Neurocomputing 122 (2013) 490–500

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multiple linear regression modeling for compositional data

Huiwen Wang, Liying Shangguan, Junjie Wu*, Rong Guan

School of Economics and Management, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 30 November 2012
Received in revised form
13 March 2013
Accepted 23 May 2013
Communicated by D. Tao
Available online 1 July 2013

Keywords:
Compositional data
Simplex space
Isometric logratio transformation
Multiple linear regression

ABSTRACT

Compositional data, containing relative information, occur regularly in many disciplines and practical situations. Multivariate statistics methods including regression analysis have been adopted to model compositional data, but the existing research is still scattered and fragmented. This paper contributes to modeling the linear regression relationship for compositional data as both dependent and independent variables. First, some operations in Simplex space, such as the perturbation operation, the power transformation, and the inner product, are defined for compositional-data vectors. The regression models are then built by the original compositional data and transformed data, respectively, after the introduction of the Isometric Logratio Transformation (*ilr*). By theoretical inference, it turns out that the two models are equivalent in essence using the ordinary least squares (OLS) method. Two measures for testing goodness of fit, i.e., the observed squared correlation coefficient R^2 and the cross validated squared correlation coefficient Q^2, are also proposed to evaluate the regression models. Besides, the estimated regression parameters are explained to indicate the notion of relative elasticity. An empirical analysis finally illustrates the usefulness of the multiple linear regression models for compositional-data variables.

1. Introduction

In recent decades, various application fields have witnessed the explosion of massive data [1,2] in various complex types, such as symbolic data [3,4], functional data [5,6], compositional data [7–9] and the like [10–12]. In this paper, our focus will be placed on compositional data, which has aroused considerable interests from academy to industry. Compositional data convey the structure information that parts organize a whole in a quantitative way. Expressed in proportions or percentages, all components are subject to non-negative and constant-sum (e.g., 100 weight percent) constraints, which contributes to closure of compositional data [9,13].

In many aspects of real life, compositional data are widely emerging as a measurement for the structure information of an activity. For example, in order to see how granted patents are affected by accepted patent applications in three categories: inventions, utility models and designs, it is not the absolute accepted and granted values but the proportional values are more meaningful. The proportions of the accepted or granted patents in the three classes thus constitute the compositional data, upon which we can establish a regression model to characterize the relationship between the accepted and granted patents. In economics, to investigate the association between gross domestic products (*GDP*) and employment across three industries, we can focus on the whole compositional data changes of *GDP* affected by the changes of employment compositional data, rather than the independent influence between the corresponding components. A regression model can then be built on the compositional data to quantify the relationship. However, classical statistical methods, including classical regression models [14,15] are mostly concerned with numerical data, and thus not directly applicable to compositional-data variables. This motivates our study on modeling a regression relationship between dependent and independent variables of compositional data.

Indeed, employing standard linear regression analysis for compositional data often leads to undesirable properties like the response variable incoherence. One plausible solution is to remove the constraints of compositional data before building the model. Inspired by Ref. [8], a lot of research efforts have been devoted for investigating transformation approaches [16,17]. The general idea is that the constraints of compositional data are removed first, and then traditional statistics methods are performed on the transformed vectors, and finally the results are transformed back into the original space. A family of logratio transformations has been introduced [16,17], including the additive logratio (*alr*) transformation, the centered logratio (*clr*) transformation, and the isometric logratio (*ilr*) transformation.

In recent years, multivariate statistics for direct modeling of compositional data has become a research focus [18–22]. As to regression models of compositional data, there exist two types of work in the literature. *Type I* refers to linear regression models for components of compositional data as explanatory variables [16,23–25]. The first model along this line can be written in terms of a conditional expected value as

$$E(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i, \tag{1}$$

* Corresponding author. Tel./fax: +86 10 8233 9983.
E-mail addresses: wujj02@gmail.com, wujj@buaa.edu.cn (J. Wu).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.05.025

or

$$E(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \beta_{ij} x_i x_j, \tag{2}$$

with unknown parameters β_0, $\beta_1, \ldots, \beta_D$ and β_{ij} $(1 \le i \le D-1, i+1 \le j \le D)$. Aitchison contributed the first study on this model [16,23]. Later on, Hron et al. argued that it is difficult to explain the estimated parameters, especially in more complicated models [25]. By using the *ilr* transformation, they discussed the regression model $E(Y|\mathbf{x}) = \beta_0 + \sum_{i=1}^{D} \beta_i x_i$, and presented parameter estimation based on classical linear regression, which improves the model interpretability [25]. The second model is to find a line fitting the set of n two-dimensional data points, which can be achieved by the *ilr* transformation for three-part compositions [24]. Statistic analysis, including fitted lines and confidence regions, was investigated for both original compositional data and the transformed real vectors [24].

Type II research centers on the linear regression model as follows:

$$\mathbf{z}(x_1, \ldots, x_p) = \mathbf{b}_0 \oplus [\oplus_{i=1}^{p} (x_i \otimes \mathbf{b}_i)] \oplus \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{z}$ is the response variable of compositional data, $x_1, \ldots, x_p$ are real explanatory variables, and the slope vectors $\mathbf{b}_1, \ldots, \mathbf{b}_p$, the intercept vector $\mathbf{b}_0$, and the residual vector $\boldsymbol{\varepsilon}$ are all compositional data. The operators $\oplus$ and $\otimes$ are, respectively, perturbation operation and power transformation in Aitchison geometry [26]. As can be seen, this model has p real explanatory variables and a compositional-data response, which was first proposed in Refs. [7,27] but gained great improvements in Refs. [26,28,29]. The residual distributions for regression models, most of which are the logistic-normal or the Dirichlet distribution, were also studied [27,30]. In addition, Egozcue et al. claimed that the logistic-normal linear regression is the simplest regression method, compared with the linear regression model with Dirichlet distributed residuals [30]. In summary, the regression models discussed in previous studies refer to either *Type I* models, with components of compositional data as explanatory variables, or *Type II* models, with dependent variable of compositional data and the independent variables of real data. Little attention, however, has been placed on the regression model for compositional data as explained and explanatory variables.

