Case Studies in Binary Dispersion

Kung-Yee Liang; Peter McCullagh

Biometrics, Volume 49, Issue 2 (Jun., 1993), 623-630.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199306%2949%3A2%3C623%3ACSIBD%3E2.0.CO%3B2-T

BIOMETRICS 49, 623–630
June 1993

THE CONSULTANT'S FORUM

Case Studies in Binary Dispersion

Kung-Yee Liang
Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, Maryland 21205, U.S.A.
and
Peter McCullagh
Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637, U.S.A.

SUMMARY

It is common in biomedical studies with binary responses that variability in the observed number of events exceeds binomial variability, a phenomenon known as overdispersion. Failure to make an adjustment to the nominal standard errors can lead to seriously misleading inference for regression analysis. In this note, we examine a series of examples drawn from the literature to see which of two commonly used variance formulas is more adequate for describing overdispersion in applications. Two methods, residual analysis and formal comparison, are introduced. We recommend that both methods be employed in seeking an appropriate variance expression for binary responses. Each of the five data sets exhibits substantial overdispersion, one favoring the beta–binomial form, another favoring a constant overdispersion factor. The remaining three examples exhibit no preference.

1. Introduction

The "ideal" or "standard" assumption in binary regression models is that the observations Y_i are independent, and that $Y_i \sim B(m_i, \pi_i)$, the binomial distribution with index m and parameter π. This is a one-parameter family of distributions, implying that the mean, variance, and all higher-order cumulants are determined by the single parameter π. In practical applications of binary regression models, it is frequently observed that variability in the observed number of successes exceeds binomial variability $m\pi(1 - \pi)$, sometimes by a substantial factor. This phenomenon is called overdispersion. It is assumed throughout this paper that interest remains in the systematic effect of explanatory variables on the response probability rather than on the nature of the excess dispersion. Since the precision of parameter estimates is degraded by overdispersion, failure to make an adjustment to the standard errors can lead to overoptimistic standard errors and seriously misleading inferences (Cox and Snell, 1989, §3.2; McCullagh and Nelder, 1989, §4.5).

The remedies that have been suggested in the literature fall into two broad categories. One option is to develop a parametric model for the excess dispersion, and to fit the more complex model by maximum likelihood. In practice, it is rarely feasible to do this in a realistic way on a routine basis. A practical difficulty is that several equally plausible models for the excess dispersion may give rise to quite different likelihood functions, each requiring nonstandard tailor-made software for fitting and testing. To take a tractable example, suppose that $Y \sim B(m, P)$, where P has the beta distribution with mean $\pi = \alpha/(\alpha + \beta)$, and variance $\tau^2\pi(1 - \pi) = \pi(1 - \pi)/(\alpha + \beta + 1)$. Then the marginal distribution of Y is

$$\Pr(Y = y;\ \alpha, \beta) = \binom{m}{y} \frac{B(y + \alpha,\ m - y + \beta)}{B(\alpha, \beta)}, \quad y = 0, 1, \ldots, m,$$

which is known as the beta–binomial distribution. For purposes of model construction it is usually appropriate to reparameterize in terms of the proportion π and a measure of excess dispersion such

Key words: Beta–binomial; Binary regression; Overdispersion; Toxicological data.

as $\tau^2 = 1/(\alpha + \beta + 1)$ or some function thereof (Griffiths, 1973; Williams, 1975; Crowder, 1978). The mean and variance of Y_i are given by $E(Y_i) = m_i\pi_i$ and

$$\text{var}(Y_i) = m_i\pi_i(1 - \pi_i)\{1 + (m_i - 1)\tau_i^2\}.$$

Model specification can now be completed in a variety of ways—for example, by taking the excess dispersion to be constant, $\tau_i^2 = \tau^2$, and assuming a linear logistic model for the effect of covariates on π_i. Note, however, that if we were to parameterize the beta–binomial by π together with, say, $\gamma = \alpha^{-1} + \beta^{-1}$, the latter being assumed constant over observations, a quite different likelihood function would be obtained.

