Far East Journal of Theoretical Statistics
Volume 25, Number 1, 2008, Pages 83-94
Published Online: May 21, 2008
This paper is available online at http://www.pphmj.com

BARTLETT ADJUSTMENT FOR SPARSE BINARY DATA

MOHAMMAD REZA ZADKARAMI

Department of Statistics
Shahid Chamran University
Ahvaz, Iran
e-mail: zadkarami_m@scu.ac.ir

Abstract

In the present paper, we investigate the effect of sparseness of data on Bartlett adjustment. The choice of statistical tools for inference is highly dependent on the nature of the data. When the real data set is sparse, it creates some problems in using asymptotic theory. This property of the data restricts the use of statistical tools for inference, especially when the data set is extensive. The Bartlett adjustment factor improves the chi-square approximation to the distribution of the likelihood ratio statistic. We introduce Bartlett adjustment for binary response with probit link function. We find that the Bartlett adjustment does not improve the results when the binary data set is sparse.

Introduction

The Bartlett adjustment framework is focused on the chi-square approximation to the likelihood ratio and other test statistic criteria. In the theory of testing statistical hypotheses, it is well known that the likelihood ratio test statistic has approximately the χ^2 distribution for large samples, e.g., McCullagh and Nelder [15], Serfling [18], Azzalini [3] and Lehmann [14]. It is possible to improve this approximation by

2000 Mathematics Subject Classification: 62E15, 62J12.

Keywords and phrases: Bartlett adjustment, binary, likelihood ratio test, generalized linear models, sparse data.

Received December 15, 2007

multiplying the likelihood ratio test statistic (L.R.T.S.) by a scale factor, as first noted by Bartlett [5, 6]. Lawley [13] developed a general approach to approximating the null distribution of the likelihood ratio test statistic by finding a correction factor so that the null distribution of the modified statistic has the same moments as the reference, χ^2, chi-square distribution, ignoring terms of order, where n is the size of the sample.

In the last two decades, there has been renewed interest in Bartlett corrections leading to better approximations of the null distribution of the likelihood ratio statistic by a chi-square distribution, e.g., Lawley [13], Barndorff-Nielsen and Cox [4], Cordeiro and Paula [11], Cordeiro and Ferrari [10] and Cordeiro [8, 9]. Cordeiro [8] developed a Bartlett correction to improve the L.R.T.S. distribution for univariate generalized linear models (GLMs) (Nelder and Wedderburn [17]). He provided a general formula for the expected value of the likelihood ratio statistic for GLMs, corrected up to terms of order n^{-1}. His formula has advantages for numerical purposes because it requires only operations on matrices.

The effect of sparseness of the data on goodness-of-fit criteria in medical studies and other applied fields has been considered by several authors, e.g., Boyle et al. [7], Spiess and Hamerle [19] and Zadkarami [21]. The sparseness of data sets depends on a range of factors and models which may, or may not, be under the control of the researchers. Sparseness of the data creates some problems in using asymptotic theory and it is consequently computationally demanding, especially when the data set is extensive (Ainswort and Dean [1] and King and Zeng [12]). Sparseness in the data affects the shape of the log likelihood function so that it does not have a quadratic form. Therefore, from a theoretical point of view, this does not allow us to compare the change in the deviance between two nested models using the chi-square distribution (Zadkarami [21]). In this paper, we investigate the effect of sparseness of the data on the Bartlett adjustment.

The structure of the paper is as follows. We introduce the Bartlett adjustment in Section 1. The correction factors for GLMs are induced in Section 2. In Section 3, the empirical results are discussed and finally conclusion is presented in Section 4.

1. Bartlett Adjustment

Let ℓ denote the log likelihood function dependent on $p + q$ unknown parameter $\theta^T = (\theta_1, \theta_2, \ldots, \theta_{p+q})$ and $\ell^{(k)}$ denote the result of maximizing ℓ with respect to $\theta_1, \theta_2, \ldots, \theta_k$. The statistic denotes the likelihood ratio test for the hypothesis

$$H_0 : \theta_{p+1} = \theta_{p+1}^{(0)}, \ldots, \theta_{p+q} = \theta_{p+q}^{(0)},$$

where $\theta_1, \theta_2, \ldots, \theta_p$ are nuisance parameters. Lawley [13], by an exceedingly complicated calculation, has shown that, under fairly general regularity conditions,

