Journal of the Korean
Statistical Society
Vol. 23, No. 1, 1994

Score Tests for Overdispersion†

Choongrak Kim[1], Meeseon Jeong[1], Meeyeong Yang[1]

ABSTRACT

Count data are often overdispersed, and an appropriate test for the existence of the overdispersion is necessary. In this paper we derive a score test based on the extended quasi-likelihood and the pseudolikelihood after adjusting to the Bartlett factor. Also, we compare it with Levene(1960)'s F-type test suggested by Ganio and Schafer (1992).

KEYWORDS: Bartlett factor, Extended quasi-likelihood, Overdispersion, Pseudolikelihood, Score test.

1. INTRODUCTION

Count data are usually analyzed by the generalized linear models (McCullagh and Nelder 1989) such as the logistic regression model or the log-linear regression model. However, by the presence of the overdispersion, variability is often greater than what is predicted by a specified model. To incorporate the overdispersion, Efron (1986) proposed the double exponential family, Nelder and Pregibon (1987) suggested the extended quasi-likelihoods, and Carroll

[1] Department of Statistics, Pusan National University, Pusan, 609-735, Korea.
† This research was supported by the Ministry of Education (BSRI-93-116).

and Ruppert (1988) considered the pseudolikelihoods. Estimation of the dispersion parameter is done by Williams (1982) for extrabinomial variation, and by Breslow (1984) for extra-Poisson variation. Asymptotic relative efficiencies of various estimators are discussed by Firth (1987), Hill and Tsai (1988), and Kim et al. (1992). Tests for the existence of the overdispersion are studied by Cox (1983), Breslow (1990), and Ganio and Schafer (1992).

Ganio and Schafer (1992) recommended using the Levene (1960) type modifications of score tests throughout simulation studies. This paper presents another score test which is more powerful and governs the level well. The idea is based on forcing the expectation of the score function as close as zero by inserting the Bartlett adjustment factor. In Section 2, models that we will use are described. Score test statistics are derived for these models in Section 3. Simulations are done in Section 4, and concluding remarks are given in Section 5.

2. MODELS FOR DISPERSION

For comparisons and corrections we take the same models and notations as in Ganio and Schafer (1992). Assume that $Y_1, \ldots, Y_n$ are independent response variable with

$$E(Y_i) = \mu_i = h(\eta_i), \quad \eta_i = \mathbf{x}_i'\boldsymbol{\beta} \tag{1}$$

and

$$\text{var}(Y_i) = \phi_i V(\mu_i), \quad \phi_i = g(\gamma_i), \quad \gamma_i = \lambda + \mathbf{z}_i'\boldsymbol{\alpha} \tag{2}$$

where $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are p-vectors of known explanatory variables and unknown parameters, $h(\cdot)$ is link function, $V(\cdot)$ is variance function, ϕ_i is a dispersion parameter, $g(\cdot)$ is a twice-differentiable positive function, λ is a scalar parameter, and $\mathbf{z}_i$ and $\boldsymbol{\alpha}$ are q-vectors of explanatory variables and unknown parameters. Our interest is to derive a score test statistic for $H : \boldsymbol{\alpha} = \mathbf{0}$. If $\boldsymbol{\alpha} = \mathbf{0}$ then ϕ_i is a constant and (1) and (2) reduce to the generalized linear model under H.

Hypotheses about $\boldsymbol{\alpha}$ may be tested with likelihood methods such as double exponential family which is equivalent to the extended quasi-likelihood function. The log-likelihood is defined as

$$l_d(\boldsymbol{\mu}, \phi; \mathbf{y}) = -\frac{1}{2} \sum \left\{ \log \phi_i + \frac{D_i}{\phi_i} \right\} \tag{3}$$

where $D_i = D(y_i, \mu_i)$ is the i-th deviance. As noted by Ganio and Schafer (1992), (3) is the log-likelihood function obtained by treating

$$\text{sign}(y_i - \mu_i)\sqrt{D_i} \sim N(0, \phi_i).$$

As is well known, D_i is a biased estimator of ϕ_i, and thus bias corrected form is more accurate, i.e., we will use

$$\text{sign}(y_i - \mu_i)\sqrt{D_i} \sim N(0, k_i\phi_i) \tag{4}$$

where $k_i = 1 + b_i$, and b_i is a Bartlett adjustment factor. Analytic form of b_i for some distributions is given in McCullagh and Nelder (1989). Then, the log-likelihood corresponding to (4) becomes

$$l_d^*(\boldsymbol{\mu}, \phi; \mathbf{y}) = -\frac{1}{2} \sum \left\{ \log(k_i\phi_i) + \frac{D_i}{k_i\phi_i} \right\}. \tag{5}$$