The second option is to avoid a fully parametric specification and to assume a particular relationship between the variance and the mean of Y_i. Second-moment methods, in particular quasi-likelihood, are then used for fitting and testing. For example, guided in part by the beta–binomial discussion, we might be prepared to assume that the variance of Y_i is

$$\text{var}(Y_i) = m_i\pi_i(1 - \pi_i)\{1 + (m_i - 1)\tau^2\}, \tag{1}$$

with $0 \leq \tau^2 \leq 1$. The dispersion factor $1 + (m_i - 1)\tau^2$ can be derived under the assumption that $\text{var}(P_i) = \pi_i(1 - \pi_i)\tau^2$; in particular, it is not necessary that P should have the beta distribution. Alternatively, and slightly more conveniently, we might prefer to assume a constant dispersion factor relative to the binomial, namely

$$\text{var}(Y_i) = m_i\pi_i(1 - \pi_i)\sigma^2. \tag{2}$$

Overdispersion is then modelled by $\sigma^2 > 1$. Underdispersion, though less common in applications, can be modelled by $\sigma^2 < 1$. Certain cluster-sampling procedures can give rise to variances approximately of the form (2) (McCullagh and Nelder, 1989, §4.5.1). While the use of the quasi-likelihood method will in general result in loss of efficiency, its loss is known to be modest in many practical situations (Firth, 1987).

From a purely mathematical perspective, since Y is restricted to the integer values $0, 1, \ldots, m$ with mean $m\pi$, we must have $\text{var}(Y) \leq m^2\pi(1 - \pi)$, implying $\sigma^2 \leq m_i$ in (2) (McCullagh and Nelder, 1989, p. 125). Thus, if all integer values $m_i \geq 1$ are permissible, we can have only $\sigma^2 = 1$ in (2). This is an unfortunate property of the dispersion model (2), but it is rarely a serious concern in applications. In practice, dispersion factors are often in the range 1–5, whereas the indices m_i are typically larger than this. The restriction that $\sigma^2 \leq m_i$ can usually be ignored with impunity.

It is ultimately an empirical question whether excess variability in applications of binary regression models is best modelled by the variable dispersion factor in (1) or the constant dispersion factor in (2). The purpose of this paper is to look at a number of examples gleaned from the literature to see which, if either, dispersion model is appropriate. It is worth pointing out that modelling via serial correlation is a viable alternative when the binary responses are observed over time.

2. Examples

In the examples that follow, all fitted values are calculated under the standard model without overdispersion.

2.1 *Weil–Williams Toxicology Data*

The observations are the numbers of rats surviving the 21-day lactation period as a fraction of the number alive at 4 days (Weil, 1970; Williams, 1975). The control group and treatment group each comprise 16 litters of rats. Since the excess dispersion appears to be larger in the treatment group than in the control group, we examine the two groups separately. Figure 1 shows the residuals $(y_i - m_i\hat{\pi})/\sqrt{m_i\hat{\pi}(1 - \hat{\pi})}$ plotted against litter size m_i. There is some evidence that the residuals tend to increase with litter size, indicating a higher survival rate for larger litters, an effect not mentioned by Williams. Both the treatment effect and the litter size effect are of borderline statistical significance. However, there is no strong evidence that the variance of the residuals is related to litter size as predicted by (1). Inclusion of litter size as a covariate in the model has little effect on this conclusion.

It has been suggested on some biological grounds that the survival rate may be associated with litter size in a nontrivial manner (Dr Keith Soper, personal communication). How this observation may affect the variance expression of Y_i is unclear. For example, model (1) implies that, in the context of toxicological experiments, τ^2 has the within-litter correlation coefficient interpretation that is constant regardless of litter size. Model (2) assumes instead that the inverse of the within-litter coefficient is linearly related to the litter size. More studies are warranted to examine which of the two assumptions is more consistent with the above observations.

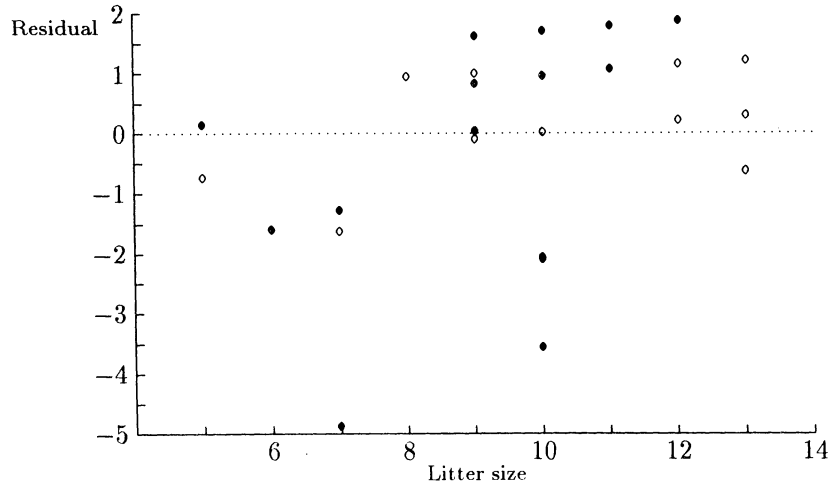

Figure 1. Residuals plotted against litter size for the Weil–Williams toxicology data. ◇ denotes a control litter; ◆ denotes treated litter.

