응용통계연구 제 6권 1호
1993년 3월 pp. 41 - 52

Adjustments of Dispersion Statistics in Extended Quasi-likelihood Models[1)]

Choongrak Kim and Meeseon Jeong[2)]

Abstract

In this paper we study numerical behavior of the adjustments for the variances of the Pearson and deviance type dispersion statistics in two overdispersed mixture models; negative binomial and beta-binomial distribution. They are important families of an extended quasi-likelihood model which is very useful for the joint modelling of mean and dispersion. Comparisons are done for two types of dispersion statistics for various mean and dispersion parameters by simulation studies.

1. Introduction

Generalized linear models (Nelder and Wedderburn, 1972) have been widely used in regression modelling. Let the i-th response Y_i belongs to an exponential family of the canonical form

$$f(y;\theta,\phi)=\exp\{(y\theta-b(\theta))/a(\phi)+c(y,\phi)\}$$

for some specific function $a(\cdot), b(\cdot)$, and $c(\cdot)$, where θ is canonical parameter and ϕ is dispersion parameter. To specify the generalized linear models, we have to define a linear predictor $\eta_i = x_i{}'\beta$ and a link function $g(\cdot)$ such that $\eta_i=g(\mu_i)$, where $\mu_i=E(Y_i)$in addition to the distribution of Y_i. It is often, however, that we have limited information for the complete specification of the generalized linear model. To avoid this difficulty, Wedderburn (1974) suggested a quasi-likelihood model requiring the first two moments of the responses. To be specific, the variance of a response has been assumed to take the form

$$\mathrm{var}(Y_i)=\phi V(\mu_i)$$

where Y_i is a response, ϕ is dispersion parameter, and $V(\mu)$ is a known

[1)] This research was supported by the Ministry of Education (BSRI-91-116)

[2)] Department of Statistics, Pusan National University, Pusan, 609-735
Internet : crkim@hyowon.pusan.ac.kr

variance function. In the simplest of generalized linear models the dispersion parameter ϕ is a constant, usually unknown, however, there are many cases those ϕ's vary in a systematic way with other measured covariates. To come up with these Pregibon (1984) suggested joint model specification in terms of the dependence on covariates of the first two moments. For the mean we have the usual specification

$$E(Y_i)=\mu_i,\ \ \eta_i=g(\mu_i)=\sum_j x_{ij}\beta_j,\ \mathrm{var}(Y_i)=\phi_i V(\mu_i)$$

where η_i is the linear predictor, g is the link function, x_{ij} is the element of the $n\times p$ design matrix. Also, for the dispersion, it is assumed that

$$E(d_i)=\phi_i,\ \ \zeta_i=h(\phi_i)=\sum_j u_{ij}\gamma_j,\ \mathrm{var}(d_i)=\tau V_D(\phi_i).$$

In this specification, d_i is a suitable statistic chosen as a measure of dispersion; $h(\cdot)$ is the dispersion link function; ζ is the dispersion linear predictor, τ is dispersion parameter for d_i, and $V_D(\phi)$ is the dispersion variance function. The dispersion covariates u_i are commonly, but not necessarily, a subset of the regression covariates x_i.

Two possible choices for the dispersion statistic are the generalized Pearson contribution $r_p^2=(Y_i-\mu_i)^2/V(\mu_i)$ and the contribution to the i-th deviance

$$r_D^2=2\sum\{y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i)\}/a(\phi)$$

where $\tilde{\theta}_i$ and $\hat{\theta}_i$ are estimates of θ_i under the maximal model and current model, respectively. The pros and cons for the performance of r_p^2 and r_D^2 as goodness-of-fit measure in the generalized linear models were discussed by Pierce and Schafer (1986). If Y is normal d_i has $\phi_i\chi_1^2$ distribution, so that a gamma model with $V_D(\phi)=2\phi^2$would be chosen. If Y is non-normal, adjustments to the dispersion model may be necessary to account for the bias in r_D^2 or for the excess variability of r_p^2. Recently, McCullagh and Nelder (1990) suggested adjustments of the estimating equations for the dispersion parameters in each ways that

$$\mathrm{var}(r_p^2)=2\phi^2(1+\rho_4/2) \tag{1}$$

and

$$\mathrm{var}(r_D^2) = 2\phi^2(1+b)^2 \tag{2}$$

where $b=(5\rho_3^2-3\rho_4)/12$ is the Bartlett adjustment, and ρ_3 and ρ_4 denotes the standardized third and fourth cumulant, respectively. Note that the nominal variance is $2\phi^2$.