Using a method of moments Carroll and Ruppert (1982,1988) suggested the pseudolikelihood defined as

$$l_p(\boldsymbol{\mu}, \phi; \mathbf{y}) = -\frac{1}{2} \sum \left\{ \log \phi_i + \frac{R_i}{\phi_i} \right\} \tag{6}$$

where $R_i = R(y_i, \mu_i) = (y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i)$ is the Pearson chi-square. Again, (6) is obtained by treating $(y_i - \mu_i)/\sqrt{V(\mu_i)} \sim N(0, \phi_i)$, and we will modify (6) based on the same idea as

$$l_P^*(\boldsymbol{\mu}, \phi; \mathbf{y}) = -\frac{1}{2} \sum \left\{ \log(k_i\phi_i) + \frac{R_i}{k_i\phi_i} \right\}. \tag{7}$$

3. DERIVATION OF SCORE TESTS

We derive score tests for (5) donoted by SD^*. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \lambda, \boldsymbol{\beta}')'$ and assume g is identity. Let $\dot{D}_i$ and $\ddot{D}_i$ be the first and second derivatives of

D_i with respect to μ_i. Let the score vector be $\mathbf{U}_\theta = (U'_\alpha, U_\lambda, U'_\beta)'$ and the information matrix be

$$\mathrm{I}(\theta) = \left[\begin{array}{c|c} \mathrm{I}_{11} & \mathrm{I}_{12} \\ \hline \mathrm{I}_{21} & \mathrm{I}_{22} \end{array}\right] = \left[\begin{array}{c|cc} \mathrm{I}_{\alpha\alpha} & \mathrm{I}_{\alpha\lambda} & \mathrm{I}_{\alpha\beta} \\ \hline \mathrm{I}_{\lambda\alpha} & \mathrm{I}_{\lambda\lambda} & \mathrm{I}_{\lambda\beta} \\ \mathrm{I}_{\beta\alpha} & \mathrm{I}_{\beta\lambda} & \mathrm{I}_{\beta\beta} \end{array}\right].$$

Then the score test for $H : \boldsymbol{\alpha} = \mathbf{0}$ is given by

$$DS^* = \hat{U}'_\alpha (\hat{\mathrm{I}}_{1\cdot 2})^{-1} \hat{U}_\alpha$$

where $\hat{U}_\alpha$ is U_α evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$ under H, and $\hat{\mathrm{I}}_{1\cdot 2} = \mathrm{I}_{11} - \mathrm{I}_{12}\mathrm{I}_{22}^{-1}\mathrm{I}_{21}$, also evaluated at the maximum likelihood estimate of $\boldsymbol{\theta}$ under H. Now, we need to compute U_α and $\mathrm{I}_{1\cdot 2}$.

$$U_\alpha = \frac{\partial l_d^*}{\partial \alpha} = \frac{\partial l_d^*}{\partial \phi_i} \cdot \frac{\partial \phi_i}{\partial \alpha} = -\frac{1}{2} \sum \left\{ \frac{1}{\phi_i} - \frac{D_i}{k_i \phi_i^2} \right\} \mathbf{z}_i$$

$$\mathrm{I}_{\alpha\alpha} = -E\left[\frac{\partial^2 l_d^*}{\partial \alpha^2}\right] = \frac{1}{2} \sum \mathbf{z}_i \mathbf{z}'_i / \phi_i^2$$

$$\mathrm{I}_{\alpha\lambda} = -E\left[\frac{\partial^2 l_d^*}{\partial \alpha \partial \lambda}\right] = \frac{1}{2} \sum \mathbf{z}_i / \phi_i^2$$

$$\mathrm{I}_{\lambda\lambda} = -E\left[\frac{\partial^2 l_d^*}{\partial \lambda^2}\right] = \frac{1}{2} \sum 1/\phi_i^2.$$

