Modelling, Measurement and Control D
Vol. 39, No. 1, December, 2018, pp. 25-29
Journal homepage: http://iieta.org/Journals/MMC/MMC_D

Empirical analysis on multiple regression modeling method of compositional data

Zhihui Zhang[1], Zhepeng Zheng[1], Chenxia Suo[1], Yong Yang[2*]

[1] Beijing Institute of Petrochemical Technology, Beijing 102617, China
[2] Postdoctoral Programme, Bank of Zhengzhou, Zhengzhou 450018, China

Corresponding Author Email: yangyonghebei@126.com

https://doi.org/10.18280/mmc_d.390104

Received: 10 November 2017
Accepted: 20 March 2018

Keywords:
compositional data, multiple regression data, partial least squares path analysis

ABSTRACT

The paper combines the logratio transformation method of compositional data with the partial least squares path analysis and puts forward the method of building the multiple linear regression model under the condition that dependent variables are compositional data and the relevant several independent variables are also compositional data. The modeling method can meet the fixed-sum constraint of compositional data, overcome the adverse effect of complete multi-collinearity on modeling in compositional data, and highlight the effect and significance of compositional data thematic meaning in modeling. As the application case, the paper used the suggested method and established the regression model among the employment demands of Beijing three industries, investments and GDP with the structural data of Beijing tertiary industry investments (including real estate), GDP and employment.

1. INTRODUCTION

In the data analysis on many fields such as society, economy and technology, compositional data is a kind of widely applied data type and can be used to reflect the investment structure, industry structure, resident consumption structure and other problems. According to the mathematical definition, compositional data refers to arbitrary nonnegative p-dimension vector quantity $X=(x_1, x_2, \ldots, x_p)'$, and the value of p components of X meets the following constraint conditions:

$$\sum_{j=1}^{p} x_j = 1, x_j \geq 0 \quad (1)$$

The formula (1) is also called the "fixed-sum constraint" and it is the basic feature of compositional data. The concept of compositional data was originated from the work of Ferrers (1866) in the nineteenth century at the earliest.

In 1897, Pearson pointed out in an authoritative article discussing spurious correlation, "In the practical compositional data analysis, fixed-sum constraint is often intentionally or unintentionally neglected, and some statistical methods designed for data without restrictive conditions are often abused inappropriately, leading to disastrous consequences." In 1986, Aitchison first published the Compositional Data Statistical Analysis, put forward the log ratio transformation of compositional data and completely established the logical normal distribution theory of compositional data [1].

The question to be discussed in the paper is if the dependent variable is a compositional data and the relevant several independent variables are also compositional data, how to build the multiple regression model among them. Generally speaking, when building multiple regression modeling for compositional data, there're mainly several difficult issues as follows:

(1) According to formula (1), compositional data features "fixed-sum constraint". When designing models and analytical methods, the constraint that the sum of the components of every compositional data equal 1 must always be met;

(2) The value of compositional data is between [0,1], so the data value of components is often very small and the reflection of the variation features of data is very insensitive, which often brings great difficulties to modeling and analysis;

(3) Due to the existence of "fixed-sum constraint", when the components of compositional data are analyzed as variables, the completely related problem of variables must exist, which makes the application condition of the classic least square regression method completely destroyed [2];

(4) If the independent variable set consists of multiple compositional data, the corresponding multiple regression analysis should include two levels. Each compositional data signifies a thematic meaning, and the compositional data consists of multiple components. Therefore, in the analysis process, the two levels should be analyzed after being divided clearly as much as possible, multiple components of compositional data cannot be briefly enumerated and processed with the analysis methods of common variables.