In this paper, we study numerical behavior of the adjustments for $\mathrm{var}(r_p^2)$ and $\mathrm{var}(r_D^2)$ in two over-dispersed mixture models; negative binomial distribution and beta-binomial distribution which are standard mixture of Poisson and binomial distribution when the overdispersion exist. Also, they are important classes of an extended quasi-likelihood models (Nelder and Pregibon, 1987). Definition and examples of overdispersion can be found in Cox(1983), Efron(1986), Jorgensen(1987) and Gelfand and Dalal(1990).

2. Model Specification

We assume that, conditional on the sampling mean θ_i, the data y_i have independent distributions belonging to a natural exponential family (NEF) with quadratic function (Morris, 1982, 1983), and that the means θ_i are independent with conjugate mixture (CM) distributions. Let $[\mu, V(\mu)]$ denote a distribution with mean μ and variance function $V(\mu)$.

2.1 Negative binomial distribution

When we have gamma-Poisson mixture, the resulting marginal distribution is negative binomial. To be more specific, let

$$y|\theta \sim \mathrm{Poisson}(\theta) = \mathrm{NEF}[\theta, \theta],$$

$$\theta \sim \Gamma\left(\frac{\mu}{\phi}, \phi\right) = \mathrm{CM}[\mu, \phi\mu],$$

$$\mathrm{V}(\mu) = \mu,$$

and

$$y \sim \mathrm{NB}\left(\frac{\mu}{\phi}, \frac{\phi}{1+\phi}\right) = \mathrm{marg}\,[\mu, (1+\phi)\mu]$$

where $\Gamma(\alpha,\beta)$ denotes a gamma distribution with parameters α and β, and $NB(\alpha,\beta)$ denotes a negative binomial distribution with probability function

$$p(y)=\binom{\alpha+y-1}{y}\beta^{y}(1-\beta)^{\alpha},\ y=0,1,\cdots$$

and $y\sim \text{marg}[\theta_1,\theta_2]$ denotes the random variable y has mean θ_1 and variance θ_2 marginally. In this set up, two types of dispersion statistics are

$$r_p^2=(y_i-\hat{\mu})^2/\hat{\mu}$$

and

$$r_D^2=-2\{(y_i\log\hat{\mu}-\hat{\mu})-(y_i\log y_i-y_i)\}.$$

Also, the adjusted variances for r_D^2 and r_p^2 given by McCullagh and Nelder (1990) are

$$\text{var}(r_p^2)=2(1+\phi)^2\left(1+\frac{1+\phi}{2\mu}\right)$$

and

$$\text{var}(r_D^2)=2(1+\phi)^2\left(1+\frac{1+\phi}{6\mu}\right)^2$$

respectively. Note that the nominal variance is just $2(1+\phi)^2$.

2.2 Beta-binomial distribution

A beta-binomial mixture can be specified as follows;

$$y|\theta\sim\frac{1}{m}\ \text{Binomial}(m,\theta)=\text{NEF}\left[\theta,\frac{\theta(1-\theta)}{m}\right]$$

$$\theta\sim\text{Beta}(\psi\mu,\psi(1-\mu))=\text{CM}[\mu,\phi\mu(1-\mu)]$$

where $\psi=1/\phi-1$, and the variance function is

$$V(\mu)=\mu(1-\mu).$$

Then, the marginal distribution of y becomes

$$y\sim\frac{1}{m}\ \text{BB}\ (m,\psi\mu,\psi(1-\mu))=\text{marg}\ [\mu,\mu(1-\mu)/w]$$