Also, it can be easily shown that $\mathrm{I}_{\alpha\beta}$ and $\mathrm{I}_{\lambda\beta}$ are approximately zero because $E(\dot{D}_i)$ is approximately zero. Therefore,

$$\begin{aligned}
\mathrm{I}_{1\cdot 2} &= \mathrm{I}_{\alpha\alpha} - (\mathrm{I}_{\alpha\lambda} \ \mathrm{I}_{\alpha\beta}) \begin{pmatrix} \mathrm{I}_{\lambda\lambda} & \mathrm{I}_{\lambda\beta} \\ \mathrm{I}_{\beta\lambda} & \mathrm{I}_{\beta\beta} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{I}_{\lambda\alpha} \\ \mathrm{I}_{\beta\alpha} \end{pmatrix} \\
&= \mathrm{I}_{\alpha\alpha} - (\mathrm{I}_{\alpha\lambda} \ 0) \begin{pmatrix} \mathrm{I}_{\lambda\lambda} & 0 \\ 0 & \mathrm{I}_{\beta\beta} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{I}_{\lambda\alpha} \\ \mathrm{I}_{\beta\alpha} \end{pmatrix} \\
&= \mathrm{I}_{\alpha\alpha} - \mathrm{I}_{\alpha\lambda}\mathrm{I}_{\lambda\lambda}^{-1}\mathrm{I}_{\lambda\alpha} \\
&= \tfrac{1}{2}\left[\sum \mathbf{z}_i \mathbf{z}'_i / \phi_i^2 - \{\sum \mathbf{z}_i / \phi_i^2\}\{\sum \mathbf{z}'_i / \phi_i^2\} / \sum \phi_i^{-2}\right].
\end{aligned}$$

Under H, ϕ_i is a constant, say ϕ, and let $\hat{D}_i = D_i(y_i, \hat{\mu}_i)$ be the i-th deviance statistic and $\hat{\phi} = \frac{1}{n}\sum \hat{D}_i/\hat{k}_i$ where $\hat{k}_i = 1 + \hat{b}_i$. Then,

$$DS^* = \frac{1}{2\hat{\phi}^2}\left\{\sum(\hat{\phi} - \frac{\hat{D}_i}{\hat{k}_i})\mathbf{z}_i\right\}' \left\{\sum \mathbf{z}_i\mathbf{z}_i' - \frac{1}{n}\sum \mathbf{z}_i \sum \mathbf{z}_i'\right\}^{-1} \left\{\sum(\hat{\phi} - \frac{\hat{D}_i}{\hat{k}_i})\mathbf{z}_i\right\}. \quad (8)$$

Note that the score test statistic by Ganio and Schafer (1992) based on (3) is

$$DS = \frac{1}{2\bar{D}^2}\sum(\hat{D}_i - \bar{D})\mathbf{z}_i'(\sum \mathbf{z}_i\mathbf{z}_i')^{-1}\sum(\hat{D}_i - \bar{D})\mathbf{z}_i \quad (9)$$

where $\bar{D} = \frac{1}{n}\sum \hat{D}_i$. In deriving DS in (9), they used $\sum \mathbf{z}_i = 0$, however, this condition is inconsistent with their simulation structure which will be stated in Section 4. Note that if $\hat{k}_i = 1$ then (8) and (9) are equivalent except the correction term $\sum \mathbf{z}_i \sum \mathbf{z}_i'/n$.

Similarly, we can derive the score test based on (7) as

$$PS^* = \frac{1}{2\hat{\phi}^2}\left\{\sum(\hat{\phi} - \frac{\hat{R}_i}{\hat{k}_i})\mathbf{z}_i\right\}' \left\{\sum \mathbf{z}_i\mathbf{z}_i' - \frac{1}{n}\sum \mathbf{z}_i \sum \mathbf{z}_i'\right\}^{-1} \left\{\sum(\hat{\phi} - \frac{\hat{R}_i}{\hat{k}_i})\mathbf{z}_i\right\} \quad (10)$$

where $\hat{R}_i = R(y_i, \hat{\mu}_i)$ and $\hat{\phi} = \frac{1}{n}\sum \hat{R}_i/\hat{k}_i$.