To solve the above problem, the paper proposes to combine the symmetrical logratio transformation with the partial least squares path analysis to realize the linear regression modeling of multiple compositional data. As the empirical study, the paper will use the cases of analyzing Beijing fixed asset investments, GDP output and labor employment according to the industrial structure to explain the working process and application value of multiple compositional data regression modeling

2. COMPOSITIONAL DATA AND LOGRATIO TRANSFORMATION

In Aitchison's book Compositional Data Statistical Analysis, the logratio transformation method of compositional data was raised. The computational process of the method is simple and convenient, and it has very good mathematical property under some conditions. For compositional data X.

$$\boldsymbol{X}=\left\{\left(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_p\right)'\in\boldsymbol{R}^p\left|\sum_{j=1}^{p}\boldsymbol{x}_j=1,\ \ 0<\boldsymbol{x}_j<1\right.\right\}$$

The following formula is normally used for Logratio transformation

$$\boldsymbol{u}_j=\log\left(\boldsymbol{x}_j/\boldsymbol{x}_p\right),\quad j=1,2,\cdots,p-1 \tag{2}$$

Zhang Yaoting once used transformation (2) and gave a regression modeling method that the independent variable is a compositional data. He pointed out that using the Logratio transformation method as the analytical variable has many advantages. The method can overcome the "fixed-sum constraint" of compositional data and partially eliminate the complete correlation among components, so as to use the least square method. Meanwhile, due to the value of u_j in $(-\infty, +\infty)$, this will bring much inconvenience to the selection of the model. However, the regret of the model is that the explanatory of the model is not strong because the new variable transformed from (2) cannot be corresponding to the original variable, so it is hard to be applied in the practical work.

To solve the problem, the paper proposed that the symmetrical Logratio transformation should be used according to the formula (3) in regression modeling:

$$\boldsymbol{v}_j=\log\frac{\boldsymbol{x}_j}{\sqrt[\boldsymbol{p}]{\prod_{\boldsymbol{i}=1}^{\boldsymbol{p}}\boldsymbol{x}_i}},\quad j=1,2,\cdots,p \tag{3}$$

Record: $V=(v_1, v_2, \ldots, v_p)'$,

Obviously: $v_j\in(-\infty, +\infty)$, j=1, 2, …, p

As the definition of $V=(v_1, v_2, \ldots, v_p)'$ is symmetrical to the components of the compositional vector $X=(x_1, x_2, \ldots, x_p)'$, using it to conduct regression modeling can better reflect the features of each component, so the explanatory of the model is also stronger. However, as the data obtained from symmetrical logratio transformation are completely relevant, it needs to use the partial least squares regression method for modeling.

On the other hand, the inverse transformation of the symmetrical Logratio transformation is expressed with formulas (4)-(6). Through the three formulas, the corresponding compositional data $X=(x_1, x_2, \ldots, x_p)'$ can be computed in reverse according to $V=(v_1, v_2, \ldots, v_p)'$.

$$w_j=v_j-v_p\ ,\ j=1,2,\cdots,p\text{-}1 \tag{4}$$

$$\boldsymbol{x}_j=\left\{\boldsymbol{e}^{\boldsymbol{w}_j}/(1+\sum_{i=1}^{p-1}\boldsymbol{e}^{\boldsymbol{w}_i})\right\},\quad j=1,2,\cdots,p-1 \tag{5}$$

$$\boldsymbol{x}_p=\left\{1/(1+\sum_{i=1}^{p-1}e^{w_i})\right\} \tag{6}$$

3. PLS PATH MODEL AND THE REGRESSION MODEL OF COMPOSITIONAL DATA

Using symmetrical Logratio transformation can ensure "fixed-sum constraint" in the regression modeling process and solve the problem of too small value of compositional data. For the linear regression that the dependent variable and independent variable are both a a compositional data, the common partial least square regression model can be directly used, which can overcome the complete multiple correlation in the variable set [3]. But for the regression issue that the independent variable is a set of multiple compositional data, the modeling process should be hierarchical. First, generalize the thematic meaning expressed by each compositional data, obtain the corresponding thematic variable, and then analyze the functional relationship among these thematic variables.

To clearly divide the two levels in the regression analysis process, the paper proposes the use of the PLS path model: First, extract the aggregate variable with the strongest explanatory to each compositional data – hidden variable (namely the thematic variable), then build the regression model of these hidden variables, analyze the causal relationship among compositional data, and reach a more accurate and reliable model [4].