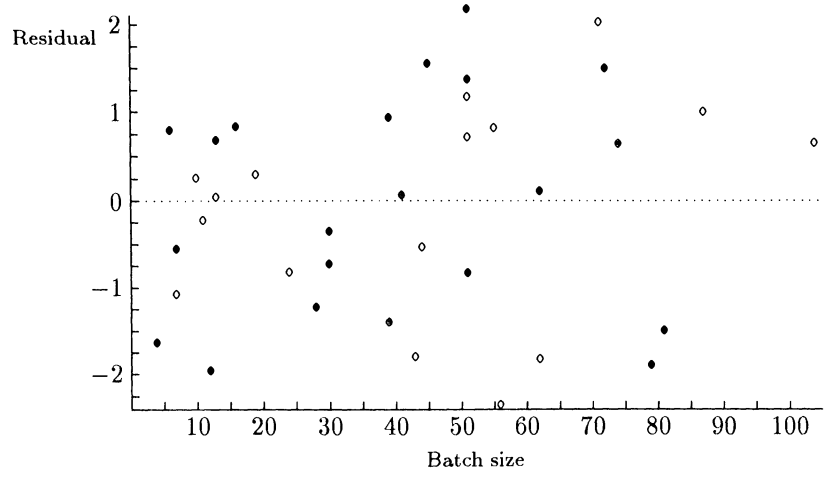

Figure 2. Residuals plotted against batch size for Crowder's seed germination data. ◇ denotes *O. cernua*; ◆ denotes *O. aegyptiaca.*

2.2 *Crowder's Germination Data*

These data, taken from Crowder (1978, Tables 1 and 3), were collected in a series of experiments concerning the effect of certain extracts on the germination rates of the seeds *O. cernua*, *O. aegyptiaca* 75, and *O. aegyptiaca* 73. Following Crowder's analysis, we treat the dilution series data in his Table 1 as a single-factor layout. The data from Table 3 are fitted using a 2 × 2 factorial model with interaction on the logistic scale to model the systematic effects. Effectively, therefore, we have seven groups with five to six batches in each group. The combined residuals from these fits are plotted in Figure 2.

Crowder's data exhibit strong evidence of overdispersion. Pearson's statistic is 54.01 on 30 degrees of freedom, giving an estimate of $\tilde{\sigma}^2 = 1.80$ based on the constant dispersion model (2). Although the range of batch sizes is large, the residual plot shows no evidence that the variance of the residuals is related in any way to the batch size as predicted by the beta-binomial model. The data, however, are consistent with both (1) and (2), so there is insufficient information to refute either form of overdispersion.

2.3 *Chen–Kodell–Howe–Gaylor's Toxicology Data*

The data, taken from Chen et al. (1991), were collected from a study on developmental effects resulting from exposure to hydroxyurea. The observations are the ratios of the number of dead or

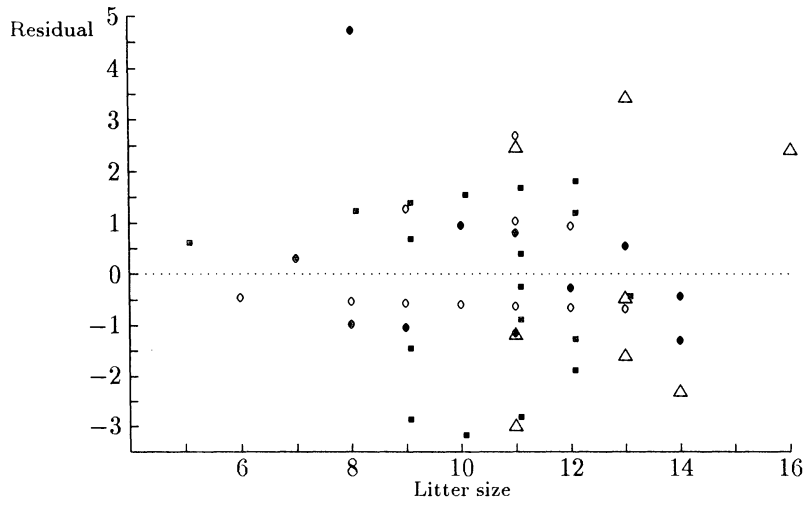

Figure 3. Residuals plotted against litter size for the Chen–Kodell–Howe–Gaylor teratology data. ◇ denotes a control litter; ◆ denotes low dose; △ denotes medium dose, ■ denotes high dose.

resorbed implants to the number of implantation sites. The experiment consisted of four treatment groups: control, low dose, medium dose, and high dose. There is strong evidence of overdispersion as Pearson's statistic is 148.7 on 56 degrees of freedom. However, there is very little evidence from Figure 3 to suggest that the variance of the residuals is linearly related to m_i. Just as in Section 2.2, the data appear to be consistent with both (1) and (2) and leave very little room to discriminate between the two forms of overdispersion.

