Communications for Statistical Applications and Methods
2013, Vol. 20, No. 2, 129–136

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.2.129

Biplots of Multivariate Data Guided by Linear and/or Logistic Regression

Myung-Hoe Huh[1,a], Yonggoo Lee[b]

[a]Department of Statistics, Korea University
[b]Department of Applied Statistics, Chung-Ang University

Abstract

Linear regression is the most basic statistical model for exploring the relationship between a numerical response variable and several explanatory variables. Logistic regression secures the role of linear regression for the dichotomous response variable. In this paper, we propose a biplot-type display of the multivariate data guided by the linear regression and/or the logistic regression. The figures show the directional flow of the response variable as well as the interrelationship of explanatory variables.

Keywords: Data visualization, biplot graph, linear regression, logistic regression, dimensional reduction.

1. Background and Aim

For the graphical display of $n \times p$ multivariate data $\mathbf{X}$, the biplot initiated by Gabriel (1971) maps n observations in a lower-rank linear space and represents p variables in the same graph. There are several versions of biplots, among which the canonical version is the principal components (PC) biplot. In the two-dimensional PC biplot, n observations and p variables are dotted respectively by the rows of $\mathbf{X}\mathbf{V}_{(2)}$ and the rows of $\mathbf{V}_{(2)}$, where $\mathbf{V}$ is $p \times p$ matrix consisting of all eigenvectors of $\mathbf{X}'\mathbf{X}$ and $\mathbf{V}_{(2)}$ is the $p \times 2$ submatrix of $\mathbf{V}$ corresponding to two principal eigenvalues. PC biplot is very useful in exploring the pattern in observations and the interrelationship among variables. Details of the methodology can be found at Lebart *et al.* (1984) and Huh (2011a). Several extensions and modern formulations can be found at Gower and Hand (1996) and Greenacre (2010).

In this study, we consider the data sets of n observations of one response variable and $p\,(\geq 2)$ explanatory variables. Response variable could be numerical or dichotomous. We suppose the data is modeled by either linear regression or logistic regression, according to the data type of the response variable. The aim of the paper is to build a biplot of the $n \times p$ dataset of explanatory observations guided by the linear regression and/or the logistic regression. Thus the graphs proposed in this paper can be regarded as a supervised biplot.

2. Biplot with Linear Regression

For the numerical response variable Y, we suppose that p numerical (or possibly dummy) variables $X_1, \ldots, X_p$ are put for explanatory purpose in a linear regression model. Assuming all the variables are standardized with means zero and standard deviations 1, we write the fitted model as

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_p X_p,$$

[1] Corresponding author: Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr

where $b_1, \ldots, b_p$ are regression coefficients of p explanatory variables. Hence

$$\mathbf{v}^{[1]} = \frac{\mathbf{b}}{\|\mathbf{b}\|}, \quad \text{for } \mathbf{b} = (b_1, \ldots, b_p)^t$$

is the unit vector indicating the flow of Y in increasing direction as modeled by a linear regression. Even though $b_0 = 0$ when the model is fitted by least squares (LS), we will retain b_0 to deal with alternative fits of the model.

Denoting $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for n observations of $(X_1, \ldots, X_p)$, we consider the projections of $\mathbf{x}_i$ onto $\mathbf{v}^{[1]}$ for $i = 1, \ldots, n$. From the process, we obtain the secondary component $\mathbf{x}_i^{[2]}$ of $\mathbf{x}_i$ that is orthogonal to $\mathbf{v}^{[1]}$.

The collection of all secondary components $\mathbf{x}_1^{[2]}, \ldots, \mathbf{x}_n^{[2]}$ can be portrayed by their projections onto a unit directional vector, say $\mathbf{v}$. In ordinary cases, the best choice of $\mathbf{v}$ is determined by maximizing the total squared lengths of the projections for the largest complimentary spread. That is,

$$\max_{\mathbf{v}} \sum_{i=1}^{n} \left\| \mathbf{x}_i^{[2]} - \left(\mathbf{v}^t \mathbf{x}_i^{[2]}\right) \mathbf{v} \right\|^2 \quad \text{subject to } \mathbf{v}^t\mathbf{v} = 1.$$

Then, the formulation is exactly same as the principal component analysis of

$$\mathbf{X}^{[2]} = \begin{pmatrix} \mathbf{x}_1^{[2]t} \\ \vdots \\ \mathbf{x}_n^{[2]t} \end{pmatrix}$$

which is an $n \times p$ data matrix, of which each row vector is orthogonal to $\mathbf{v}^{[1]}$. Therefore, the optimal $\mathbf{v}$, denoted by $\mathbf{v}^{[2]}$, is the principal eigenvector of $\mathbf{X}^{[2]t}\mathbf{X}^{[2]}$. Clearly, $\mathbf{v}^{[2]}$ is orthogonal to $\mathbf{v}^{[1]}$.

