.831	0.973	0.315	0.911
logit(p_d)	2.712	–1.177	–0.137	0.798
logit(p_e)	0.253	–1.280	–0.786	0.832

$$p_d = \frac{\exp(2.712 - 1.177y^* - 0.137x^*)}{1 + \exp(2.712 - 1.177y^* - 0.137x^*)} \tag{3.5}$$

So for $y^* = x^* = 0$, exp(2.712) = 15.06 and the estimated probability of d is 15.06/16.06 = 0.938 (from Exhibit 2.1 species d occurs at 27 out of 30 sites, so its probability of a presence is high, but comes down mainly when y^* increases). So the origin of the map corresponds to an estimated p_d of 0.938. Where would the

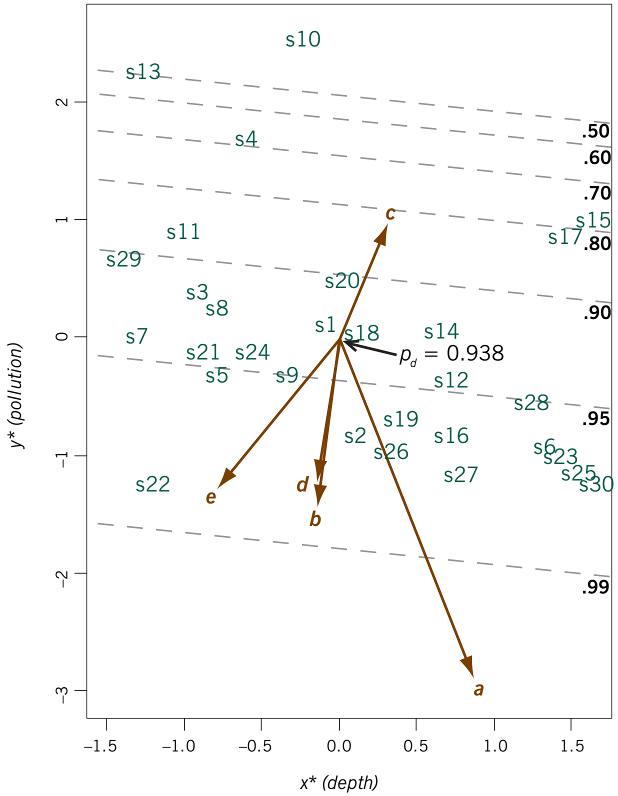


Exhibit 3.6:
*Logistic regression biplot of the presence/absence data of the five species. The calibration for species **d** is shown in the form of contours in units of predicted probability of presence. The scale is linear on the logit scale but non-linear on the probability scale, as shown*

tic mark be placed for 0.95? The corresponding logit is log(0.95/0.05) = 2.944, which is 2.944 – 2.712 = 0.232 units higher on the logit scale from the origin. The unit length is once more the inverse of the length of the biplot vector $1/\sqrt{1.177^2 + 0.137^2} = 0.844$, so the tic mark for 0.95 is at a distance 0.232 × 0.844 = 0.196 from the origin in the positive direction of *d*. Exhibit 3.6 shows the logistic regression biplot with the contours of the probability function for species *d*.

In a similar way the logistic regression surfaces could be indicated for each of the other species as a sequence of probability contour lines at right angles to the biplot vectors in Exhibit 3.6, where the origin corresponds to the probability for the means of the explanatory variables and the probability contours increase in the direction of the respective biplot vectors.

SUMMARY: Generalized Linear Model Biplots

1. A regression biplot can still be made if a nonlinear transformation of the response variable is performed: the effect is that the tic marks on the biplot axes are not at equal intervals, that is, the calibration is nonlinear.
2. *Generalized linear models* generalize linear regression to include different relationships between the conditional mean of the response variable and the explanatory variables as well as different distributions for the response variable. In each generalized linear model the conditional mean, transformed by the *link function*, is modelled as a linear function of the explanatory variables.
3. Examples of generalized linear models are *Poisson regression* (for count data), where the link function is the logarithm and the assumed distribution is Poisson; and *logistic regression* (for discrete responses), where the link function is the logit and the assumed distribution is binomial.