where $BB(m,\alpha,\beta)$ denotes a beta-binomial random variable with probability function

$$p(r)=\binom{m}{r}\frac{\Gamma(\alpha+\beta)\Gamma(r+\alpha)\Gamma(m+\beta-r)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(m+\alpha+\beta)},r=0,1,\cdots,m$$

and $w=m/\{1+(m-1)\phi\}$. In this situation,

$$r_p^2=(y_i-\hat{\mu})^2/\hat{\mu}(1-\hat{\mu})$$

and

$$r_D^2=-2\left\{y_i\log\left(\frac{\hat{\mu}}{1-\hat{\mu}}\right)+\log(1-\hat{\mu})-y_i\log\left(\frac{y_i}{1-y_i}\right)-\log(1-y_i)\right\},$$

and the corresponding adjusted variances are

$$\mathrm{var}(r_p^2)=\frac{2}{w^2}\left(1+\frac{1}{2mw}\frac{1-6\mu(1-\mu)}{\mu(1-\mu)}\right)$$

and

$$\mathrm{var}(r_D^2)=\frac{2}{w^2}\left(1+\frac{1}{6mw}\frac{1-\mu(1-\mu)}{\mu(1-\mu)}\right)^2.$$

Also, note that the nominal variance is $2/w^2$.

3. Simulation

In the negative binomial distribution simulations are done for $n=100,1000$; $\mu=1(1)10$; $\phi=0.1,0.2$, and 1000 replications are allowed. To be more specific, we generate a sample of size n for a given μ and ϕ from the IMSL library (GGDA), and repeat 1000 times to obtain true variances (TV; i.e., average of 1000 sample variances) of r_p^2 and r_D^2, and compared them with the nominal variance (NV) and their adjusted variances (AV). Figure 1 shows the simulation result for n=100, and Figure 2 shows them for n=1000. When n=100, r_D^2 is overly adjusted for $\mu\leq 2$, while r_p^2 is properly adjusted. When n=1000, r_p^2 is under-adjusted especially for $\phi=0.2$ case while r_D^2 is almost perfectly adjusted except $\mu\leq 2$. Simulation for other n, μ, and ϕ than listed showed similar results. Based on these results, we note that

i) True variances of r_P^2 and r_D^2 are away from their nominal variance $2\phi^2$, unless μ is large (at least $\mu \geq 5$ in our experience) regardless of the sample size n. Therefore, adjustment is necessary.

ii) For small sample size adjustment of r_P^2 is better than that of r_D^2, and the converse is true for the large sample size.

iii) As the overdispersion parameter ϕ increases, adjustment of r_D^2 shows better performance unless μ is too small.

For fixed sample size n=1000, simulations are done for m=5,10; μ=.05(.05).50 ; ϕ=.01,.02, and 1000 replications are allowed in the beta-binomial distribution. Simulation results are given in Figure 3 and 4 for m=5 and m=10, respectively. As shown in these Figures both adjustments for m=5 are too small to explain the actual variances. Adjustment of r_D^2 is so bad in the sense that the actual variance increases while the adjusted variance decreases. In r_P^2 case both decreasing, however, the adjustment should be made larger than it is. For m=10, the discrepancies become much smaller but similar phenomenon to m=5 case occur. We note that

i) Adjustment of r_D^2 is very poor.

ii) Adjustment of r_P^2 is smaller than it should be.

4. Remarks and Further Studies

Conclusively, adjustments for the variances of the dispersion statistics are necessary, but the existing adjustments are not good enough to be used safely in every situation. We have done simulations on other situations than listed in this paper, and they showed similar trend. It is therefore required that refined adjustments must be studied. Also, as pointed by referees, analytic form of adjustments might be possible and application to real data sets improve and rectify this study.