We propose a biplot on the two-dimensional plane for the linear regression case:

$$\text{For the observations, plot } \left(\mathbf{x}_i^t\mathbf{v}^{[1]}, \mathbf{x}_i^t\mathbf{v}^{[2]}\right), \quad \text{for } i = 1, \ldots, n.$$
$$\text{For the variables, plot } c\left(v_j^{[1]}, v_j^{[2]}\right), \quad \text{for } j = 1, \ldots, p.$$

On the above line, $v_j^{[k]}$ is the j^{th} component of $\mathbf{v}^{[k]}$ and c is a constant such as 1, 2, or 3, suitably chosen not to make the graph over-crowded. In the figures, we set $c = 3$ and the vertical axis as the first dimension and the horizontal axis as the second dimension, to imply that the response (or the response variable) increases along the vertical axis. Obviously, we can extend the plot to be displayed on three or more dimensional plane.

As a numerical illustration, consider the stack loss data (Brownlee, 1965) of which the response variable is the loss of ammonia (= Y) and the explanatory variables are air flow (= X_1), water temperature (= X_2) and acid concentration (= X_3). The number of observations is 21 (= n).

The LS fitted linear regression is

$$\hat{Y} = 0.645X_1 + 0.403X_2 - 0.080X_3$$

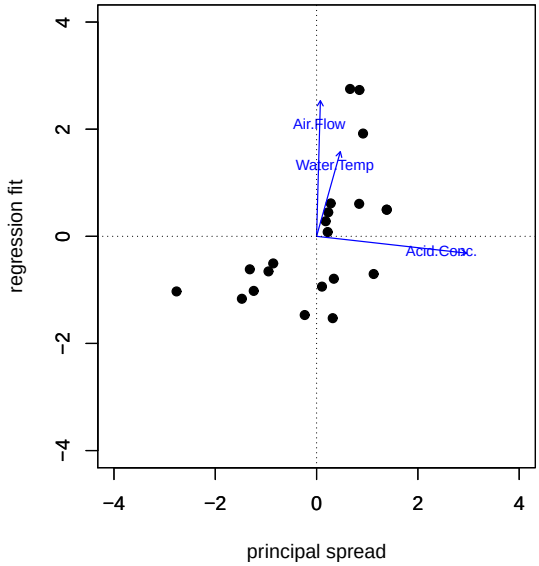

Figure 1: *A biplot of the stack loss data, with the linear regression fit along the vertical axis*

on the standardized variables. Thus first directional vector in the explanatory space is

$$\mathbf{v}^{[1]} = \frac{\begin{pmatrix} 0.645 \\ 0.403 \\ -0.08 \end{pmatrix}}{\left\| \begin{matrix} 0.645 \\ 0.403 \\ -0.08 \end{matrix} \right\|} = \begin{pmatrix} 0.844 \\ 0.526 \\ -0.105 \end{pmatrix}.$$

The second directional vector in the explanatory space is obtained by eigen-decomposing $\mathbf{X}^{[2]t}\mathbf{X}^{[2]}$, where

$$\mathbf{X}^{[2]} = \mathbf{X} - \mathbf{X}\,\mathbf{v}^{[1]}\mathbf{v}^{[1]t}.$$

The principal eigenvalue of $\mathbf{X}^{[2]t}\mathbf{X}^{[2]}$ occupies 87% of total eigenvalues. Corresponding eigenvector is given as

$$\mathbf{v}^{[2]} = (0.118, 0.009, 0.993)^t.$$

With these two directional vectors $\mathbf{v}^{[1]}$ and $\mathbf{v}^{[2]}$, the two-dimensional biplot is produced. See Figure 1.

In the figure, we see that the regression fit is determined primarily by the air flow (= X_1) and secondly by water temperature (= X_2). The influence of acid concentration (= X_3) is negative but apparently weak. The plot shows that observations spread to equal degree along the vertical axis (= regression fit) and the horizontal axis (= principal spread). Major determinant of the horizontal axis is the acid concentration (= X_3).

