

Regression Analysis of Compositional Data When Both the Dependent Variable and Independent Variable Are Components

L.A. van der Ark¹

¹Tilburg University, The Netherlands; a.vdark@uvt.nl

Abstract

It is well known that regression analyses involving compositional data need special attention because the data are not of full rank. For a regression analysis where both the dependent and independent variable are components we propose a transformation of the components emphasizing their role as dependent and independent variables. A simple linear regression can be performed on the transformed components. The regression line can be depicted in a ternary diagram facilitating the interpretation of the analysis in terms of components. An example with time-budgets illustrates the method and the graphical features.

Key words: compositional data, regression analysis, time budgets, visualization.

1 Introduction

Compositional data are collected in an $N \times D$ two-way table \mathbf{X} , such that its elements, x_{ri} ($r = 1, \dots, N; i = 1, \dots, D$) are nonnegative and the rows add up to 1 (see, e.g., Aitchison, 1986/2003, p. 1). The latter constrained is often referred to as the *sum constrained*. The rows of \mathbf{X} , $\mathbf{x}_1, \dots, \mathbf{x}_N$ are called compositions. In the social sciences the rows of \mathbf{X} are often called *budgets* (see, e.g., De Leeuw and others, 1990; Van der Heijden and others 1992; Van der Ark and others, 1999). Sometimes, compositional data are formed by the *closure* of an $N \times D$ two-way table \mathbf{W} with nonnegative elements. Closure means that all elements are divided by the row sum; that is, $x_{ri} = w_{ri} / \sum_j w_{rj}$.

Table 1 shows an (fictitious) example of so-called time-budgets. Students were asked to indicate how they would divide a weeks time in in six mutually exclusive and exhaustive components: studying (S), paid work (PW), domestic work (DW), personal needs (PN), social past time (SP) other activities (AO). The question of interest in this paper is how the time spent on paid work is related to the time spent on studying. It is assumed that if students students spend a large part of their available time on paid work, they will spend relatively more time on studying in the remaining period than students that spend little time on paid work.

2 A regression model

A regression analysis is the most natural way to investigate the relation between paid work and studying. However, an ordinary regression analysis,

$$S = \beta_0 + \beta_1 \times PW,$$

is problematic. First, it yields estimates for S that are not range preserving and may violate the nonnegativity constraints. Second, coefficient β_1 is related to the correlation coefficient, r ,

$$\beta_1 = r_{S,PW} \times \frac{\text{SD}(S)}{\text{SD}(PW)}$$

Table 1: Example of time-budgets.

Respondent	Activities						Total
	S	PW	DW	PN	SP	AO	
Respondent 1	0.12	0.18	0.01	0.33	0.12	0.24	1.00
Respondent 2	0.14	0.06	0.02	0.36	0.09	0.33	1.00
Respondent 3	0.20	0.00	0.02	0.30	0.08	0.40	1.00
Respondent 4	0.14	0.10	0.05	0.30	0.12	0.30	1.00
Respondent 5	0.31	0.12	0.01	0.36	0.06	0.15	1.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note: S = studying; PW = paid work;
 DW = domestic work; PN = personal needs;
 SP = social past time; O = other activities

It is well known that the correlational structure of compositional data is problematic (see Aitchison, 1985/2003, pp. 52-58). For example, $\beta_1 = 0$, which implies that $r_{S,PW} = 0$, does not imply linear independence.

Transforming the compositional data to unconstrained data before the regression analysis may solve the problem. Well-known transformations are the *logratio transformation* (Aitchison, 1985/2003, p. 77) where

$$y_{ri} = \log \frac{x_{ri}}{rD}, \text{ for } r = 1, \dots, N; i = 1, \dots, D - 1$$

and the *centered logratio transformation*. Let $g(\mathbf{x}_i)$ denote the geometric mean of budget \mathbf{x}_i , then the centered logratio transformation is

$$z_{ri} = \log \frac{x_{ri}}{g(\mathbf{x}_i)}, \text{ for } r = 1, \dots, N; i = 1, \dots, D.$$

We argue that all transformations are arbitrary in the sense that no transformation can give back the information that was lost in closure of the data. Hence, the underlying data matrix \mathbf{W} can not be recovered. Therefore, the mathematical and substantive properties of the transformed data should be decisive in choosing a transformation.