4. SIMULATIONS

We follow the same simulation structure as in Ganio and Schafer (1992) for comparisons. Observations were generated according to the logit model, $\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij}$, with $x_{ij} = i$ for $i = 1, \ldots, 5$ and $j = 1, \ldots, 8$. $z_{ij} = 1$ for j even and 0 for j odd, and with the following conditions ;

A. m and β's

(a) $m = 16$; $\beta_0 = -4, \beta_1 = 1$ (small sample size ; moderate p)

(b) $m = 16$; $\beta_0 = -4, \beta_1 = .6$ (small sample size ; small p)

(c) $m = 64$; $\beta_0 = -4, \beta_1 = 1$ (large sample size ; moderate p)

(d) $m = 64$; $\beta_0 = -4, \beta_1 = .6$ (large sample size ; small p)

B. distribution of Y

1. $Y \sim B(m, p)$

2. $Y|\tilde{p} \sim B(m, \tilde{p})$, $\tilde{p} \sim$ Beta with mean p and variance $p(1-p)/16$

3. $Y|\tilde{p} \sim B(m, \tilde{p})$, $\text{logit}(\tilde{p}) \sim N(\text{logit}p, 1/2)$
4. $Y|\tilde{p} \sim B(m, \tilde{p})$, $\tilde{p} \sim$ Beta with mean p and variance $p(1-p)z/16$
5. $Y|\tilde{p} \sim B(m, \tilde{p})$, $\text{logit}(\tilde{p}) \sim N(\text{logit}p, z/2)$

The test statistics used in this simulation are

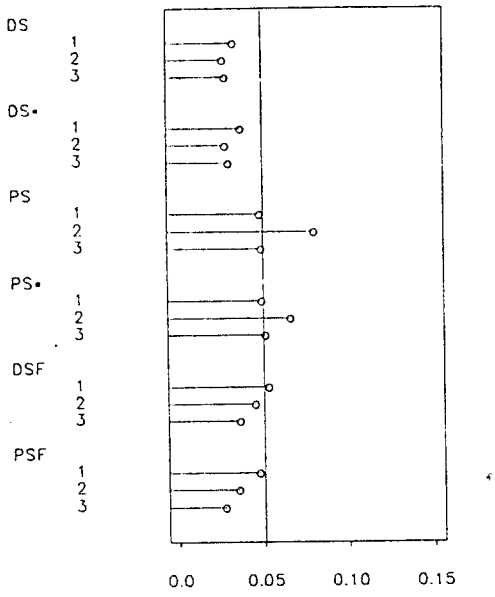
DS : score test under (3)
PS : score test under (6)
DSF : Levene test under (3)
PSF : Levene test under (6)
DS^* : score test under (5)
PS^* : score test under (7).

2000 replications are done for each set of conditions with nominal 5% level. Distributions 1, 2, and 3 are for the performance of level control (Figure 1), and 4 and 5 are for power (Figure 2). We obtain almost the same results for DS, PS, DSF, and PSF as in Ganio and Schafer (1992). They recommended using DSF or PSF to DS or PS because DS is quite conservative and PS is too liberal. In our simulation, PS^* is also liberal, while DS^* is quite satisfactory in most cases. For the performance of power, DS^* is slightly better than DS in all cases. Also, DS^* is better than DSF or PSF except when both the sample size and p are small (Figure 2(b)). Theoretically DS^* should be much better than DS in the level control and power. However, the improvement of DS^* over DS is not appreciable. We believe that this phenomenon stems from the poor adjustment of the Bartlett factor. The performance of the Bartlett factor was investigated by Kim and Jeong (1992) throughout the simulation study. Therefore, the refinement of the Bartlett factor such as second order expansion is necessary to improve DS^* up to the theoretical level.

