Computational Statistics and Data Analysis 55 (2011) 375–388

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models

Jouni Kuha [a,*], David Firth [b]

[a] Department of Statistics, London School of Economics, UK

[b] Department of Statistics, University of Warwick, Coventry, UK

ARTICLE INFO

Article history:
Received 15 June 2009
Received in revised form 10 May 2010
Accepted 10 May 2010
Available online 19 May 2010

Keywords:
Bias reduction
Delta
Iterative proportional fitting
Model selection
Raking
Stratified sampling
Total variation distance

ABSTRACT

The index of dissimilarity, often denoted by Delta, is commonly used, especially in social science and with large datasets, to describe the lack of fit of models for categorical data. The definition and sampling properties of the index for general loglinear and log-multiplicative models are investigated. It is argued that in some applications a standardized version of the index is appropriate for interpretation. A simple, approximate variance formula is derived for the index, whether standardized or not. A simple bias reduction formula is also given. The accuracy of these formulae and of confidence intervals based upon them is investigated in a simulation study based on large-scale social mobility data.

1. Introduction

In the presence of a large amount of informative data, even 'good' models are typically rejected by conventional lack-of-fit tests based on statistical significance. In such cases attention turns to assessment of the extent to which a model's lack of fit is important from the subject-matter point of view. For example, in situations where a statistical model is to be used mainly or entirely as a basis for forecasting, predictive performance will usually be a more important criterion than formal goodness of fit.

In the context of models for categorical data a commonly used statistic is the so-called *index of dissimilarity* or *dissimilarity index* (e.g. Agresti, 2002, pp. 329–330), which aims to quantify lack of model fit by estimating the smallest fraction of the population under study that would need to be re-classified in order to make the fitted model exactly correct. The index of dissimilarity thus has a fairly direct interpretation in terms of the magnitude of departures from the model, and the statistic itself is simple to compute from model residuals (Section 2 below). These appealing properties have led to routine use of the index of dissimilarity for model assessment, especially in social science where the computed statistic is sometimes referred to simply as 'Delta' (e.g. Erola and Moisio, 2007; Jonsson et al., 2009; Pfeffer, 2008; Uggen and Blackstone, 2004; Wells et al., 2003). The Delta statistic is used as a supplement to, rather than a replacement for, model-selection criteria such as those based on the log likelihood. The present work explores the definition and estimation of the index of dissimilarity in a fairly general setting.

* Corresponding address: Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK. Tel.: +44 20 7955 6835; fax: +44 20 7955 7005.
E-mail addresses: j.kuha@lse.ac.uk (J. Kuha), d.firth@warwick.ac.uk (D. Firth).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.005

Previously the sample index of dissimilarity has been used mainly as a descriptive device, and in particular no measure of the associated uncertainty, such as an estimated standard error, has been available. A primary aim of this paper is to rectify this deficiency. An approximate variance formula is derived in Section 4, based on large-sample arguments. The same large-sample considerations also suggest an approximate bias correction, given in Section 5. These results apply to general parametric models for data collected by multinomial or Poisson sampling, including any loglinear and log-multiplicative models; and in keeping with the simplicity of the index itself, the variance and bias formulae are straightforward to compute from quantities made available by standard software for such models. As a preliminary to all of this work, the definition of the index of dissimilarity itself is placed under scrutiny, and it is suggested that often a standardized version of the index will be of most interest (Section 3). The accuracy of the various approximations and the effects of standardization are investigated in Section 6 using simulation, based on data from six countries in a large-scale study of intergenerational social class mobility.

2. Index of dissimilarity and models for categorical data

Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_K)'$ is a vector of K observed cell counts corresponding to the cells of a possibly multidimensional contingency table with a sample size $N = \sum_{i=1}^{K} Y_i$ and observed cell proportions $\mathbf{p} = (p_1, \ldots, p_K)' = \mathbf{Y}/N$. The corresponding vectors of fitted counts and proportions for some model M of interest estimated from the data are denoted by $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K)'$ and $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_K)' = \hat{\mathbf{Y}}/N$. The index of dissimilarity for the fitted model is defined as