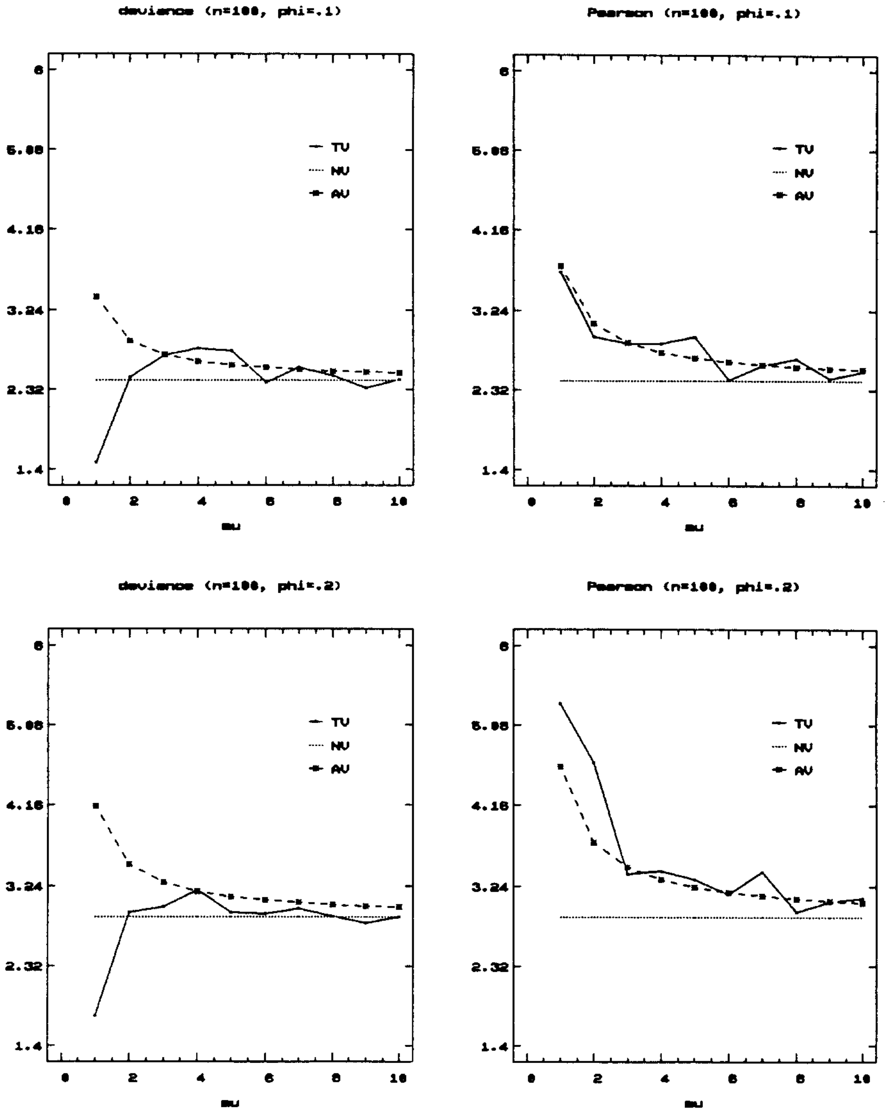

Figure 1. Simulations for the variances of dispersion statistics r_D^2 and r_p^2 with their nominal and adjusted variances with respect to μ when n=100, ϕ=.1, .2 in negative binomial distribution.