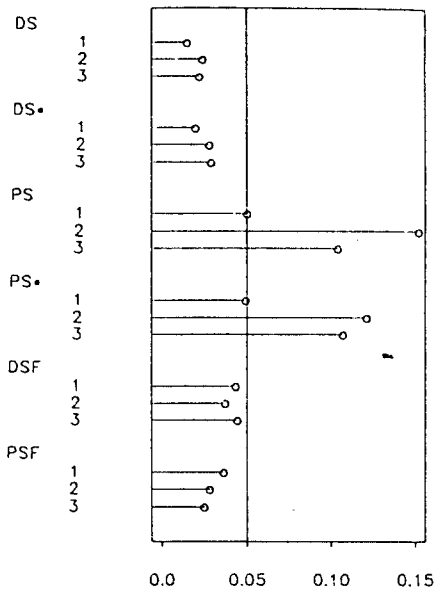
5. CONCLUDING REMARKS

Count data are often overdispersed, and an appropriate test for the existence of the overdispersion is necessary. In this paper we derived a score test based on the extended quasi-likelihood and the pseudolikelihood after adjusting to the Bartlett factor. Also, we compared them with other tests via the level control and power through the simulation study. In most cases, the adjusted score test was best. We can improve the adjusted score test by refining the Bartlett factor which will be a future research area.

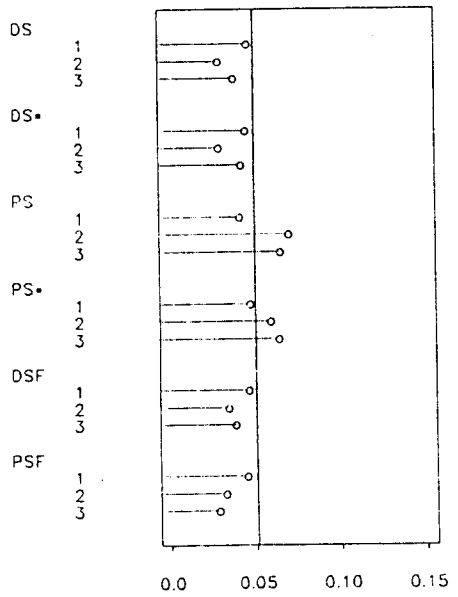
Figure 1. Simulation results. Type I error rates for nominal 5% tests based on 2000 simulated samples of size $n = 40$.



(a) $m = 16$, large p



(b) $m = 16$, small p



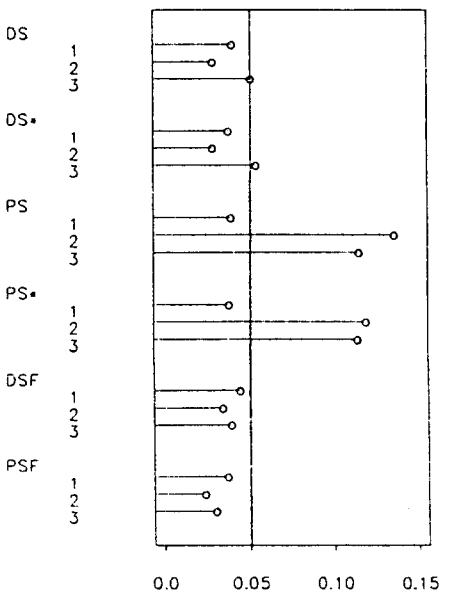
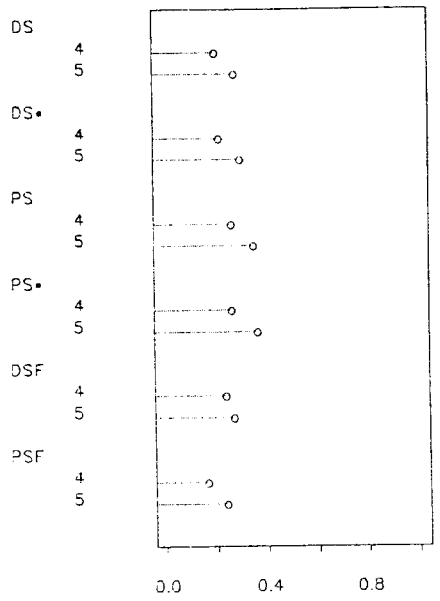
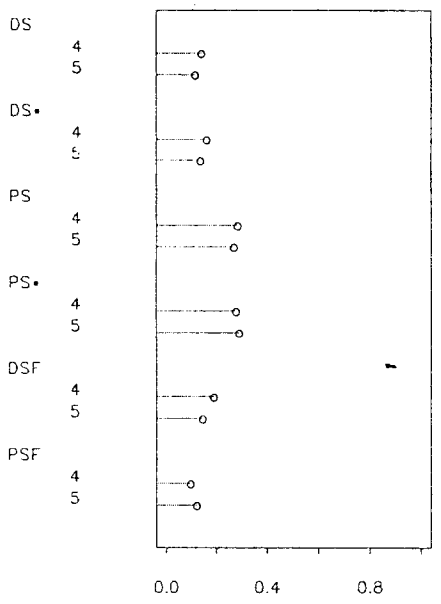


