The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 269-276

Estimating Parameters in Overdispersed Binary Data[1)]

Sunho Lee [2)]

Abstract

There are several methods available for estimating parameters in overdispersed binary response data with the litter effect. Simulations are performed to compare methods for estimating an overall mean and an overdispersion parameter using moments, a maximum likelihood under a beta-binomial distribution, a maximum quasi-likelihood and a maximum extended quasi-likelihood.

1. Introduction

In toxicology and other biomedical research areas, it is common to encounter littermate data in the form of binary responses. Suppose there are k litters and each litter contains varying numbers of fetuses, n_i, each of which is classified as having some fatal abnormalities or not. The ideal method of analysis of binary data is to assume $X_i \sim B(n_i, p)$, a binomial random variable with parameters n_i and p, where X_i is the number of abnormal fetuses among n_i fetuses in the ith litter $(i=1, \cdots, k)$ and p is the proportion of abnormal fetuses. But the binary responses of animal litters from reproductive experiments often exhibit variation greater than predicted by a simple binomial model. This kind of overdispersion is generally recognized as a litter effect, the tendency for animals from the same litter to respond more alike than animals from different litters. When this overdispersion is not appropriately taken into account, it has little effect on estimation of the mean(Cox, 1983), but standard errors, tests, and confidence intervals may be seriously in error. Thus, it is preferable to assume that within each litter the binary responses form a set of Bernoulli trials whose abnormal proportion is a random variable from some distributions with a mean μ and a variance σ^2.

1) This research is supported by a Korea Science and Engineering Foundation Grant, 1998 (Project No. 981-0105-027-1).

2) Assistant Professor, Department of Applied Mathematics, Sejong University, Seoul, 143-747, Korea.

Let $R_i = X_i/n_i$ and $\sigma^2 = \mu(1-\mu)\phi$ where ϕ is a nonnegative dispersion factor relative to the binomial. Then the unconditional mean of R_i is

$$E(R_i) = \mu \tag{1}$$

and the variance of R_i is

$$Var(R_i) = \mu(1-\mu)/n_i + (n_i-1)\mu(1-\mu)\phi/n_i, \tag{2}$$

the terms of a binomial variance component and an extra binomial component with an overdispersion parameter ϕ.

In the analysis of a biomedical effect, estimating the proportion of a certain abnormality and testing homogeneity are fundamental works to do(Paul and Islam, 1995). Also, if we ignore the difference of the litter sizes, the analysis will be flawed. Therefore, in the presence of overdispersion with unequal litter sizes, estimating an overall mean μ and an overdispersion parameter ϕ are very important.

Various techniques for estimating parameters are discussed in this paper. When we are not willing to make assumptions about the form of the underlying distribution of a proportion, the method of moments will be a good choice. For a completely parametric method, the proportion of each litter is further assumed to be a beta random variable and parameters can be estimated by maximum likelihood. However, this method has a defect when the variance model is misspecified (Kupper et al. 1986). Instead of the full distributional assumption about a proportion, it is desirable to use the quasi-likelihood of a weaker assumption with only the first and second moments.

2. Method of Moments

To find point estimators of μ and ϕ, the first two moments of the distribution are equated to the corresponding moments of the sample. When litter sizes are all equal, it is easy to find moment estimators. But, in an unequal case, it causes a problem of weighting to obtain good estimators. Kleinman(1973) showed the empirical weighting moment method by weighting each R_i as the inverse of its variance. Let

$$\overline{R} = \sum w_i R_i \text{ and } S = \sum w_i (R_i - \overline{R})^2$$

be equal to their expected values and solve the equations for μ and ϕ with weights $\{w_i\}$. Moment estimators are obtained as follows:

$$\hat{\mu} = \overline{R} \tag{3}$$

and

$$\hat{\phi} = \frac{S - \overline{R}(1-\overline{R})\sum\{w_i(1-w_i)/n_i\}}{\overline{R}(1-\overline{R})\{\sum w_i(1-w_i) - \sum w_i(1-w_i)/n_i\}} \quad . \tag{4}$$

When the weights are inversely proportional to the variance of R_i such as

$$w_i = \frac{\frac{n_i}{1+(n_i-1)\phi}}{\sum_l \frac{n_l}{1+(n_l-1)\phi}},$$

a linear unbiased estimate of μ with a minimum variance is obtained. But ϕ is rarely known in practice. In this case, starting either with $w_i = 1/k$ or $w_i = n_i / \sum n_i$ is recommended and then estimates of μ and ϕ can be obtained from (3) and (4). With these values, one can calculate the empirical weights and find new estimates of μ and ϕ. If a negative $\hat{\phi}$ is obtained, it must be set to zero. The efficiency of the moment estimator of ϕ relative to maximum likelihood is often low(Kleinman, 1973), but the computation is very simple. Moore(1986) has shown that the moment estimates have the desirable properties of consistency and asymptotic normality under reasonable conditions.

