Journal of Multivariate Analysis 123 (2014) 311–329

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Transformed goodness-of-fit statistics for a generalized linear model of binary data

Nobuhiro Taneichi [a,*], Yuri Sekiya [b], Jun Toyama [c]

[a] Department of Mathematics and Computer Science, Graduate School of Science and Engineering, Kagoshima University, 1-21-35 Korimoto, Kagoshima 890-0065, Japan

[b] Kushiro Campus, Hokkaido University of Education, Kushiro 085-8580, Japan

[c] The Institute for Use of Mathematics, Sapporo 063-0001, Japan

ARTICLE INFO

Article history:
Received 13 October 2012
Available online 15 October 2013

AMS 2000 subject classifications:
62E20
62H10

Keywords:
Asymptotic expansion
Binary data
ϕ-divergence statistics
Generalized linear model
Improved transformation

ABSTRACT

In a generalized linear model of binary data, we consider models based on a general link function including a logistic regression model and a probit model as special cases. For testing the null hypothesis H_0 that the considered model is correct, we consider a family of ϕ-divergence goodness-of-fit test statistics C_ϕ that includes a power divergence family of statistics R^a. We propose a transformed C_ϕ statistics that improves the speed of convergence to a chi-square limiting distribution and show numerically that the transformed R^a statistic performs well. We also give a real data example of the transformed R^a statistic being more reliable than the original R^a statistic for testing H_0.

1. Introduction

We consider generalized linear models [11] in which the response variables are measured on a binary scale. Let N independent random variables $Y_\alpha,\ \alpha = 1, \dots, N$ corresponding to the number of successes in N different subgroups be distributed according to binomial distributions $B(n_\alpha, \pi_\alpha),\ \alpha = 1, \dots, N$. If we use a monotone and differentiable function $g(\cdot)$ as a link function, we obtain a generalized linear model for binary data as follows.

$$g(\pi_\alpha) = \boldsymbol{x}_\alpha' \boldsymbol{\beta}, \quad (\alpha = 1, \dots, N), \tag{1}$$

where $\boldsymbol{x}_\alpha = (x_{\alpha 1}, \dots, x_{\alpha p})'$, $(\alpha = 1, \dots, N)$, are covariate vectors and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is an unknown parameter vector and $p < N$. When $g(t)$ is a canonical link function, that is,

$$g(t) = \log\left(\frac{t}{1-t}\right), \tag{2}$$

model (1) is a logistic regression model. When

$$g(t) = g_P(t) = \Phi^{-1}(t), \tag{3}$$

* Corresponding author.
E-mail addresses: taneichi@sci.kagoshima-u.ac.jp (N. Taneichi), sekiya.yuri@k.hokkyodai.ac.jp (Y. Sekiya), mandheling@nifty.com (J. Toyama).

0047-259X/$ – see front matter
http://dx.doi.org/10.1016/j.jmva.2013.09.014

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution, model (1) is a probit model. When

$$g(t) = \log\{-\log(1-t)\}, \tag{4}$$

model (1) is a complementary log–log model. Aranda-Ordaz [1] considered a family of link functions,

$$g(t) = g_c(t) = \log\left\{\frac{(1-t)^{-c}-1}{c}\right\}, \quad (c \geq 0), \tag{5}$$

that depend on parameter c. By this family of link functions, we obtain a family of models that includes a logistic regression model when $c = 1$ and a complementary log–log model when $c = 0$ in the limit.