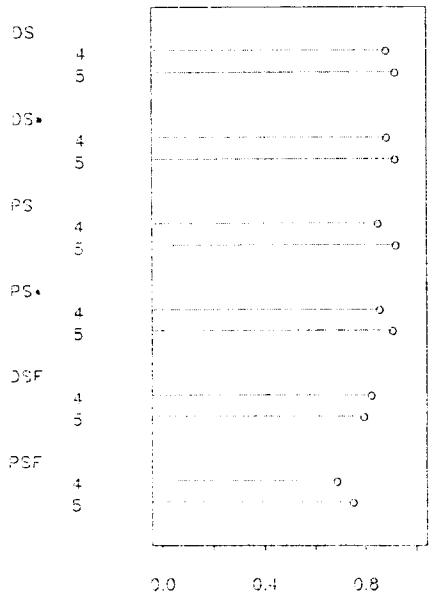
Figure 2. Simulation results. Power of 5% level tests under two nonnull conditions, based on 2000 simulated samples.

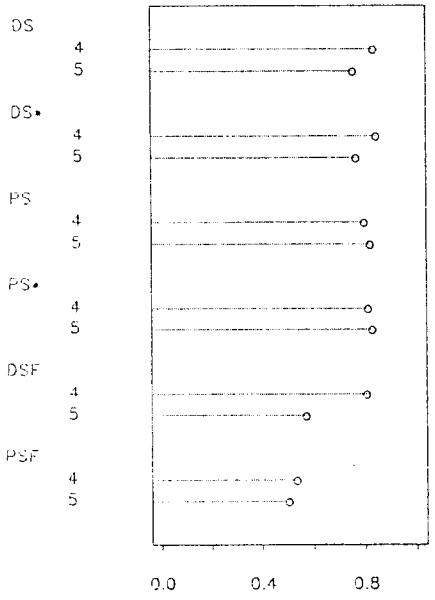


(a) $m = 16$, large p



(b) $m = 16$, small p





REFERENCES

(1) Breslow, N. E. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, **85**, 565-571.

(2) Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, **10**, 429-441.

(3) Carroll, R. J. and Ruppert, D. (1988). *Transformations and Weighting in Regression*, London : Chapman and Hall.

(4) Cox, D. R. (1983). Some remarks on over-dispersion. *Biometrika*, **70**, 269-274.

(5) Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **82**, 1079-1091.

(6) Firth, D. (1987). On the efficiency of quasi-likelihood estimation. *Biometrika*, **74**, 233-245.

(7) Ganio, L. M. and Schafer, D. W. (1992). Diagnostics for overdispersion. *Jounal of the American Statistical Association*, **87**, 795-804.

(8) Hill, J. R. and Tsai, C. (1988). Calculating the efficiency of maximum quasi-likelihood estimation. *Applied Statistics*, **37**, 219-230.

(9) Kim, C., Lee, K., Chung, Y., and Choi, K. (1992). Extended quasi-likelihood estimation in overdispersed models. *Journal of the Korean Statistical Society*, **21**, 187-200.

(10) Kim, C. and Jeong, M. (1992). Adjustments of dispersion statistics in extended quasi-likelihood models. *The Korean Journal of Applied Statistics.*

(11) Levene, H. (1960). "Robust tests for equality of variances" in *Contributions to Probability and Statistics.* Stanford University Press.

(12) McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London : Chapman and Hall.

(13) Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221-232.

(14) Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, **31**, 144-148.