3. Maximum likelihood method under a beta binomial model

For describing the variation of the proportion parameter in a binomial distribution, the beta distribution, which exhibits varieties of shapes and variation on the unit interval, is widely used. It can be bell-shaped, U-shaped, J-shaped, reverse J-shaped or a straight line according to two parameters.

Suppose that $X_i \sim B(n_i, P_i)$, where P_i is a beta random variable satisfying a mean μ and a variance $\mu(1-\mu)\theta/(1+\theta)$. Then unconditionally, X_i is a random variable having a beta-binomial distribution and the probability density function can be written as

$$P(X_i = x_i) = \frac{n_i! \prod_{r=0}^{x_i-1}(\mu + r\theta) \prod_{r=0}^{n_i-x_i-1}(1-\mu+r\theta)}{x_i!(n_i-x_i)! \prod_{r=0}^{n_i-1}(1+r\theta)}.$$

With $\phi = \theta/(1+\theta)$, the mean and variance of R_i are the same as (1) and (2), and the binomial model is a special case of a beta binomial model when $\phi = 0$.

The likelihood score equations of μ and θ for estimating μ and $\phi = \theta/(1+\theta)$ can be solved iteratively and Smith(1983) suggested the algorithm using a damped Newton-Raphson method with moment estimates as initial values.

Among several parametric models for analyzing overdispersed binary data, the above model already has been shown to provide a much better fit than a simple binomial model(Paul, 1982; Nakashima and Ohtaki, 1994) and used in a dominant lethal study(Aeschbacher et al. 1987). Also Cox(1983) had shown that the maximum likelihood method retains high efficiency for modest amounts of overdispersion, even if the overdispersion is not explicitly accounted for in

a parametric model.

4. Quasi-likelihood method

The quasi-likelihood is a semi-parametric model with assumptions for only the form of the first two moments of a random variable. Let $R_1, R_2, \cdots, R_k$ be binomial outcomes satisfying (1) and (2). Then the quasi-likelihood is given by $Q = \sum_{i=1}^{k} Q(R_i, \mu, \phi)$, where

$$Q(R_i, \mu, \phi) = \int_{R_i}^{\mu} n_i(R_i - t)/[t(1-t)\{1+(n_i-1)\phi\}]dt$$

and the quasi-likelihood score equation is defined,

$$U(\mu) = dQ/d\mu = \sum n_i(R_i - \mu)/[\mu(1-\mu)\{1+(n_i-1)\phi\}] = 0 . \quad (5)$$

When ϕ is known, a maximum quasi-likelihood estimator of μ is the solution of the estimating equation, $U(\hat{\mu}) = 0$, and can be obtained by the iterated reweighted least squares algorithm. But, in practice, an overdispersion parameter ϕ is usually unknown and it must be estimated. Two methods are proposed for obtaining a reasonable estimator of ϕ.

4.1 Quasi-likelihood/Method of moments(QL/M)

For estimating an overdispersion parameter ϕ, a conventional moment equation, which equates the Pearson chi-squared statistic $\sum\{(R_i - \hat{\mu})^2 / Var(R_i)\}$ to its expected value can be adopted:

$$\sum \frac{n_i(R_i - \hat{\mu})^2}{\hat{\mu}(1-\hat{\mu})\{1+(n_i-1)\phi\}} = k-1 . \quad (6)$$

QL/M estimates $\hat{\mu}$ and $\hat{\phi}$ can be obtained as the joint solution of (5) and (6). If the assumed model is correct, $(\hat{\mu}, \hat{\phi})$ is consistent and $\sqrt{k}((\hat{\mu}, \hat{\phi}) - (\mu_0, \phi_0))$ has a limiting multivariate normal distribution, where (μ_0, ϕ_0) is the true value of the parameters(Moore, 1986). Even though the variance function is misspecified, $\hat{\mu}$ is still asymptotically consistent Moore and Tsiatis, 1991).