3.1 PLS path model

The section will first give a simple narration to the PLS path model. Suppose there're j sets of variables $X_j=\{x_{j1}, x_{j2}, \ldots, x_{jk}\}$ for n observation sample points, variable x_{jh} is called "manifest variable", and suppose they are all centralized variables (namely the mean value of variables is zero)[5]. In addition, suppose each set of variables is roughly "unidimensional", that is each manifest variable in the set is mainly influenced by the same standardized "hidden variable" ξ_j. The relationship of the manifest variable and the hidden variable is expressed by the simple regression model:

$$\boldsymbol{x}_{jh}=\pi_{jh}\xi_j+\varepsilon_{jh} \tag{7}$$

The mean value of ξ_j is 0, the standard deviation is 1, and the mean value of the error term ε_{jh} is 0 and unrelated with the hidden variable ξ_j.

On the other hand, the structural model can be used to describe the relationship among hidden variables, and the form is as follows:

$$\xi_j=\sum_{i\neq j}\beta_{ji}\xi_i+\nu_j \tag{8}$$

The error term v_j should meet the supposition that the mean value is 0 and unrelated with $\xi_i(i\neq j)$. The causal relationship described by the structural model can be generalized into a 0/1 matrix, of which, the dimension is the number of hidden variables. If the hidden variable j explained the hidden variable i, the value of the element (i, j) in the matrix is 1; if not, it is 0. The matrix is called the internal design matrix.

To estimate the hidden variable ξ_j, on the one hand, it is thought the hidden variable ξ_j can be estimated by the linear combination of the manifest variable x_{jh}, recorded as Y_j:

$$\mathbf{Y}_j = \sum_h w_{jh} x_{jh} = \mathbf{X}_j \mathbf{w}_j \tag{9}$$

X_j is a matrix with ξ_j as the manifest variable and x_{jh} as the column vector.

On the other hand, if Yi is the estimated value of the hidden variable ξ_i related with ξ_j, Yi can be used to estimate the hidden variable ξ_j. The estimated value is recorded as Z_j:

$$Z_{\mathbf{j}} \propto \sum_{\mathbf{i}:\xi_i \text{ is connected with } \xi_j} e_{\mathbf{ji}} Y_{\mathbf{i}} \tag{10}$$

In the above formula, the meaning of the sign $\propto$ is: The left variable is the value of the right variable after being standardized. The internal weight of e_{ji} equals the sign function value of Y_j and connected correlation coefficients of Y_j, namely:

$$e_{ij} = \mathrm{sign}\left(\mathrm{cor}\left(\mathbf{Y}_j, \mathbf{Y}_i\right)\right) \tag{11}$$

The two methods put forward by Wold can calculate the weight w_{jh} in the formula (9). It is thought in the pattern A that w_{jh} is the covariance coefficient of x_{jh} about Z_j:

$$\mathbf{w}_j = \frac{1}{n} \mathbf{X}'_j \mathbf{Z}_j \tag{12}$$

In the pattern B, it is thought the weight vector w_j is the regression coefficient vector of Z_j about the manifest variable x_{jh} of ξ_j:

$$\mathbf{w}_j = \left(\mathbf{X}'_j \mathbf{X}_j\right)^{-1} \mathbf{X}'_j \mathbf{Z}_j \tag{13}$$

So, the iterative algorithm steps of PLS path analysis are as follows:

Step 1: The initial value of the vector Y_j is x_{j1}. By the formula (10), the estimated value of Z_j can be reached;

Step 2: According to the estimated value of Z_j, the weight vector w_j can be calculated by the formula (12) or (13);

Step 3: For w_j obtained through calculation, the new Y_j can be got via the formula (9); again, back to the step 1, it is until the computed convergence. Y_j got finally is taken as the estimated value $\hat{\xi}_j$ of the hidden variable ξ_j;

Step 4: At last, after using the estimated value $\hat{\xi}_j$ to replace the hidden variable ξ_j, use the multiple regression method of the common least squares to estimate the coefficients in the model (8).