For a suitable transformation for the situation that both the dependent and independent variable are components, the research question needs to be investigated. We are interested in how students spend the remainder of their time when the paid work is done. Our special interest is which part of the remaining time, consisting of S , DW , PN , SP , and AO , is spent on studying. If the components DW , PN , SP , and AO are added into one large component *other* O , the dependent variable of this research would be (the logarithm of) the ratio of S and O ; hence $S^* = \log\left(\frac{S}{O}\right)$. We call S the *primary dependent component*, O the *reference dependent component*, and the ratio of S and O the *dependent ratio*. The independent component is the time spent on paid work. A so-called logit transformation (see, e.g., Agresti, 2002), $PW^* = \log\left(\frac{PW}{1-PW}\right)$, makes sure that the domain of the independent variable is the real line. Then, the proposed regression equation is

$$S^* = \beta_0 + \beta_1 \times PW^* + \text{error}. \tag{1}$$

The expected value of the dependent variable, denoted \hat{S}^* equals

$$\hat{S}^* = \beta_0 + \beta_1 \times PW^*. \tag{2}$$

After β_0 and β_1 have been estimated, the primary dependent component and the reference dependent component can be computed from Equation 2. First take the inverse natural logarithm of

Equation 2 which yields the estimated dependent ratio

$$\left(\frac{\widehat{S}}{\widehat{O}}\right) = \exp \left[\beta_0 + \beta_1 \times \log \left(\frac{PW}{1 - PW} \right) \right]. \quad (3)$$

Hence, the part of the time budget without PW (i.e., $1 - PW$), is spent on S and O according to the dependent ratio in Equation 3. The part of $1 - PW$ that is spent on studying is

$$\hat{S} = \frac{\left(\frac{\widehat{S}}{\widehat{O}}\right)}{1 + \left(\frac{\widehat{S}}{\widehat{O}}\right)} \times (1 - PW)$$

and the part that is spent on the remaining activities is

$$\hat{O} = \frac{1}{1 + \left(\frac{\widehat{S}}{\widehat{O}}\right)} \times (1 - PW).$$

It may be noted that this procedure is range preserving and that for each respondent $\hat{S} + \hat{O} + PW = 1$.

3 Graphic features on the simplex

The proposed regression can be depicted in ternary diagrams. A ternary diagram (see, e.g., Aitchison, 1985/2003, pp. 5-6; Van der Ark and Van der Heijden, 1998; Magidson and Vermunt, 2001) is a triangle used to depict budgets with three components. This visualization of the regression facilitates the interpretation. Each vertex of the triangle represents one component. Figure 1 is a ternary diagram where the upper vertex represent the independent component (PW), the right-hand vertex represents the primary dependent component (S), and the left-hand vertex represents the reference dependent component (O). In a time budget depicted in the upper vertex all time is spent on paid work; in a time budget depicted in the right-hand vertex all time is spent on studying. Budgets depicted in the triangle are mixtures of the three components. The time budgets of Table 1 are displayed in Figure 1. Note that the black square which is most to the right represents the budget of respondent 5 who spends more time on studying than the other respondents, the black square on the edge of the ternary diagram represents the budget of respondent 3, who spends no time at all on paid work.

The estimated regression curves in Equation 2 can also be depicted in a ternary diagram. Figure 2 shows the regression curves of for several values of β_0 and β_1 .

It may be noted that all regression curves start at the lower edge of the ternary diagram and end in the upper vertex. The lowest edge represents the dependent ratio, when the value of the independent component equals zero; that is, the ratio of time spent on studying and time spent on the other components when no time is spent on paid work. The shape of the curves depends on the values of β_0 and β_1 , but the interpretation of the parameters is not as straightforward as in linear regression. Therefore, visualizing the regression curves may help understand the results. Figure 2 shows the following. If $\beta_1 = 0$ (Fig. 2D, Fig. 2E, Fig. 2F) the dependent ratio does not change as the independent component increases. If $\beta_1 < 0$ (Fig. 2A, Fig. 2B, Fig. 2C) the dependent ratio decreases as the independent component increases and if $\beta_1 > 0$ (Fig. 2G, Fig. 2H, Fig. 2I) the dependent ratio increases as the independent component increases. As β_1 becomes closer to 0 (not shown in Fig. 2), the change in the dependent ratio as a function of the independent component becomes more gradual.

Coefficient β_0 determines at what value of the independent component the changes in the dependent ratio occur. If $\beta_0 < 0$ (Fig. 2A, Fig. 2C, Fig. 2G) the change in the dependent ratio, expressed

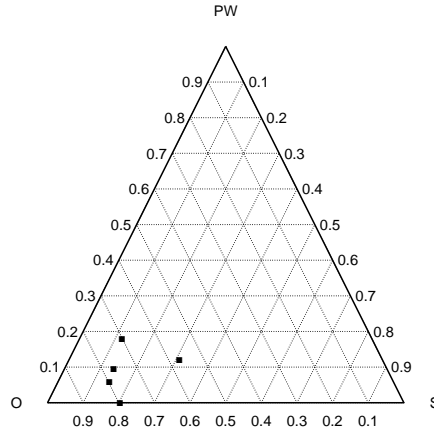


Figure 1: Graphic display of the time-budgets of the respondents in Table 1.