Note that the QL/M method allows the estimation of only one single dispersion parameter and it can be generalized to more than one using pseudo-likelihood(Carroll and Ruppert, 1982). And QL/M does not need more information than the second order moment, and is, therefore, robust.

4.2 Extended quasi-likelihood(EQL) method

Quasi-likelihood behaves like an ordinary log likelihood with respect to μ, but not with

respect to ϕ. To obtain simultaneous estimation of μ and ϕ, the EQL, which is essentially the same as the quasi-likelihood Q for known ϕ and exhibits the properties of a log likelihood with respect to ϕ-derivative, is introduced by Nelder and Pregibon(1987). EQL with a mean and a variance specified by (1) and (2) is

$$Q^+(\mu,\phi) = Q - \sum \log(1+(n_i-1)\phi)/2 \ ,$$

ignoring constants involving only the observations. The estimates are obtained by maximizing EQL. Taking derivatives with respect to μ, $dQ^+/d\mu$ yields the quasi- likelihood score function. Similarly, $dQ^+/d\phi$ leads to the use of the deviance for estimating ϕ.

For finding joint solutions in the above QL/M and EQL methods, the usual iterative weighted least squares algorithm, which alternately iterates between holding μ fixed to estimate ϕ, and holding ϕ fixed to get a new estimate μ, will lead to estimates $\hat{\mu}$ and $\hat{\phi}$ maximizing the likelihood function. It is computationally much less intensive than the maximum likelihood method of a beta-binomial model.

Davidian and Carroll(1987) pointed out that estimating equations based on QL/M are unbiased, and hence consistency and asymptotic normality are obtained under very general conditions. Also, they discussed the inconsistency of the EQL estimators. However, the study by Nelder and Lee(1992) showed that EQL could often be superior.

5. Simulation and Results

Simulations are performed for various combinations of the overall mean, $\mu =$ 0.1, 0.2, 0.5, an overdispersion parameter, $\phi =$0.0, 0.01, 0.05, 0.1, 0.2, 0.3 and the number of litters, $k=10$, 27, 50. Table I shows the litter sizes used in this simulation.

Table I. Data of litter sizes

Group	Litter sizes
$k=10$	5 6 7 7 8 8 8 9 9 10
$k=27$	1 2 4 6 6 6 7 7 7 7 7 7 7 8 8
	8 8 8 9 9 10 10 10 11 11 12 12

For $k=10$, a moderate spread of litter sizes are chosen from the control group($k=10$) in Kupper and Haseman(1978) and for $k=27$, litter sizes of a wide range between 1 and 12 are chosen from the control group($k=27$) in Paul and Islam(1995). And, for $k=50$, each half of the litters are set as sizes of 10 and 25. Although the form of litter size may or may not affected by the estimating method, it is not discussed in this paper.For each combination of

these elements, the IMSL random number generator generates 10000 samples and (μ, ϕ) are estimated by the previously mentioned methods of the moments(MM), maximum likelihood(ML), QL/M and EQL. These four methods are compared in terms of bias and mean squared error(MSE) in Table II.

Consider the estimators of μ. All methods give very similar results, except when there is a small overdispersion corresponding to binomial distribution. When ϕ is near to zero, the ML estimator, which is derived from the wrong assumption of a beta-binomial, is the least efficient, as expected. When μ and ϕ are both near to zero, maximization with boundary conditions makes ML estimators for μ and ϕ worse.

For estimating ϕ, EQL performs remarkably well in the small overdispersion and a small number of litters(similar results have been reported by Nelder and Lee(1992)). But, as both number of litters and overdispersion get larger, the EQL method becomes worse because of the biased estimating equation and it does not warrant further consideration. Except in the small overdispersed case, the ML method always provides a good estimator of ϕ and MSE of the ML estimator is almost uniformly smaller than those of the other estimators. The QL estimator is not recommendable for an overdispersion parameter, but its bias is small compared with others in a large overdispersion. Usually, the larger values of ϕ are underestimated by the MM and ML methods.