$$E(2(\ell^{(k)} - \ell)) = k + \varepsilon_k + O(n^{-2}),$$

where

$$\varepsilon_k = \lambda^{rs}\lambda^{tu}\left\{\frac{1}{4}\lambda_{rstu} - (\lambda_{rst})_u + (\lambda_{rt})_{su}\right\} - (\varepsilon'_k + \varepsilon''_k),$$

$$\varepsilon'_k = \lambda^{rs}\lambda^{tu}\lambda^{vw}\left\{\lambda_{rtv}\left(\frac{1}{6}\lambda_{suw} - (\lambda_{sw})_u\right) + (\lambda_{rt})_v(\lambda_{sw})_u\right\},$$

$$\varepsilon''_k = \lambda^{rs}\lambda^{tu}\lambda^{vw}\left\{\lambda_{rtu}\left(\frac{1}{6}\lambda_{svw} - (\lambda_{sw})_v\right) + (\lambda_{rt})_u(\lambda_{sw})_v\right\},$$

$$\ell_r = \partial\ell/\partial\theta_r,$$

$$\ell_{rs} = \partial^2\ell/\partial\theta_r\partial\theta_s, \qquad \ell_{rst} = \partial^3\ell/\partial\theta_r\partial\theta_s\partial\theta_t, \text{ etc.}$$

$$\lambda_{rs} = E(\ell_{rs}), \qquad \lambda_{rst} = E(\ell_{rst}), \text{ etc.}$$

$$\ell_{rs} = \ell_{rs} - \lambda_{rs}, \qquad \ell_{rst} = \ell_{rst} - \lambda_{rst}, \text{ etc.}$$

$$(\lambda_{rs})_t = \partial\lambda_{rs}/\partial\theta_t, \qquad (\lambda_{rst})_u = \partial\lambda_{rst}/\partial\theta_u, \text{ etc.}$$

$$(\lambda_{rs})_{tu} = \partial^2\lambda_{rs}/\partial\theta_t\partial\theta_u, \tag{1}$$

where all subscripts are summed over the values 1 to k. Furthermore, he has shown that

$$E(2(\ell^{(p+q)} - \ell^{(p)})) = q + \varepsilon_{p+q} - \varepsilon_p + O(n^{-2}) = q(1 + b + O(n^{-2})),$$

where b is a constant of order n^{-1}. In fact, the statistic

$$W_l = 2q(\ell^{(p+q)} - \ell^{(p)})/(q + \varepsilon_{p+q} - \varepsilon_p)$$

has the same moments as χ^2 with q degrees of freedom, neglecting quantities of order n^{-1}.

2. Correction Factors for Generalized Linear Models

Let the random variables $Y_1, \ldots, Y_n$ be independent, and each Y_i has a density function of the form

$$f(y;\ \theta_i,\ \psi_i) = \exp\{\psi_i[y\theta_i - b(\theta_i) + c(y)] + d(\psi_i,\ y)\}, \tag{2}$$

where $b(.)$, $c(.)$ and $d(.)$ are known functions. The expectation and variance of Y_i can be expressed as

$$E(Y_i) = \mu_i = \partial b(\theta_i)/\partial\theta_i \quad \text{and} \quad \mathrm{Var}(Y_i) = V_i/\psi_i,$$

respectively, where $V_i = \partial\mu_i/\partial\theta_i$ is called the *variance function* and $\theta_i = q(\mu_i)$ is a *known function* of μ_i only. It is assumed that ψ_i is known for each observation and then the Y_i's have distributions belonging to the same one-parameter exponential family (McCullagh and Nelder [15]).

We assume that $X_i = (X_{i1}, \ldots, X_{ip})$ are the explanatory variables associated with individual i and $\mu_i = g(\eta_i)$ is a one-to-one function so that $\eta = X\beta$, where $\eta^T = (\eta_1, \ldots, \eta_n)$ and $X^T = (X_1, \ldots, X_n)$ is an $n \times p$ $(p < n)$ known matrix of the explanatory variables and $\beta^T = (\beta_1, \ldots, \beta_p)$ denotes a set of unknown model parameters to be estimated.

As Nelder and Wedderburn [17] showed, the estimation of β can be obtained by the maximum likelihood method which is equivalent to an iteratively re-weighted least squares procedure. Wedderburn [20] proved the finiteness, existence in the interior of the parameter space and uniqueness of the maximum likelihood estimates of β for various distributions and link functions.