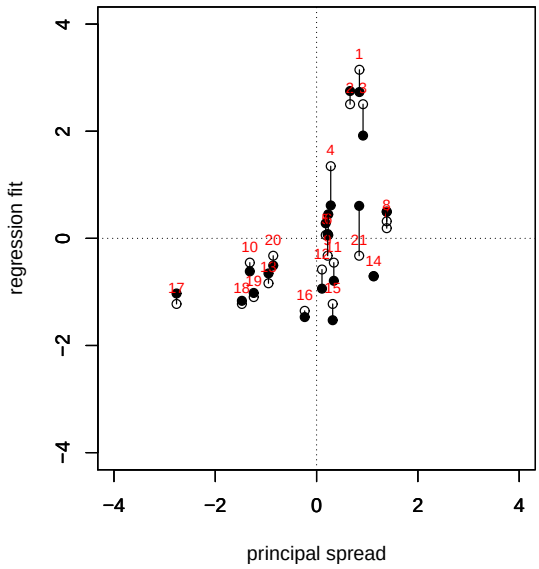

Figure 2: *A biplot of the stack loss data, with the linear regression fit and the observed response values along the vertical axis*

In the proposed plot, we may add the observed values of Y by noting that the vertical coordinate for an arbitrary case $\mathbf{x}$ is given by

$$\mathbf{x}^t\mathbf{v}^{[1]} = \frac{\mathbf{x}^t\mathbf{b}}{\|\mathbf{b}\|} = \frac{\hat{Y} - b_0}{\|\mathbf{b}\|}.$$

Hence, the case observation Y can be marked at $(Y - b_0)/\|\mathbf{b}\|$. See Figure 2. In the figure, open circles represent observed response values. We see that the case 21 has a conspicuously large size of negative residual and that the case 4 has a large size of positive residual.

3. Biplot with the Logistic Regression

For the dichotomous response variable $Y(= 0, 1)$, we suppose that p numerical variables $X_1, \ldots, X_p$ are put for explanatory purpose in the logistic regression model. Assuming all explanatory variables are standardized with means zero and standard deviations 1, we write the fitted model as

$$\log \frac{\hat{P}}{1 - \hat{P}} = b_0 + b_1X_1 + \cdots + b_pX_p,$$

where $\hat{P} = \hat{P}\{Y = 1|X_1, \ldots, X_p\}$ and $b_1, \ldots, b_p$ are regression coefficients of p explanatory variables. Hence

$$\mathbf{v}^{[1]} = \frac{\mathbf{b}}{\|\mathbf{b}\|}, \quad \text{for } \mathbf{b} = \left(b_1, \ldots, b_p\right)^t$$

is the unit vector indicating the direction of increasing $\hat{P}$ as modeled by the logistic regression. Hence the building elements of the graphics are exactly same as those of the linear regression case.

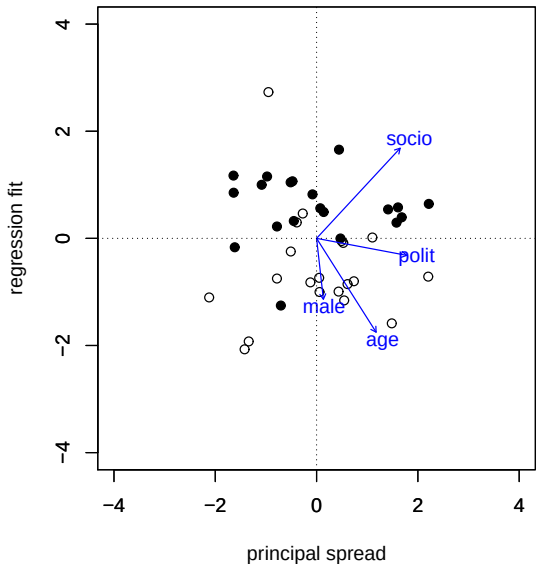

Figure 3: *A biplot of the magazine data, with the logistic regression fit along the vertical axis*

As a numerical illustration, we consider the "magazine data" of which the response variable is the intention to subscribe a particular magazine (= Y) and the explanatory variables are gender (= X_1), age (= X_2), social affinity (= X_3), and political propensity (= X_4). X_1 is dummy (1 = male, 0 = female), and X_2 to X_4 are numerical (Huh, 2011b). Number of cases is 40 (= n).

The biplot is shown in Figure 3, of which the filled/open circles represent the cases with/without the intention to subscribe the magazine. In the plot, we see that the social affinity and the age are two major determinants of Y and that the males are less likely to respond positively compared to females. Influence of the political propensity is weakly negative.

Horizontal axis line does not necessarily indicate the fitted probability 50%. Rather, it represents the probability level $\exp(b_0)/\{1 + \exp(b_0)\}$ (= $\hat{P}_0$). In the numerical example, $b_0 = -0.019$ and, thus, $\hat{P}_0 = 0.495$.