6. Discussions

From the limited comparison, no definite advantage of one estimating method for (μ, ϕ) over the other has been found. For small overdispersion under the small and moderate number of litters, the EQL method performs best, and the ML method is the worst especially in terms of bias and MSE of μ. In the other cases, the ML method is superior to others. However, it has apparent drawbacks in that the computation is difficult and the initial values of the starting points are needed. Although the MM method is not that superior to the ML method, it always provides reasonable estimators over a wide range of conditions, even in the small overdispersion case. In addition, MM has much of its own merits in terms of computational ease, efficiency, robustness, and the desirable asymptotic properties. Therefore, the MM method should be the first choice to estimate parameters when we have the binary response data without any ideas about the amount of overdispersion and the form of overdispersion parameters.

Bibliography

[1] Aeschbacher, H.U., Milon, H. and Wurzner, H.P.(1978). Caffeine concentrations in mice plasma and testicular tissue and the effect of caffeine on the dominant lethal test,

Mutation Research, **57**, 193-200.

[2] Carroll, R.J., and Ruppert, D.(1982). Robust estimation in heteroscedastic linear models, *Annals of Statistics,* **10**, 429-41.

[3] Cox, D.R.(1983). Some remarks on overdispersion, *Biometrika,* **70**, 269-74.

[4] Davidian, M., and Carroll, R.J.(1987). Variance function estimation, *Journal of American Statistical Association,* **82**, 1079-91.

[5] Kleinman, J.C.(1973). Proportions with extraneous variance: single and independent samples, *Journal of American Statistical Association,* **68**, 46-54.

[6] Kupper, L., Portier, C., Hogan, M. and Yamamoto, E(1986). The impact of litter effects on dose response modeling in teratology, *Biometrics,* **42**, 85-98.

[7] Kupper, L. and Haseman, J.K. (1978). The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics,* **35**, 281-93.

[8] Liang, K.Y. and McCullagh, P.(1993). Case studies in binary dispersion, *Biometrics,* **49**, 623-30.

[9] Moore, D.(1986). Asymptotic properties of moment estimators for overdispersed counts and proportions, *Biometrika,* **73**, 583-88.

[10] Moore, D. and Tsiatis, A.(1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation, *Biometrics,* **47**, 383- 401.

[11] Nakashima, E. and Ohtaki, K.(1994). Two methods for the analysis of chromosome aberration data from the atomic-bomb survivors: quasi-likelihood moment method and beta-binomial method, *Journal of Japanese Statistical Society,* **24**, 209-219.

[12] Nelder, J.A. and Lee, Y.(1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons, *Journal of Royal Statistical Society,* Series B, **54**, 273-84.

[13] Nelder, J.A. and Pregibon, D.(1987). An extended quasi-likelihood function, *Biometrika,* **74**, 221-32.

[14] Paul, S. (1982). Analysis of proportions of Affected Foetuses in teratological experiments, *Biometrics,* **38**, 361-70.

[15] Paul, S. and Islam, A.(1995). Analysis of proportions in the presence of over-/ under-dispersion, *Biometrics,* **51**, 1400-10.

[16] Smith, D.M.(1983). Maximum likelihood estimate of the parameters of the beta-binomial distribution, *Applied Statistics,* **32**, 196-204.