3.2 The multiple regression modeling method of compositional data

The multiple regression modeling method can be obtained as follows according to the above-mentioned method of compositional data logratio transformation and PLS path analysis:

(1) For the dependent variable Y and independent variable $X_1, \ldots, X_p$, the symmetrical logratio transformation is made for them according to the formula (3) first; the vectors after transformation are $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_p$;

(2) Use the PLS path analysis model to extract the hidden variable of $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_p$ respectively;

(3) Build the multiple regression model of hidden variables to analyze the correlation means among composition data;

(4) In the prediction application, the symmetrical logratio transformation is made for the independent variables $X_1, \ldots, X_p$ of prediction time points to get $\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_p$, and then it is put into the PLS path analysis model to get the predicted value of $\widetilde{\mathbf{Y}}$. Then according to the inverse transformation formulas (4)-(6) of logarithm transformation, the predicted value of the original dependent variable Y can be got.

3.3 The unidimensional guarantee of the variable set

It is noteworthy that in apply the PLS path analysis model, it needs to verify that each variable set is unidimensional. In the application, the simplest verification method is to make the main component analysis on a variable set [6]. If only the first characteristic value is greater than 1 and other characteristic values are all less than 1, it is thought the variable set is roughly unidimensional.

When a component vector cannot meet the condition, the components should be further grouped, making each group of variables unidimensional. For this end, the method of variable clustering analysis can be used. If more than one characteristic value of a set of variables is greater than 1, the variable set should be classified to make the sub-variable group of each category unidimensional. Then the variable group after being grouped is transformed to new compositional data and then the modeling method raised in the paper can be used.

4. CASE ANALYSIS

In the case, Beijing's fixed asset investments (including real estate), output and employment status are taken as hidden variables, and the structural proportions of them in the three industries are taken as the manifest variables. The investments, output and employment proportion data of Beijing three industries are collected in the chronological order, and the causal relationship among hidden variables and between manifest variables and hidden variables is analyzed through building the PLS path model.

Table 1. The meaning of hidden variables and their manifest variables

Hidden variable	Investment status (inv., ξ1)	Output status (GDP,ξ2)	Employment status(emp.,ξ3)
Manifest variable	a) Primary industry investment proportion (inv1) b) Secondary industry investment proportion(inv2) c)Tertiary industry investment proportion(inv3)	a) Primary industryoutput proportion(GDP1) b) Secondary industry output proportion (GDP2) c)Tertiary industry output proportion (GDP3)	a) Primary industryemploymetn proportion(emp1) b) Secondary industry employmetn proportion (emp2) c)Tertiary industryemploymetn proportion (emp3)

According to the law of macroeconomic operation, the investment status directly influences output and the employment status and indirectly influences employment through output. It is thought that the investment status is the exogenous hidden variable and the employment status and output status are endogenous variables. The internal design matrix of the initially established structure model is:

Table 2. Internal design matrix of structure model (1)

	Investment	Output	Employment
Investment	0	0	0
Output	1	0	0
Employment	1	1	0

The corresponding path diagram is:

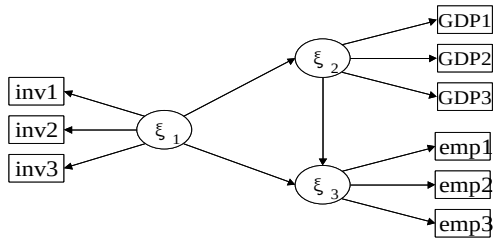

Figure 1. The relationship diagram of investments, output and employment

Table 3 is the proportion data of the investment, output and employment of three industries in Beijing from 2003 to 2015 and the value through logratio transformation. The range of the data has been unfolded:

Table 3. The logratio transformation data of Beijing investment, output and employment proportion