We may add contour lines for specified probability levels for $0 < \hat{P} < 1$, noting that the covariate $\mathbf{x}$ associated to $\hat{P}$ is dotted on the vertical axis at

$$\mathbf{x}^t\mathbf{v}^{[1]} = \frac{\mathbf{x}^t\mathbf{b}}{\|\mathbf{b}\|} = \frac{1}{\|\mathbf{b}\|}\left(\log\frac{\hat{P}}{1-\hat{P}} - b_0\right).$$

Figure 4 shows the contour levels for $\hat{P} = 0.05, 0.25, 0.5, 0.75, 0.95$ for the magazine data. The case 32 with no intention ($Y = 0$) is rather exceptional, since it is predicted that the case is very likely to subscribe the magazine ($Y = 1$) with probability larger than 95%.

4. Subset Regression

Sometimes, we select a subset of all available explanatory variables to be included in the regression model by various methods. Then, having a subset of unselected variables at hand, we may be interested in how they are related to the selected variables. Unselected variables could be related to

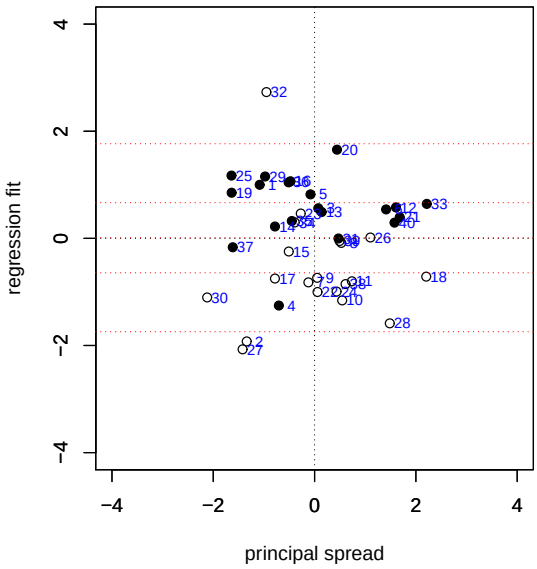

Figure 4: *A biplot of the magazine data, with the logistic regression fit and the predicted probability contours along the vertical axis*

the response variable, but they are masked by selected ones. Or, they are not related to the response variable nor to the selected explanatory variables.

To study the interrelationship among all explanatory variables in the setting of subset regression, we propose a projection scheme of p explanatory vectors of length n on the factor score vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ derived from the subset of k explanatory variables, where

$$\mathbf{u}_1 = \sqrt{n-1}\,\frac{\mathbf{X}_{(k)}\mathbf{v}^{[1]}}{\left\|\mathbf{X}_{(k)}\mathbf{v}^{[1]}\right\|}, \qquad \mathbf{u}_2 = \sqrt{n-1}\,\frac{\mathbf{X}_{(k)}\mathbf{v}^{[2]}}{\left\|\mathbf{X}_{(k)}\mathbf{v}^{[2]}\right\|},$$

denoting that $\mathbf{X}_{(k)}$ is the $n \times k$ submatrix of $\mathbf{X}$ with k selected variables, $\mathbf{v}^{[1]}$ is the $k \times 1$ unit vector derived from the subset regression coefficients, $\mathbf{v}^{[2]}$ is the $k \times 1$ unit vector determining the primary principal component of complementary explanatory vectors after the regression. Even though $\mathbf{v}^{[1]}$ and $\mathbf{v}^{[2]}$ are orthogonal, $\mathbf{u}_1$ and $\mathbf{u}_2$ may not be so. Thus, we rewrite $\mathbf{u}_2$ by

$$\sqrt{n-1}\left(\mathbf{u}_2 - \frac{\mathbf{u}_1^t\mathbf{u}_2}{n-1}\mathbf{u}_1\right)\Bigg/\left\|\left(\mathbf{u}_2 - \frac{\mathbf{u}_1^t\mathbf{u}_2}{n-1}\mathbf{u}_1\right)\right\|,$$

to secure the orthogonality of $\mathbf{u}_1$ and $\mathbf{u}_2$. Then, we project all p explanatory vectors of length n onto the linear space spanned by $\mathbf{u}_1$ and $\mathbf{u}_2$, to obtain the map of both selected and unselected explanatory variables.

Now, we illustrate our procedure with the aerobic fitness data (SAS Inc., 2009). Being measured from thirty one males, the response variable is the oxygen uptake rate (= Y), that could be explained by six variables: age (= X_1), running time (= X_2), run pulse (= X_3), weight (= X_4), max pulse (= X_5), and rest pulse (= X_6). We suppose that the data modeler decided to include X_1, X_2 and X_3 in the linear regression ($p = 6, k = 3$).