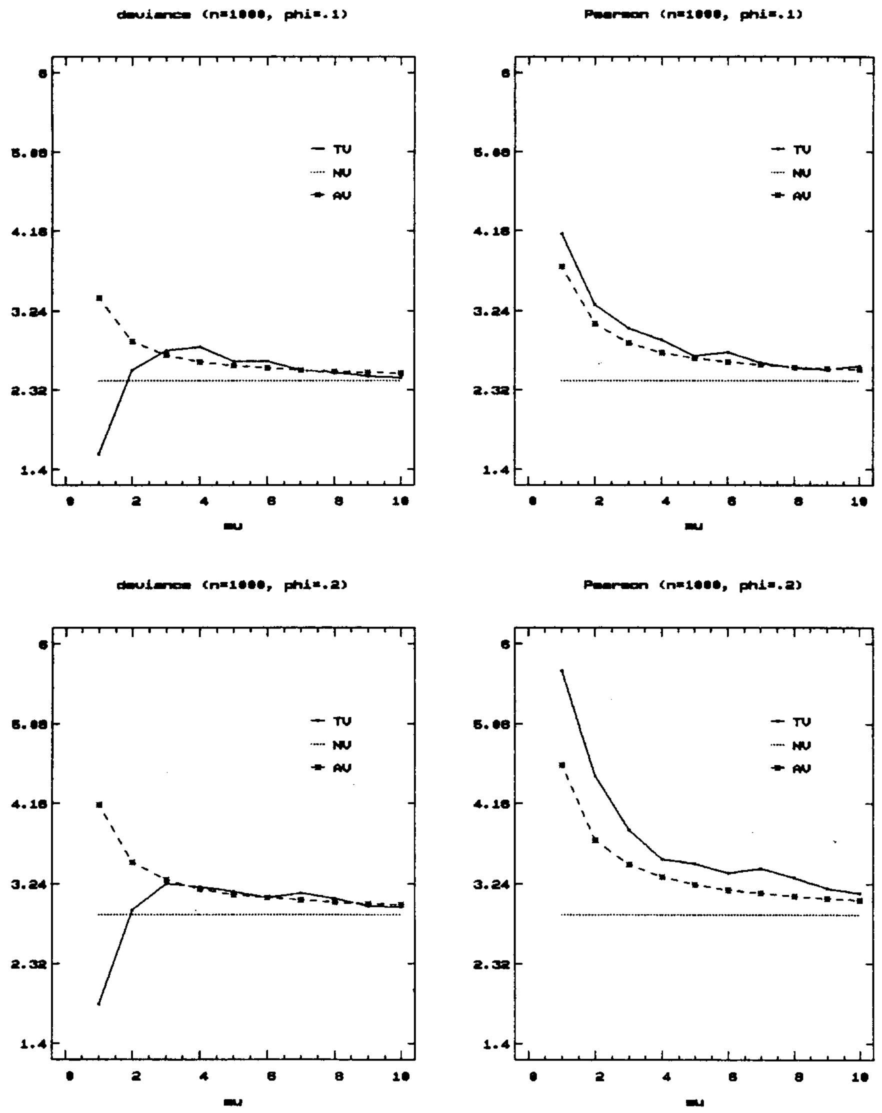

Figure 2. Simulations for the variances of dispersion statistics r_D^2 and r_P^2 with their nominal and adjusted variances with respect to μ when n=1000, ϕ=.1, .2 in negative binomial distribution.