2.1. Correction factors for generalized linear models

We now consider the issue of the Bartlett adjustment for GLMs. Cordeiro [8] provided the matrix form of the Bartlett adjustment, for GLMs based on Lawley [13]. Lawley showed that, under general regularity conditions,

$$E(2(\ell^{(p)} - \ell)) = p + \varepsilon_p + O(n^{-2}),$$

where ℓ denotes the log likelihood function and $\ell^{(p)}$ denotes the result of maximizing ℓ with respect to $\beta_1, \ldots, \beta_p$. The general matrix form for ε_p, which is obtained by Cordeiro [8], is given by

$$\varepsilon_p = \frac{1}{4}\operatorname{tr}(\Psi H Z_d^2) - \frac{1}{3}1^T \Psi G Z^{(3)}(F + G)1$$
$$+ \frac{1}{12}1^T \Psi F(2Z^{(3)} + 3Z_d Z Z_d)F\Psi 1, \tag{3}$$

where $Z = X(X^T W \Psi X)^{-1} X^T$ is an $n \times n$ positive semi-definite matrix of rank p with elements z_{ij} and $Z^{(3)}$ denotes an $n \times n$ matrix with elements z_{ij}^3, while W, H, F, G and Z_d are diagonal matrices of order n with elements which are given by

$W = \operatorname{diag}\{V^{-1}(\mu')^2\} = \operatorname{diag}\{w\}$,

$\Psi = \operatorname{diag}\{\psi\}$,

$H = \operatorname{diag}\{V^{-1}\mu''[\mu'' - 4w(\partial V/\partial\mu)] + w^2[2V^{-1}(\partial V/\partial\mu) - (\partial^2 V/\partial\mu^2)]\}$,

$F = \operatorname{diag}\{V^{-1}\mu'\mu''\}$,

$G = \operatorname{diag}\{V^{-1}\mu'\mu'' - V^{-2}(\partial V/\partial\mu)(\mu')^3\}$,

$Z_d = \operatorname{diag}\{z_{11}, \ldots, z_{nn}\}$,

$1 = (1, \ldots, 1)^T$ as an $n \times 1$ vector,

where $\mu' = \partial\mu/\partial\eta$ and $\mu'' = \partial^2\mu/\partial\eta^2$ denote the first and second derivatives of the link function g, respectively.

Deviance as a goodness-of-fit criterion for GLMs was suggested by Nelder and Wedderburn [17]. The expectation of the deviance of the model, D_p, considering equation, can be expressed as

$$\begin{aligned} E(D_p) &= 2E(\hat{\ell}_n - \hat{\ell}_p) \\ &= 2E(\hat{\ell}_n - \ell) + 2E(\ell - \hat{\ell}_p) \\ &= 2E(\hat{\ell}_n - \ell) - (p + \varepsilon_p) + O(n^{-2}), \end{aligned} \tag{4}$$

where, using equation (2), we have

$$2E(\hat{\ell}_n - \ell) = 2\phi \sum_{i=1}^{n} E\{v(y_i) - v(\mu_i)\},$$

where the sum is over all observations and $v(x) = xq(x) - b(q(x))$.

The deviance may be used to test the adequacy of a GLM. However, it is uninformative regarding lack of fit in the case of the binary distribution, see e.g., McCullagh and Nelder [15] and Aitkin et al. [2].

Furthermore, the deviance has an exact chi-square distribution for normal linear models. The lack of an exact theory for the deviance distribution depends upon its departure from normality. However, the deviance may not follow the chi-square distribution, even for approximately large n.

Moreover, the deviance is most directly useful, not as an absolute measure of goodness of fit, but for comparing two nested models. The chi-square approximation is usually quite accurate for a difference between deviances, which is equivalent to the likelihood ratio statistic, at least in large samples (Aitkin et al. [2], Cordeiro [9] and McCullagh and Nelder [15]).

The expectation of the likelihood ratio test statistic for the two nested models, based on equation (4), is given by

$$\begin{aligned} E(W_l) &= E(D_p - D_{p+q}) \\ &= (p + q) - p + \varepsilon_{p+q} - \varepsilon_p + O(n^{-2}) \end{aligned}$$

$$= q(1 + (\varepsilon_{p+q} - \varepsilon_p / q)) + O(n^{-2}),$$

where the Bartlett correction factor to test H_0 is denoted by $c = (1 + (\varepsilon_{p+q} - \varepsilon_p / q))^{-1}$.