This paper will investigate the linear regression modeling for compositional-data variables, which provides an innovative way for parameter estimation, model evaluation and interpretation. From the viewpoint of modeling, operators of compositional-data vectors, including perturbation operation, power transformation, and inner product, are proposed in Aitchison geometry. Based on these Simplex operations, a regression model for compositional-data vectors is then built in Aitchison geometry. We also establish a regression model in orthonormal coordinates after transforming the compositional data into real vectors by the *ilr* transformation. The most beautiful part regarding these two modeling methods is, they are completely equivalence in essence, which implies that our inner product definition for compositional-data vectors is reasonable.

Furthermore, we demonstrate how to evaluate and interpret the modeling results. On one hand, the indexes, including the observed and the cross-validated squared correlation coefficients, are deduced to evaluate the regression models. Again we prove that the proposed measures have equivalence in the two methods. On the other hand, the built models have an explicit economic explanation to the estimated parameters; that is, these parameters indicate the relative elasticity for variables involved in modeling. This is considered valuable, which provides clear explanations to the relationship between variables. In addition, it is worth noting that our models work only for compositional data without zero components. If that is not the case, we can manipulate the data in advance by replacing the zero elements with small amounts or circumventing zero problems by rank transformation before modeling [16,31].

The remainder of this paper is organized as follows. In Section 2, the Aitchison geometry and various transformations are briefly introduced. Section 3 proposes the concepts and algorithms for compositional-data vectors. The linear regression models of original compositional data and transformed data are built in Section 4, with the deduction of evaluation indexes and the interpretation of estimated parameters. A case study in Section 5 illustrates the usefulness of the linear regression models. We finally conclude our work in Section 6.

2. Preliminaries

In this section, we briefly introduce some basic knowledge about the Aitchison geometry in Simplex space, and the ways to transform Simplex space to the familiar real space for modeling compositional-data variables. Note that we use "$[\cdots]$" to denote a compositional-data unit in Simplex space, and "$(\cdots)$" to denote a transversal vector in real space. The transposition operation on a vector or a matrix is indicated by the prime symbol.

2.1. Aitchison geometry

The D-part Simplex space on real number field $\mathbb{R}$ is defined as

$$S^D = \left\{ \mathbf{u} | \mathbf{u} = [u_1, u_2, \ldots, u_D]', u_j > 0 \forall j, \sum_{j=1}^{D} u_j = 1 \right\}. \tag{4}$$

Basic operations in Simplex space and their properties were summarized in Refs. [9,26]. In particular, the structure of S^D is based on the *perturbation operation*, defined for any two compositional-data units $\mathbf{u}, \mathbf{v} \in S^D$ as follows:

$$\mathbf{u} \oplus \mathbf{v} = \mathcal{C}(u_1 v_1, u_2 v_2, \ldots, u_D v_D), \tag{5}$$

and on an external operation called *power transformation*, given for $\mathbf{u} \in S^D$ and $\beta \in \mathbb{R}$ by

$$\beta \otimes \mathbf{u} = \mathcal{C}(u_1^{\beta}, u_2^{\beta}, \ldots, u_D^{\beta}), \tag{6}$$

where $\mathcal{C}$ denotes the closure of a vector, i.e.,

$$\mathcal{C}(x_1, x_2, \ldots, x_D) = \left[\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right]', \tag{7}$$

which guarantees that the operation result is still in S^D [32,33]. Apparently, the zero element with respect to the perturbation operation is $\mathcal{C}(1, 1, \ldots, 1)$.

Based on these two operations, $\mathbf{u} \ominus \mathbf{v}$ can be defined as [34]

$$\mathbf{u} \ominus \mathbf{v} = \mathbf{u} \oplus ((-1) \otimes \mathbf{v}) = \mathcal{C}\left(\frac{u_1}{v_1}, \frac{u_2}{v_2}, \ldots, \frac{u_D}{v_D} \right). \tag{8}$$

The inner product is defined as

$$(\mathbf{u}, \mathbf{v})_S = \frac{1}{D} \sum_{i<j} \log \frac{u_i}{u_j} \log \frac{v_i}{v_j} = \sum_{i=1}^{D} \log \frac{u_i}{g(\mathbf{u})} \log \frac{v_i}{g(\mathbf{v})}, \tag{9}$$

where $g(u) = \sqrt[D]{u_1 u_2 \ldots u_D}$ is the geometric mean of all parts of $\mathbf{u}$. It is straightforward to show that $(\mathbf{u}, \mathbf{v})_S$ is an inner product, and therefore S^D is also the $(D\text{-}1)$-dimensional Hilbert space [32,33]. The norm and the distance in S^D can be defined accordingly as follows:

$$\|\mathbf{u}\|_S^2 = (\mathbf{u}, \mathbf{u})_S, \tag{10}$$

$$d_S^2(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} \ominus \mathbf{v}\|_S^2 = \sum_{i=1}^{D} \left(\log \frac{u_i}{g(\mathbf{u})} - \log \frac{v_i}{g(\mathbf{v})} \right)^2, \tag{11}$$

where the subscript S indicates that the definitions are given in S^D.