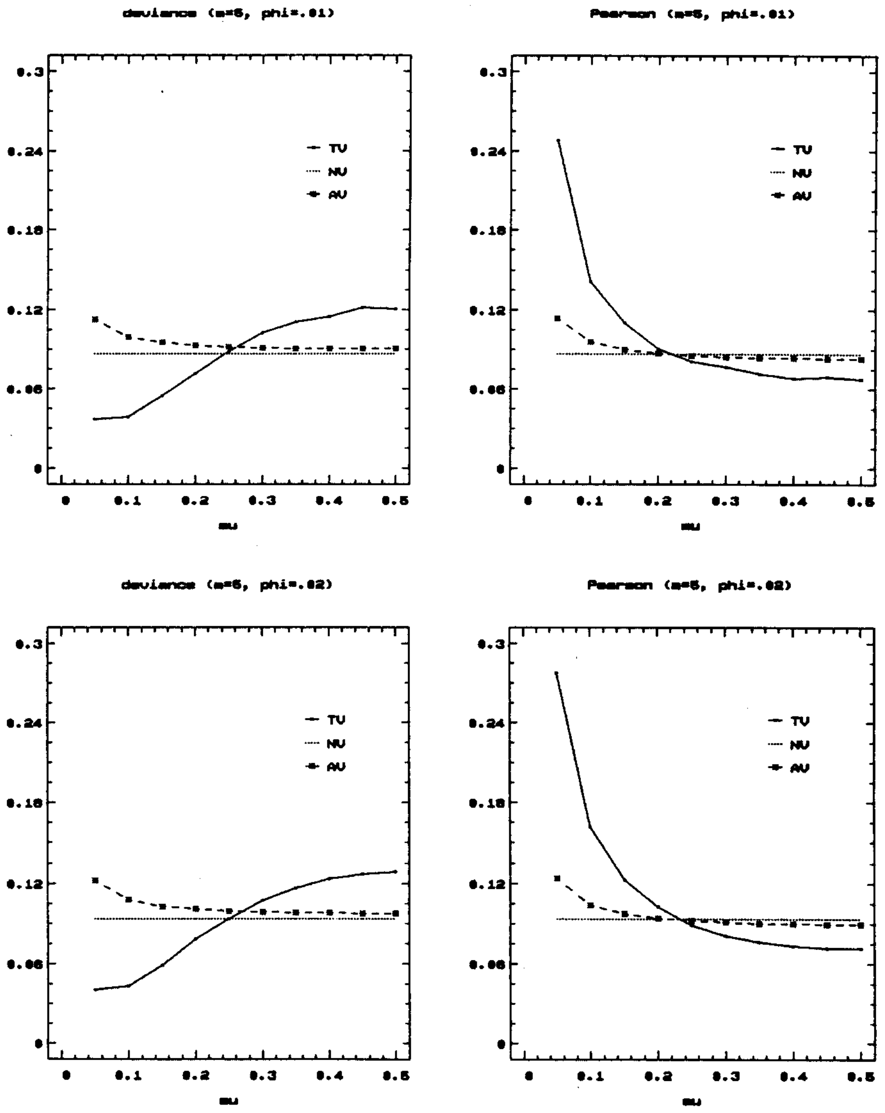

Figure 3. Simulations for the variances of dispersion statistics r_D^2 and r_p^2 with their nominal and adjusted variances with respect to μ when m=5, ϕ=.01, .02 in beta-binomial distribution.