2.2. Bartlett adjustment for binary responses

Let Y_i, conditional on the explanatory variables X, be a Bernoulli variable with probability function $f(y) = p^y (1-p)^{1-y} = \exp(y\theta - \log(1 + \exp(\theta)))$, where $y \in \{0, 1\}$ and $\theta = \log(p/(1-p))$. Comparing this with equation (2), we have $b(\theta) = \log(1 + \exp(\theta))$, $c(y, \psi) = 0$ and $\psi = 1$. Therefore, the matrix $\Psi = \mathrm{diag}\{1\} = I_{n \times n}$, the identity matrix, and we can write equation (3) in the form

$$\varepsilon_p = \frac{1}{4}\mathrm{tr}(HZ_d^2) - \frac{1}{3}1^T GZ^{(3)}(F + G)1 + \frac{1}{12}1^T F(2Z^{(3)} + 3Z_d ZZ_d)F1.$$

Zadkarami [21] showed that the probit link function produced the best results in comparison with the logit and complementary log-log link function. The assumption of the probit link function gives the mean as

$$E(Y \mid X) = \mu = p = \Phi(\eta),$$

where $\eta = \beta^T X$ denotes the linear predictor and Φ denotes the cumulative distribution function of the standard normal. Consequently, the first and second derivatives of the link function are given by $\mu' = \phi(\eta)$ and $\mu'' = -\eta\phi(\eta)$, respectively, where $\phi(.)$ denotes the density function of the standard normal distribution. Furthermore, the variance function can be expressed as $V = \partial\mu/\partial\theta = p(1-p)$ and $\partial V/\partial\mu = \partial V/\partial p = 1 - 2p = 1 - 2\mu$.

In this case, the structure of the matrices W, H, F and G, which is defined in Subsection 2.1, can be obtained by using equations, and so that

$$W = \mathrm{diag}\{(\phi(\eta))^2 / p(1-p)\} = \mathrm{diag}\{w\},$$

$$H = \mathrm{diag}\{w\eta(\eta + 4\phi(\eta)[(1-2p)/p(1-p)]) + 2w^2[((2p-1)^2/p(1-p)) + 1]\},$$

$F = \text{diag}\{-w\eta\},$

$G = \text{diag}\{-w(\eta + [\phi(\eta)(1-2p)/p(1-p)])\}.$

3. Empirical Results

The National Child Development Survey (NCDS) data set is used to investigate the effect of sparseness of binary data on Bartlett adjustment. This data set was collected on babies born in the same week (3-9 March 1958) in England, Wales and Scotland. The interval between 28th week (196 days) of gestation and 4th week after birth is divided into four subintervals, allowing for the possibility of death before delivery (antepartum stillbirth), during delivery (fresh stillbirth), in the first week after birth (early neonatal deaths), and between the first and fourth weeks after birth (late neonatal deaths). We call these stages 1, 2, 3 and 4, respectively. The survivors beyond stage 1 are further divided into two groups, those with assisted delivery and those with natural delivery to allow for the differential effects of the type of delivery in the two groups (Zadkarami [21]). We selected a sample of 10,141 individuals for whom we have complete information on forty variables associated with perinatal mortality. The data in stage 4 of the assisted delivery cases are selected as an illustration in the investigation of the effect of sparseness of data on the parameter estimates for the binary models. Because, our data set is extensive, it raises some computational problems. As an example, we cannot calculate the Bartlett adjustment for the natural delivery cases in our model because many matrices are involved in the calculation of ε_p. Those matrices need a large amount of computer memory, with the result that we cannot run the program for the natural delivery cases. Therefore, we use only the assisted delivery data at stage 4 as an example to demonstrate the effect of the Bartlett adjustment on the magnitude of the likelihood ratio test statistics as defined in Subsection 2.1. There are 1120 assisted delivery babies who are survived after stage 3. Four of them died during stage 4 and 1116 of babies are survived behind stage 4. The results in Table 1 indicate that "birthweight" is negatively significant in stage 4 for assisted delivery cases.

Table 1. The results of fitting model to stage 4 (assisted delivery)

Term		**Estimate (S.E.)**	***p*-value**
Intercept		- 1.98(5.96)	0.74
Birthweight		- 0.723(0.33)	0.03
Social class mother's husband	Re: Class I	-	-
	Class II-IV	- 0.82(0.52)	0.11
Baby birth order	Re: First baby	-	-
	2nd or later baby	2.47(5.98)	0.69
	Miscarriage	- 0.085(36.6)	0.99

We adjust the likelihood ratio test for the models M_1, M_2, M_3 and M_4 which are defined as

$$M_1 : \eta_1 = \beta_0,$$

$$M_2 : \eta_2 = \beta_0 + \beta_1 X_1,$$

$$M_3 : \eta_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

$$M_4 : \eta_4 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

where X_1 denotes the variable "birthweight", X_2 denotes the categorical variable "social class of mother's husband" which has two levels "class I" and "class II-IV" and X_3 denotes the categorical variable "baby birth order" which has three levels "first baby", "2nd or later baby" and "miscarriage". The Bartlett correction factor c, as defined in Subsection 2.1, is used to modify the magnitude of the likelihood ratio to give a better approximation to the chi-square distribution.