$$\hat{\Delta} = \frac{\sum_{i=1}^{K} |Y_i - \hat{Y}_i|}{2N} = \frac{\sum_{i=1}^{K} |p_i - \hat{\pi}_i|}{2} = \frac{\hat{\boldsymbol{\delta}}'\hat{\mathbf{e}}}{2}, \tag{1}$$

where $\hat{\mathbf{e}} = (\hat{e}_1, \ldots, \hat{e}_K)' = (p_1 - \hat{\pi}_1, \ldots, p_K - \hat{\pi}_K)'$ are the raw residuals, and their signs are indicated by $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)' = (\mathrm{sgn}(\hat{e}_1), \ldots, \mathrm{sgn}(\hat{e}_K))'$. We assume that the fitted model has the property that $\sum_i \hat{Y}_i = N$, so that the residuals sum to zero and $\hat{\boldsymbol{\pi}}$ are the fitted cell proportions. This will be true, in particular, for hierarchical loglinear and log-multiplicative models fitted by maximum likelihood, which are the main focus of this article. The index $\hat{\Delta}$ can then be interpreted as the smallest proportion of observations in $\mathbf{Y}$ that would need to re-classified to other cells to make all observed cell counts exactly equal to the fitted values.

A version of the index of dissimilarity has a long history in sociology and human geography as a measure of residential and geographic segregation. According to Goodman and Kruskal (1959), the index was first suggested by Gini (1914); other early references include Jahn et al. (1947) and Duncan and Duncan (1955); see also White (1986) for a more recent introduction and further references on segregation indices. Suppose that we want to compare the distributions of two groups, for example blacks and whites, across C locations such as schools or neighborhoods. The populations are completely segregated if no members of the two groups share the same location, and completely unsegregated if the proportions of the two groups are the same in every location. The segregation index of dissimilarity is defined as

$$D = \frac{1}{2} \sum_{j=1}^{C} \left| \frac{Y_{1j}}{N_1} - \frac{Y_{2j}}{N_2} \right|, \tag{2}$$

where Y_{ij} is the number of members of group i in location j, and N_i is total number of members of group i in the sample. The index (2) can be interpreted as the general measure (1) where the observed data are one of the rows of the group-by-location table and the 'model' is given by the column proportions in the other row. Alternatively, D is $\hat{\Delta}$ for the independence model for the two-way table, divided by $N^2/(2N_1N_2)$, which is the maximum value that $\hat{\Delta}$ can achieve for the independence model when the row totals are regarded as fixed; for more general models the maximum achievable $\hat{\Delta}$ is usually not available in a closed form.

The statistic $\hat{\Delta}$ estimates a corresponding population quantity. Suppose that the true distribution of $\mathbf{Y}$ given N is multinomial with cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, so that $\mathbf{p}$ converges to $\boldsymbol{\pi}$ as N increases. It is assumed that $\pi_i > 0$ for all i. If the table has structural zeros for which the cell probability is known to be zero, these contribute nothing to (1) and can be omitted from $\mathbf{Y}$.

Suppose further that the fitted proportions $\hat{\boldsymbol{\pi}}$ converge to $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_K^*)'$, which differ from $\boldsymbol{\pi}$ unless model M is the true model. The population value of the index of dissimilarity is then defined as

$$\Delta = \frac{\sum_{i=1}^{K} |\pi_i - \pi_i^*|}{2} = \frac{\boldsymbol{\delta}'\mathbf{e}}{2}, \tag{3}$$

where $\mathbf{e} = (e_1, \ldots, e_K)' = (\pi_1 - \pi_1^*, \ldots, \pi_K - \pi_K^*)'$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)' = (\mathrm{sgn}(e_1), \ldots, \mathrm{sgn}(e_K))'$. This is consistently estimated by $\hat{\Delta}$. The index (3) is the *total variation distance* between the two discrete distributions with probabilities $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$ (Feller, 1971, the definition there is for 2Δ). It can also be interpreted as $\Delta = \sup_{A \subset \mathcal{K}} |\pi(A) - \pi^*(A)|$ where $\pi(A) = \sum_{i \in A} \pi_i$, $\pi^*(A) = \sum_{i \in A} \pi_i^*$ and $\mathcal{K} = \{1, 2, \ldots, K\}$.