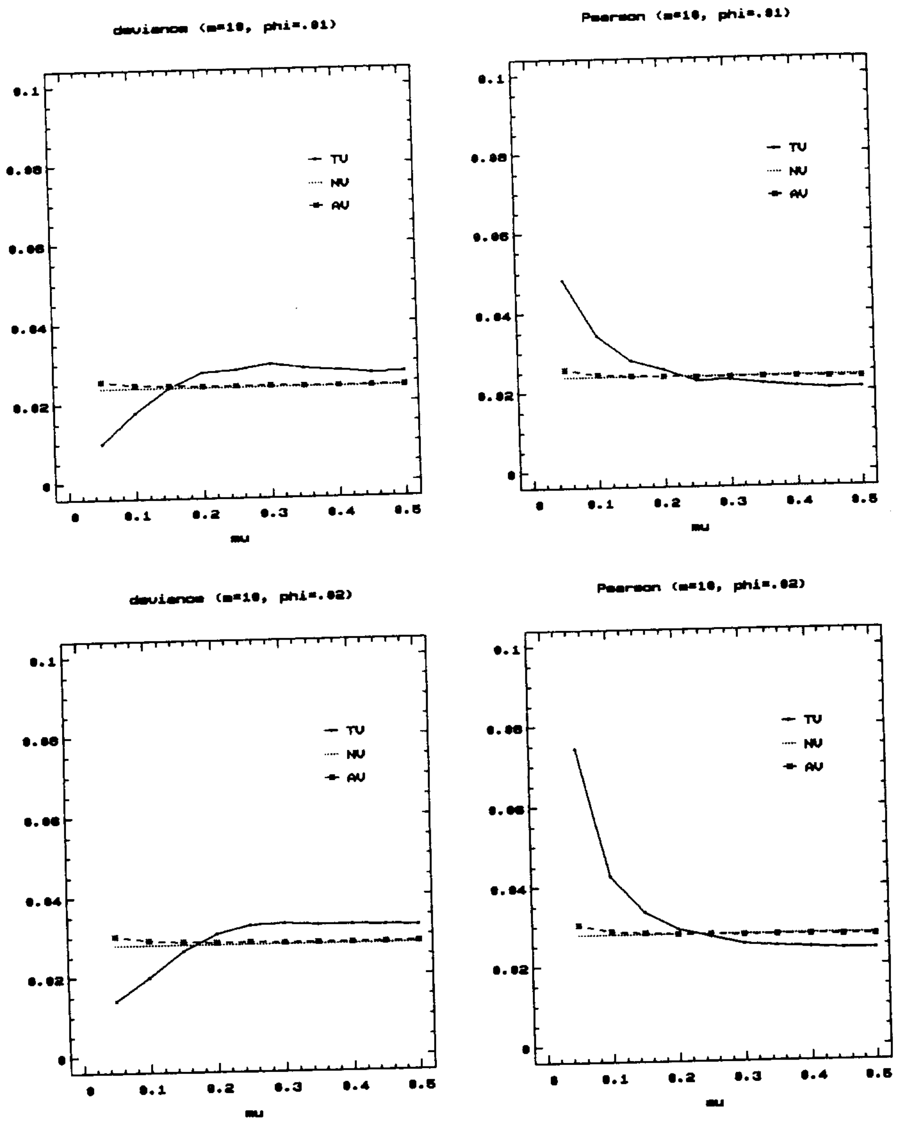

Figure 4. Simulations for the variances of dispersion statistics r_D^2 and r_P^2 with their nominal and adjusted variances with respect to μ when m=10, φ=.01, .02 in beta-binomial distribution.

References

[1] Cox, D.R. (1983), "Some remarks on overdispersion", *Biometrika*, **70**, 269-274.

[2] Efron, B. (1986), "Double exponential families and their use in generalized linear regression", *Journal of the American Statistical Association*, **81**, 709-721.

[3] Gelfand, A. E. and Dalal, S. R. (1990), "A note on overdispersed exponential families", *Biometrika*, **77**, 55-64.

[4] Jorgensen, B. (1987), "Exponential dispersion models (with discussion)", *Journal of the Royal Statistical Society*, B, **49**, 127-162.

[5] McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, J. Wiley and Sons, New York

[6] Morris, C. N. (1982), "Natural exponential families with quadratic variance functions", *The Annals of Statistics*, **10**, 65-80.

[7] Morris, C. N. (1983), "Natural exponential families with quadratic variance functions : statistical theory", *The Annals of Statistics*, **11**, 515-529.

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

[9] Nelder, J. A. and Wedderburn, R. W. M (1972), "Generalized linear models", *Journal of the Royal Statistical Society*, A, **135**, 370-384.

[10] Pierce, D. A. and Schafer, D. W. (1986), "Residuals in generalized linear models", *Journal of the American Statistical Association*, **81**, 977-986.

[11] Pregibon, D. (1984), "Review of Generalized Linear Models", *The Annals of Statistics*, **12**, 1589-1596.

[12] Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models and the Gauss-Newton method", *Biometrika*, **61**, 439-447.

준우도 함수의 분산치 교정[1]

김충락, 정미선[2]

요 약

본 논문에서는 과산포 혼합 모형인 음이항 분포와 베타이항 분포에서 피어슨 형태 및 데비언스 형태의 분산치 교정에 대한 효과를 수리적으로 비교했다. 이들 과산포 혼합 모형은, 평균과 분산을 동시에 모형화 하는데 매우 유용한 준우도함수의 중요한 구성원이다. 모의실험을 통해서 분산치의 교정이 평균, 산포모수에 따라 어떻게 달라지는지 비교 연구하였다.

[1] 본연구는 교육부 기초과학연구소 특성화 (BSRI-91-116) 연구비에 의해 수행되었음.

[2] (609-735) 부산직할시 금정구 장전동 부산대학교 통계학과