Year	Fixed asset investment proportion			GDP proportion			Employment proportion		
	Primary industry (inv1)	Secondary industry (inv2)	Tertiary industry (inv3)	Primary industry (GDP1)	Secondary industry (GDP2)	Tertiary industry (GDP3)	Primary industry (emp1)	Secondary industry (emp2)	Tertiary industry (emp3)
2003	-2.31	1.00	1.31	-1.09	0.70	0.40	-1.04	0.59	0.45
2004	-2.28	1.05	1.23	-1.20	0.65	0.55	-1.05	0.57	0.48
2005	-2.26	1.14	1.12	-1.28	0.69	0.59	-1.16	0.58	0.58
2006	-2.99	1.61	1.38	-1.35	0.70	0.65	-1.41	0.69	0.71
2007	-2.93	1.41	1.52	-1.27	0.63	0.65	-1.34	0.56	0.78
2008	-3.21	1.89	1.31	-1.39	0.63	0.76	-1.38	0.56	0.82
2009	-3.31	1.88	1.43	-1.47	0.63	0.85	-1.34	0.50	0.84
2010	-4.14	2.24	1.90	-1.54	0.62	0.91	-1.36	0.50	0.85
2011	-3.91	2.05	1.86	-1.59	0.61	0.98	-1.28	0.38	0.90
2012	-3.83	2.04	1.79	-1.64	0.62	1.02	-1.22	0.31	0.91
2013	-3.52	1.90	1.62	-1.71	0.64	1.07	-1.24	0.27	0.97
2014	-3.95	2.28	1.68	-1.77	0.63	1.14	-1.29	0.31	0.98
2015	-4.02	2.38	1.63	-1.82	0.62	1.20	-1.43	0.37	1.05

Note: Data source before transformation: Beijing Statistical Yearbook, 2004-2015.

According to the model of Figure 1, the PLS path analysis is used to analyze the data of Table 3 and the path coefficients of the hidden variables are shown in the following table:

Table 4. The path coefficients of hidden variables (1)

	Investment	Output	Employment
Investment	0	0	0
Output	0.914	0	0
Employment	0.151	0.815	0

The direct influence coefficient of investments on employment is 0.151, non-significant. Apart from the path, the additional internal design matrix is shown in Table 5. The PLS path analysis is again used for the model, and the path coefficients among hidden variables are shown in Table 6.

It can be judged that the direct influence coefficient of Beijing investments for GDP output is 0.913 and the direct influence coefficient of the output status for the employment status is 0.965. The indirect influence coefficient of the investment status for the employment status through the GDP output is 0.913*0.965=0.881. The goodness of fit of endogenous hidden variables of the above model is 0.8345and 0.9318, better fit the observation data.

The corresponding path diagram is showin Figure 2.

The paths shown in Figure 2 are all prominent and the endogenous hidden variables have very high goodness of fit, which indicates the model has very strong explanatory ability for original data. Meanwhile, the values of the yearly hidden variables of the model shown in Figure 2 are shown in the following Table 7.

Table 5. Internal design matrix (2)

	Investment	Output	Employment
Investment	0	0	0
Output	1	0	0
Employment	0	1	0

Table 6. Path coefficients of hidden variables (2)

	Investment	Output	Employment
Investment	0	0	0
Output	0.913	0	0
Employment	0	0.965	0

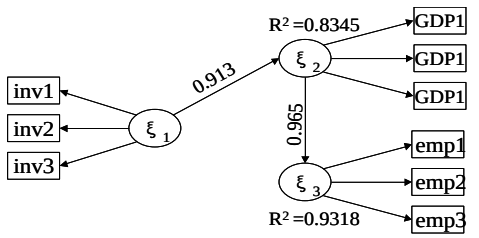

Figure 2. Beijing investments, output and employment relationship diagram

Table 7. The hidden variables of the yearly investment, output and employment of Beijing