Table II. Bias and MSE for estimators of μ and ϕ

μ	ϕ	variance	10^3 x bias and MSE of μ								10^3 x bias and MSE of ϕ							
			MLE		MM		QL/M		EQL		MLE		MM		QL/M		EQL	
number of litters k = 10																		
0.1	0.0	0.009	-31.77	3.93	2.51	1.15	2.61	1.15	2.35	1.16	51.51	5.42	53.88	5.53	73.22	6.98	47.53	3.09
	0.01	0.010	-25.02	3.44	3.97	1.25	4.04	1.25	3.79	1.25	48.00	5.55	48.83	5.43	68.98	8.90	38.96	2.36
	0.05	0.013	-8.45	2.65	7.39	1.60	7.43	1.60	7.19	1.60	35.53	7.01	31.56	6.31	55.21	10.02	3.30	1.05
0.2	0.05	0.023	-16.11	5.23	5.83	2.76	5.88	2.76	5.73	2.76	30.97	5.52	32.04	5.51	56.16	9.07	19.97	1.56
	0.1	0.030	-5.50	4.40	5.96	3.37	6.00	3.38	5.84	3.38	15.05	7.85	13.24	7.70	41.64	11.41	-25.60	1.82
	0.3	0.059	2.86	6.10	4.85	6.09	4.87	6.09	4.75	6.10	-24.54	24.27	-33.92	25.77	12.77	30.86	-221.40	50.54
0.5	0.0	0.025	0.38	5.91	0.02	3.18	0.03	3.19	0.04	3.20	50.99	4.48	66.01	6.34	88.59	10.54	88.62	11.30
	0.1	0.048	-0.28	6.38	0.26	5.38	0.25	5.38	0.24	5.39	9.44	6.17	14.01	6.28	42.78	9.65	53.34	13.56
	0.3	0.093	0.39	9.90	-0.03	9.65	-0.04	9.65	-0.03	9.65	-27.20	17.52	-25.11	17.51	22.86	21.49	-25.11	17.51
number of litters k = 27																		
0.1	0.0	0.003	-25.37	2.86	1.03	0.44	1.29	0.44	0.79	0.44	35.13	2.38	38.83	2.76	46.30	3.91	38.20	2.07
	0.1	0.012	1.22	0.79	1.94	0.73	2.20	0.73	1.61	0.73	7.68	5.02	0.92	5.03	12.35	6.15	-8.36	2.73
	0.2	0.021	1.03	1.05	0.92	1.04	1.25	1.05	0.56	1.05	-2.35	10.68	-15.59	11.52	0.17	12.81	-52.90	9.27
0.2	0.05	0.014	-9.78	2.37	2.49	1.03	2.62	1.04	2.31	1.04	10.35	1.93	13.02	2.11	21.01	2.70	35.07	3.39
	0.1	0.021	-1.82	1.65	1.67	1.34	1.82	1.34	1.44	1.34	-1.19	3.34	-1.21	3.56	8.93	4.10	25.03	4.32
	0.3	0.052	-0.99	2.36	-0.45	2.42	-0.17	2.41	-0.65	2.42	-13.58	9.92	-18.74	10.76	-0.07	11.36	7.70	12.07
0.5	0.0	0.009	-0.02	2.32	-0.17	1.18	0.17	1.18	0.18	1.19	29.24	1.42	40.47	2.32	46.67	2.99	54.69	4.12
	0.1	0.033	-0.79	2.12	-0.75	2.02	-0.75	2.02	-0.75	2.03	-3.93	2.77	-1.96	2.83	7.80	3.21	31.81	5.82
	0.2	0.057	0.74	2.93	0.73	2.92	0.72	2.92	0.73	2.93	-11.11	5.13	-9.46	5.23	4.71	5.63	55.46	12.34
number of litters k = 50																		
0.1	0.0	0.002	-28.56	2.95	0.14	0.12	0.15	0.12	0.12	0.12	9.63	0.18	12.60	0.26	13.78	0.31	18.30	0.46
	0.01	0.003	-9.45	1.12	0.70	0.14	0.71	0.14	0.66	0.14	6.09	0.19	7.61	0.24	9.06	0.28	14.34	0.39
	0.05	0.006	0.12	0.21	0.45	0.21	0.47	0.21	0.39	0.21	-1.90	0.62	-1.90	0.68	0.65	0.71	4.68	0.51
	0.2	0.019	-0.04	0.47	0.08	0.47	0.12	0.47	0.06	0.47	-5.72	3.59	-10.29	3.97	-4.50	4.06	-38.77	3.91
0.2	0.01	0.005	-44.09	8.86	0.73	0.22	0.75	0.22	0.69	0.22	0.38	0.08	2.95	0.11	4.15	0.13	7.29	0.20
	0.1	0.019	0.20	0.53	0.46	0.53	0.48	0.53	0.43	0.53	-4.42	1.04	-3.89	1.13	-0.18	1.16	10.47	1.27
	0.3	0.050	-0.62	1.10	0.01	1.13	0.05	1.12	0.02	1.13	-8.53	3.63	-9.41	4.02	-1.64	4.10	7.20	4.56
0.5	0.0	0.005	0.14	0.53	-0.04	0.33	-0.04	0.33	-0.04	0.33	10.94	0.28	14.74	0.34	16.05	0.39	16.62	0.41
	0.1	0.030	-0.50	0.82	-0.50	0.82	-0.50	0.82	-0.50	0.82	-4.14	0.90	-3.52	0.92	0.17	0.95	12.95	1.46
	0.3	0.079	0.38	1.69	0.23	1.73	0.23	1.73	0.24	1.73	-6.60	2.42	-5.18	2.54	2.62	2.63	67.83	9.04