We consider the null hypothesis

$$H_0^g : \pi_\alpha = \pi_\alpha(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}'_\alpha\boldsymbol{\beta}), \quad (\alpha = 1, \ldots, N). \tag{6}$$

Here, we assume that nuisance parameters of (6) are only $\boldsymbol{\beta}$. That is, when we use g_c of (5) as a link function, we consider c to be fixed. In order to test the null hypothesis H_0^g, we consider the family of ϕ-divergence statistics [13]

$$C_\phi = 2\sum_{\alpha=1}^{N} n_\alpha \left\{\hat{\pi}_\alpha^g \phi\left(\frac{\frac{Y_\alpha}{n_\alpha}}{\hat{\pi}_\alpha^g}\right) + (1-\hat{\pi}_\alpha^g)\phi\left(\frac{1-\frac{Y_\alpha}{n_\alpha}}{1-\hat{\pi}_\alpha^g}\right)\right\}, \tag{7}$$

where $\hat{\pi}_\alpha^g = \pi_\alpha(\hat{\boldsymbol{\beta}}^g)$, $(\alpha = 1, \ldots, N)$, $\hat{\boldsymbol{\beta}}^g = (\hat{\beta}_1^g, \ldots, \hat{\beta}_p^g)'$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under H_0^g given by (6) and $\phi(\cdot)$ is a real convex function in $(0, \infty)$, satisfying $\phi(1) = \phi'(1) = 0$ and $\phi''(1) = 1$. Here, we note that test statistics C_ϕ vary according to link function g. When we choose a convex function

$$\phi_a(t) = \begin{cases} \{a(a+1)\}^{-1}\{t^{a+1} - t + a(1-t)\} & (a \neq 0, -1) \\ t\log t + 1 - t & (a = 0) \\ -\log t - 1 + t & (a = -1) \end{cases}$$

as $\phi(t)$, C_{ϕ_a} becomes a power divergence statistic [5]

$$R^a = 2\sum_{\alpha=1}^{N} n_\alpha \left\{I^a\left(\frac{Y_\alpha}{n_\alpha},\ \hat{\pi}_\alpha^g\right) + I^a\left(1-\frac{Y_\alpha}{n_\alpha},\ 1-\hat{\pi}_\alpha^g\right)\right\}, \tag{8}$$

where

$$I^a(e,f) = \begin{cases} \{a(a+1)\}^{-1}e\left\{\left(\dfrac{e}{f}\right)^a - 1\right\} & (a \neq 0, -1) \\ e\log\left(\dfrac{e}{f}\right) & (a = 0) \\ f\log\left(\dfrac{f}{e}\right) & (a = -1). \end{cases}$$

Under H_0^g, all members of the class of statistics C_ϕ have a χ^2_{N-p} limiting distribution, assuming the condition that

$$n_\alpha/n \to \mu_\alpha, \quad (\alpha = 1, \ldots, N) \text{ as } n \to \infty, \tag{9}$$

where $n = \sum_{\alpha=1}^{N} n_\alpha$, $0 < \mu_\alpha < 1$, $(\alpha = 1, \ldots, N)$, and $\sum_{\alpha=1}^{N}\mu_\alpha = 1$. Using the results, we can use C_ϕ as a goodness-of-fit test statistic for model (1). Statistic R^0 (log likelihood ratio statistic or deviance) and statistic R^1 (Pearson's X^2 statistic) are used frequently.

In the case of the goodness-of-fit test for a multinomial distribution, Yarnold [22] obtained an approximation based on asymptotic expansion for the null distribution of Pearson's X^2 statistic. The expansion consists of a term of multivariate Edgeworth expansion for continuous distribution and discontinuous terms. Approximations based on asymptotic expansions for null distributions of some kinds of multinomial goodness-of-fit statistics have been investigated [16,14,10]. Edgeworth approximations of the distributions of some kinds of multinomial goodness-of-fit statistics under alternative hypotheses have also been investigated [18,17,15,12].

On the other hand, Taneichi et al. [19] obtained an approximation based on asymptotic expansion of the distribution of deviance for testing H_0^g given by (6) when link function g is defined by (2), that is, in a logistic regression model. Using the continuous term of the expression of the approximation, Taneichi et al. [19] proposed a Bartlett-type transformed statistic and showed that it improves the speed of convergence to a chi-square limiting distribution of the deviance.