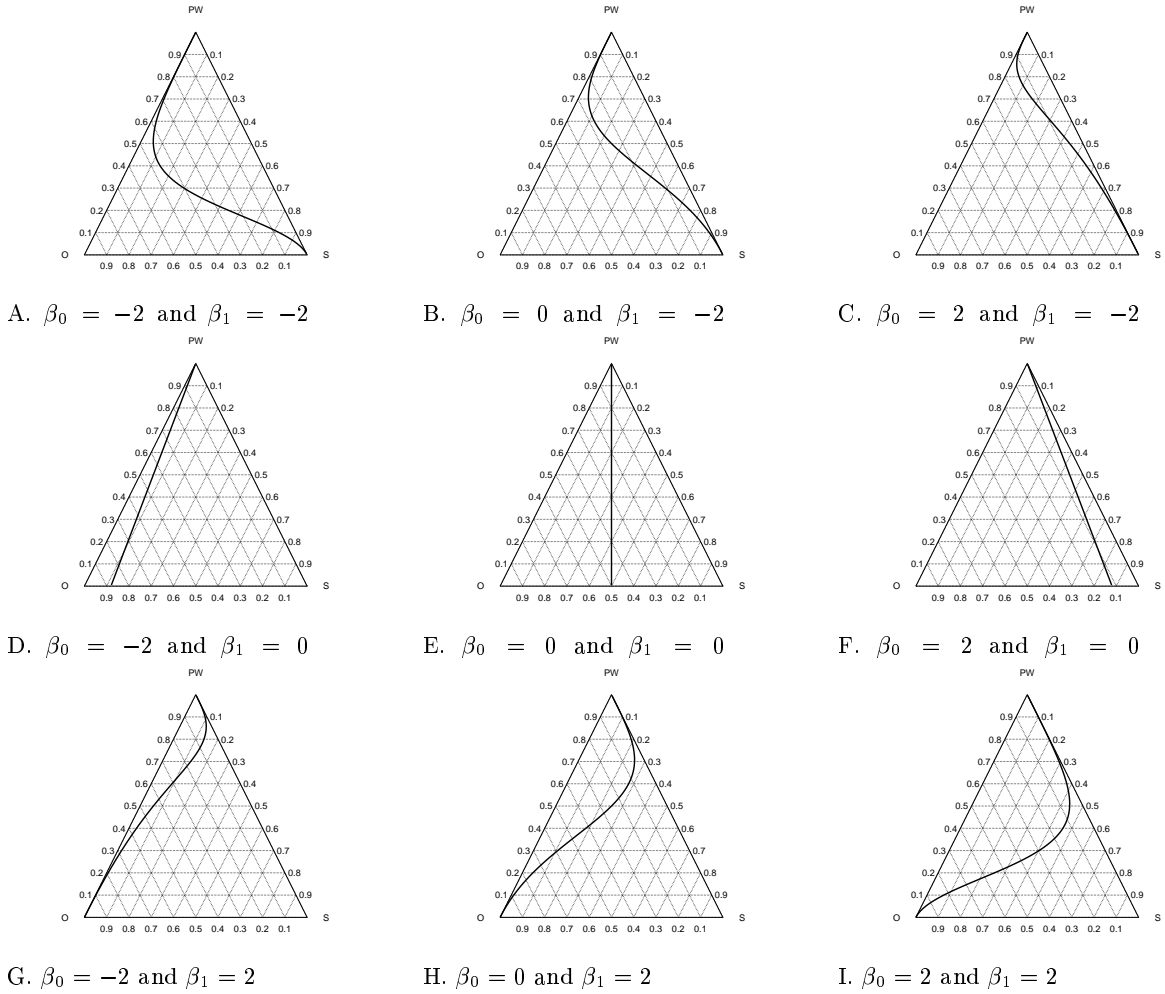


Figure 2: Examples of regression curves with the following values of the intercept: $\beta_0 = -2$, $\beta_0 = 0$, and $\beta_0 = 2$; and the following values of the slope: $\beta_1 = -2$, $\beta_1 = 0$, and $\beta_1 = 2$.

by β_1 , is most prominent for budgets with a large independent component. If $\beta_0 = 0$ (Fig. 2B, Fig. 2D, Fig. 2F) the change in the dependent ratio, expressed by β_1 , is most prominent for budgets with a medium sized independent component. If $\beta_0 > 0$ (Fig. 2C, Fig. 2F, Fig. 2I) the change in the dependent ratio, expressed by β_1 , is most prominent for budgets with a small independent component.