Table 2. The comparison of fitting different models to stage 4 (assisted delivery)

Comparison models	L.R.T.S.	D.F.	Bartlett factor	L.R.T.S. (modified)
M_1 and M_2	5.171	1	0.965	4.99
M_2 and M_3	2.227	2	0.835	1.859
M_3 and M_4	4.826	3	0.000162	0.0008

As the results in Table 2 show, the Bartlett adjustment performs well for the continuous explanatory variable. However, it does not yield an improvement for the discrete explanatory variables in the model. We found that the Bartlett adjustment does not improve the likelihood ratio test statistic for testing the discrete variables in the sparse data set.

4. Conclusion

This paper is concerned mainly with the effect of the sparseness of the data on the asymptotic theory, especially Bartlett adjustment. The Bartlett adjustment is used to improve the asymptotic theory for the homogeneous model. However, the empirical results showed that the Bartlett adjustment did not improve the likelihood ratio test statistic for testing the discrete explanatory variables for sparse data in the homogeneous binary models. Therefore, asymptotic theory cannot be applied to this type of model.

In fact, the sparseness of the data not only raises some theoretical problems but also creates some computational problems.

References

[1] L. M. Ainswort and C. B. Dean, Approximate inference for disease mapping, Comput. Statist. Data Anal. 50 (2006), 2552-2570.

[2] M. Aitkin, D. A. Anderson, B. Frances and J. P. Hinde, Statistical Modeling in GLIM4, 2nd ed., Oxford University Press, 2005.

[3] A. Azzalini, Statistical Inference, Based on the Likelihood, Chapman & Hall, 1996.

[4] O. E. Barndorff-Nielsen and D. R. Cox, Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator, J. Roy. Statist. Soc. Ser. B 46 (1984), 483-495.

[5] M. S. Bartlett, Properties of sufficiency and statistical tests, J. Roy. Statist. Soc. Ser. A 160 (1937), 268-282.

[6] M. S. Bartlett, A note on the, multiplying factors for various χ^2 approximations, J. Roy. Statist. Soc. Ser. B 16 (1954), 296-298.

[7] P. Boyle, F. Flowerdew and A. Williams, Evaluating the goodness of fit in models of sparse medical data: a simulation approach, Internat. J. Epidemiol. 26 (1997), 651-656.

[8] G. M. Cordeiro, Improved likelihood ratio statistics for generalized linear models, J. Roy. Statist. Soc. Ser. B 45 (1983), 404-413.

[9] G. M. Cordeiro, Performance of a Bartlett-type modification for the deviance, J. Stat. Comput. Simul. 5 (1995), 385-403.

[10] G. M. Cordeiro and S. L. P. Ferrari, Generalized Bartlett correction, Comm. Statist. Theory Methods 27 (1998), 509-527.

[11] G. M. Cordeiro and G. A. Paula, Improved likelihood ratio statistics for exponential family nonlinear models, Biometrika 1 (1989), 93-100.

[12] G. King and L. Zeng, Logistic regression in rare events data, Political Anal. 9 (2001), 137-163.

[13] D. N. Lawley, A general method for approximating to the distribution of likelihood ratio criteria, Biometrika 43 (1956), 295-303.

[14] E. L. Lehmann, Testing Statistical Hypotheses, 2nd ed., Chapman & Hall, 1994.

[15] P. McCullagh and J. A. Nelder, Generalized Linear Models, 2nd ed., Chapman & Hall, 1989.

[16] NAG, NAG Fortran Library Manual, The Numerical Algorithm Group Library Limited; 1, Mark 16 1993, Oxford, U. K.

[17] J. A. Nelder and R. W. M. Wedderburn, Generalised linear models, J. Roy. Statist. Soc. Ser. A 135 (1972), 370-384.

[18] R. J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.

[19] M. Spiess and A. Hamerle, A comparison of different methods for the estimation of regression models with correlated binary responses, Comput. Statist. Data Anal. 33 (2000), 439-455.

[20] R. W. M. Wedderburn, On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, Biometrika 63 (1976), 27-32.

[21] M. R. Zadkarami, Longitudinal data analysis: some of the statistical issues arising in the analysis of perinatal mortality, Unpublished Ph.D. Thesis, Lancaster University, 2000.

■