Year	Investment status (inv.,ξ1)	Output status (GDP, ξ2)	Employment status (emp. ,ξ3)
2003	1.415	1.906	1.797
2004	1.495	0.969	1.629
2005	1.575	1.177	1.159
2006	0.462	1.063	0.642
2007	0.475	0.486	0.172
2008	0.207	0.106	-0.021
2009	0.005	-0.174	-0.189
2010	-1.344	-0.491	-0.259
2011	-1.014	-0.79	-0.621
2012	-0.872	-0.838	-0.753
2013	-0.371	-0.833	-1.09
2014	-0.989	-1.149	-1.09
2015	-1.043	-1.432	-1.378

As there's stronger pertinence between hidden variables and manifest variables, the PLS regression model of hidden variables for manifest variables can be built. With the hidden variables as the dependent variables and the manifest variables as the independent variables, PLS regression can be conducted, and the relationship between hidden variables and manifest variables can be reached:

ξ_1 =0.3636 inv1 - 0.3661 inv2 - 0.3071 inv3

ξ_2 =0.3764 GDP1 + 0.3111 GDP2 - 0.3826 GDP3

ξ_3 =0.2666 emp1 + 0.4320 emp2 - 0.4786 emp3

In addition, there's a certain dependency among investments, output and employment, the regression model among components can be sought or built. For the case, the polynomial of hidden variables can be built. After conducting polynomial regression, the relationship of employment (ξ3), investments (ξ1) and output (ξ2) can be got:

$\xi 3 = -0.2407 + 0.2406(\xi 1)^2 + 0.8893\xi 2$

t-statistical magnitude: (-2.61) (3.32) (14.47)

p-value:0.02 0.008 0.000

F-statistical magnitude=148.89, adjusted determination coefficient $\bar{R}^2$=0.96.

The model fits the data well and the coefficients and the overall model all passed the test, with the statistical significance. From another perspective, it indicated that Beijing employment status is influenced by the investment status, output status and the quantitative relation among the three factors.

5. SUMMARY

The paper puts forward a method of building the multiple regression models under the condition that dependent variables and independent variables are all compositional data. As the compositional data have the fixed-sum constraint, the application condition of the classic least square regression method is completely destroyed, so the classic regression method cannot be used for modeling. The paper spreads the compositional data to the whole real number field through logratio transformation, reduces the dimension and extracts new aggregate variables – hidden variable for the multi-dimensional compositional data after transformation through building the PLS path model, and researches the multiple regression relationship among hidden variables. In the modeling process, the method can meet the fixed-sum constraint of compositional data, overcome the adverse effect of complete multicollinearity on modeling in the compositional data, and highlight the thematic meaning of compositional data and its effect and significance in modeling. To further specify the working process of the multiple compositional data regression modeling method, the paper applied the suggested method, used the structural data of the investments, GDP and employment of Beijing three industries, and built the regression model among the employment status, investment status and GDP of Beijing three industries. The case study indicated that the modeling method raised in the paper can provide an effective technological approach to solve such problems, with important application value.

ACKNOWLEDGEMENTS

This paper was funded by three projects: BIPT-POPME; Development Research Centre of Beijing New Modern Industrial Area (2016); BIPT-ER (2014); URT2017J00013.

REFERENCES

[1] Aitchison J. (1986). The statistical analysis of compositional data. London: Chapman and Hall.

[2] Chin WW. (1998). The partial least squares approach for structural equation modeling. in: G.A. Marcoulides (Ed.) Modern Methods for Business Research, Lawrence Erlbaum Associates, 295-336.

[3] Guinot C, Latreille J, Tenenhaus M. (2001). PLS path modellind and multiple table analysis. Application to the cosmetic habits of women in Ile-de-France. Chemometrics and Intelligent Laboratory Systems 58: 247-259.

[4] Lohmöller JB. (1989). Latent variables path modeling with partial least squares. Physica-Verlag, Heildelberg 34(1): 110-111.

[5] Bayol MP, Foye ADL, Tellier C, Tenenhaus M. (2000). Use of PLS path modeling to estimate the European consumer satisfaction index (ECSI) model. Statistica Applicata – Italian Journal of Applied Statistics 12(3): 361-375.

[6] Wang HW, Huang W. (2013). Linear regression model of compositional data. System Engineering (2).