4 An Example

To illustrate the procedure, a sample of 500 observations were drawn from a multivariate normal distribution with mean vector μ and covariance matrix σ , where

$$\mu = \begin{bmatrix} -1.00 \\ -0.87 \\ -4.27 \\ -0.95 \\ -0.55 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 2.00 & 1.08 & 2.58 & 0.66 & 1.49 \\ 1.08 & 1.59 & -0.81 & 0.74 & 1.94 \\ 2.58 & -0.81 & 42.11 & 1.41 & 0.61 \\ 0.66 & 0.74 & 1.41 & 0.44 & 0.96 \\ 1.49 & 1.94 & 0.61 & 0.96 & 2.76 \end{bmatrix}.$$

An inverse logratio transformation yielded compositional data. It is assumed that these data are realistic time-budgets for the components in Table 1. Figure 3 shows a ternary diagram depicting the primary dependent component (S), the reference dependent component ($O = DW + PN + SP + AO$), and the independent component (PW) of the 500 observations.

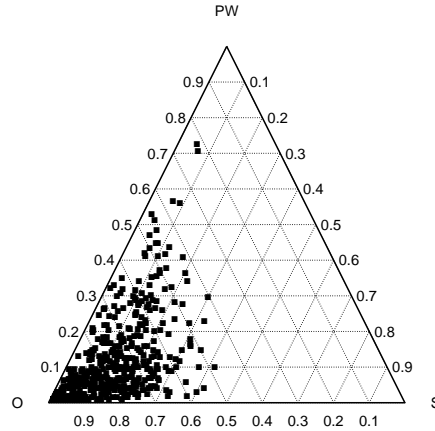


Figure 3: Graphic display of 500 time budgets.

The parameter estimates with standard errors of the linear regression are $\hat{\beta}_0 = -.8911(.1052)$ and $\hat{\beta}_1 = 0.5338(.0359)$. Both statistics are statistically significant ($p < .00005$). The multiple R-squared statistic, indicating the proportion of the total variance explained by the model, equals .3051. The regression curve is depicted in Figure 4, showing that if the component paid work increases, then the dependent ratio also increases. The expected amount of time spent on studying never exceeds 0.16.

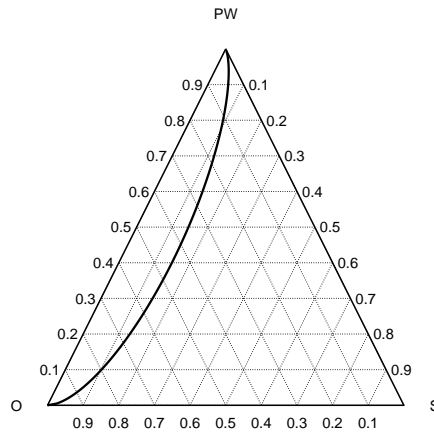


Figure 4: Estimated regression curve.

5 Discussion

The proposed method for performing a regression when both the dependent and the independent variable are components need not be restricted to a univariate regression. Other variables (not components) can be incorporated in Equation 1. This procedure is not meant as a solution for all research questions. Some research question may require another transformation of the compositional data to conduct a meaningful regression analysis. Computer code in Splus for plotting compositional data and regression curves in ternary diagrams is available free of charge from <http://spitswww.uvt.nl/~avdrark/research.html>

A topic that needs to be investigated is the effect of budgets with some very small components. These budgets have a large influence on the outcome of the regression analysis. Possible solutions are adding a small amount to these components or deleting them as unwanted outliers. Components which are exactly zero cause problems when taking the logarithm.

References

- Agresti, A. (2002). *Categorical Data Analysis*, second edition. Wiley (2002). *Categorical data analysis* (second edition). New York: Wiley.
- Aitchison, J. (1986/2003). *The statistical analysis of compositional data*. London: Chapman & Hall/ Caldwell, NJ: Blackburn.
- De Leeuw, J., Van der Heijden, P. G. M., and Verboon, P. (1990). A latent time budget model. *Statistica Neerlandica* 44(1), pp. 1-21.
- Magidson, J. and Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology* 31(1), pp. 223-264.
- Van der Ark, L. A., Van der Heijden, P. G. M., and Sikkel, D. (1999). On the identifiability in the latent budget model. *Journal of Classification* 16, pp. 117-137.
- Van der Ark, L. A. and Van der Heijden, P. G. M. (1998). Graphical display of latent budget analysis and latent class analysis, with special reference to correspondence analysis. In J. Blasius and M. Greenacre (Eds.), *Visualization of categorical data*, pp. 489-508. Boston: Academic

Press.

Van der Heijden, P. G. M., Mooijaart, A., and De Leeuw, J. (1992). Constrained latent budget analysis. *Sociological Methodology* 22(1), pp. 279-320.