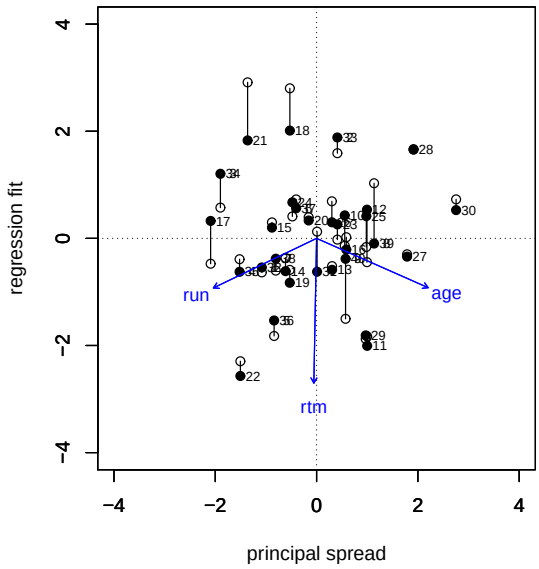

Figure 5: *A biplot of the aerobic fitness data, with the linear regression fit by three explanatory variables along the vertical axis*

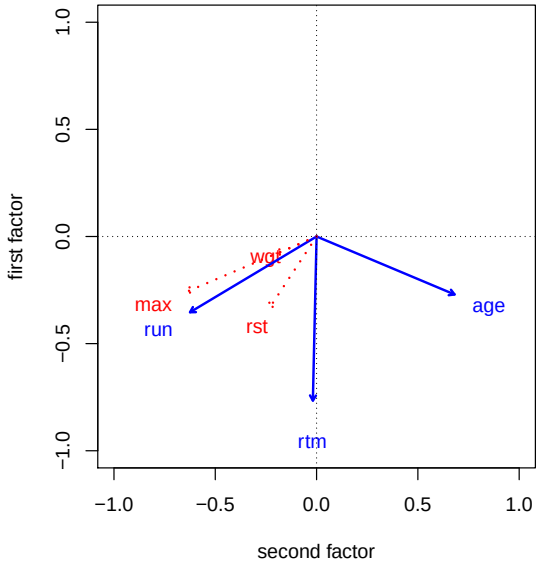

Figure 6: *Selected and unselected explanatory variables of the aerobic fitness data*

Figure 5 is the biplot guided by the subset regression. Vertical axis indicates that the response variable (oxygen uptake rate) is determined firstly by (negative) “rtm” (runtime), secondly by (negative) “age” and (negative) “run” pulse. Horizontal axis shows that the observation units are widely spread by “age” and “run” pulse.

Figure 6 visualizes the interrelationship among selected and unselected variables. We find that 1) the "wgt" (weight) is not related to all the variables of the model including the response variable, and that 2) the "max" pulse and the "rst" (rest) pulse are represented by the run pulse fairly well.

5. Concluding Remarks

Proposed graphical methods can be applied to visualize the datasets guided by any statistical model based on the linear combination of explanatory variables. For example, the models could be generalized linear models other than the logistic regression and/or the linear support vector machines. Huh and Park (2009) suggested the latter case.

Since proposed methods use a dimensional reduction tool to visualize the multidimensional data, usefulness of the output picture could be limited for the cases in which the number of explanatory variables is large, *i.e.* $p \geq 10$. In such cases, one may pursue further exploration of $\mathbf{X}^{[2]}$, the explanatory data matrix after linear/logistic regression, with dynamic graphical methods.

References

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, Second Edition, Wiley, New York.

Gabriel, K. R. (1971). The biplot display of matrices with the application to principal component analysis, *Biometrika*, **58**, 453–467.

Gower, J. C. and Hand, D. J. (1996). *Biplots*, Chapman and Hall, London.

Greenacre, M. (2010). *Biplots in Practice*, BBVA Foundation, Madrid.

Huh, M. H. (2011a). *Exploratory Multivariate Data Analysis*, Freedom Academy, Seoul.

Huh, M. H. (2011b). *Statistical Concepts, Methods and Applications Using R*, Freedom Academy, Seoul.

Huh, M. H. and Park, H. M. (2009). Visualizing SVM classification in reduced dimensions, *Communications of the Korean Statistical Society*, **16**, 881–889.

Lebart, L., Morineau, A. and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New York.

SAS Inc. (2009). *SAS/STAT V9.2 Users Guide*, Second Edition. NC: Cary.

Received February 6, 2013; Revised March 11, 2013; Accepted March 12, 2013