2.4 *Soper's Teratology Data*

The observations shown in Table 1 are the numbers of resorbed fetuses from a teratological experiment for which the data are kindly provided by Dr Keith Soper of Merck, Sharp and Dohme. Dams were

Table 1
Data from a teratological experiment provided by Dr Keith Soper.
Y: number of resorbed fetuses;
M: number of fetuses per dam; D = dose (1: control, 2: 2 mg/kg, 3: 4 mg/kg, 4: 8 mg/kg).

$D = 1$		$D = 2$		$D = 3$		$D = 4$	
Y	M	Y	M	Y	M	Y	M
1	17	1	11	1	1	6	17
1	17	1	17	1	3	4	21
1	17	1	17	1	14	2	16
1	18	1	18	1	15	2	17
0	13	1	20	1	15	1	14
0	14	0	15	1	16	1	19
0	14	0	15	0	12	0	14
0	15	0	16	0	14	0	14
0	16	0	16	0	16	0	15
0	16	0	16	0	16	0	15
0	16	0	16	0	16	0	15
0	16	0	17	0	17	0	16
0	17	0	17	0	17	0	16
0	17	0	17	0	17	0	16
0	17	0	17	0	17	0	17
0	17	0	17	0	17	0	17
0	18	0	18	0	18	0	17
0	18	0	19	0	19	0	17
0	18	0	21	0	19	0	17
0	18			0	19	0	18
0	19			0	20	0	20

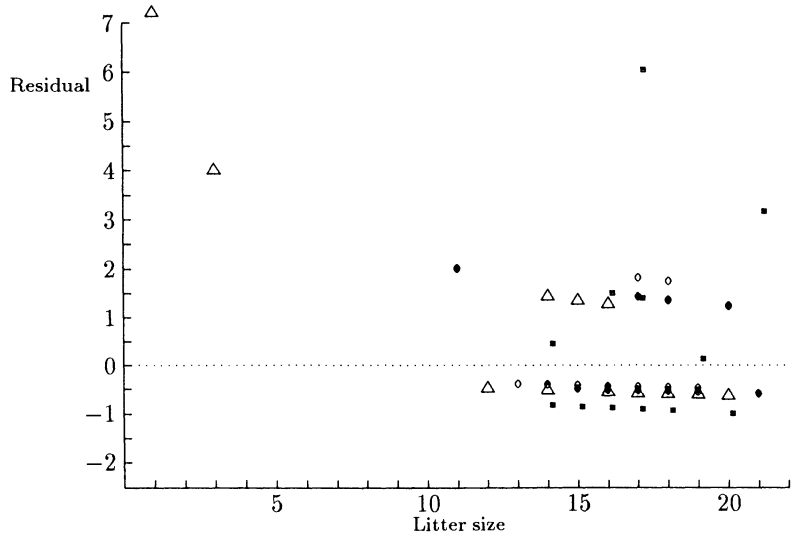

Figure 4. Residuals plotted against litter size for Soper's teratology data. ◇ denotes a control litter; ◆ denotes a dose of 2 mg/kg, △ denotes a dose of 4 mg/kg, ■ denotes a dose of 8 mg/kg.

assigned at random to vehicle control or dose 2, 4, or 8 mg/kg of a test compound, 20 dams per treatment group. Dosing was done once daily during days 6–17 of gestation, and caesarean section was on day 20. The treatment effect is evident only among the highest dose group. This finding was unaltered when the linear and quadratic terms of litter size were adjusted. Figure 4 shows little, if any, relationship between residuals and litter size, though the small number of large positive residuals deserves further scrutiny.

2.5 *Shepard–Mackler–Finch's Teratology Data*

The final example uses data from a study (Shepard, Mackler, and Finch, 1980) of the dietary regimen effects on fetal development in laboratory rats reported by Moore and Tsiatis (1991). Fifty-eight female rats were put on iron-deficient diets and divided into four groups. Group 1 is the untreated (low-iron) group; group 2 received injections on day 7 or day 10 only; group 3 received injections on days 0 and 7; and group 4 received injections weekly. There is a strong treatment effect that is unaltered even when litter size is included as a covariate. There is strong evidence judging from Figure 5 that the variance of the residuals is related to litter size as described by (1).

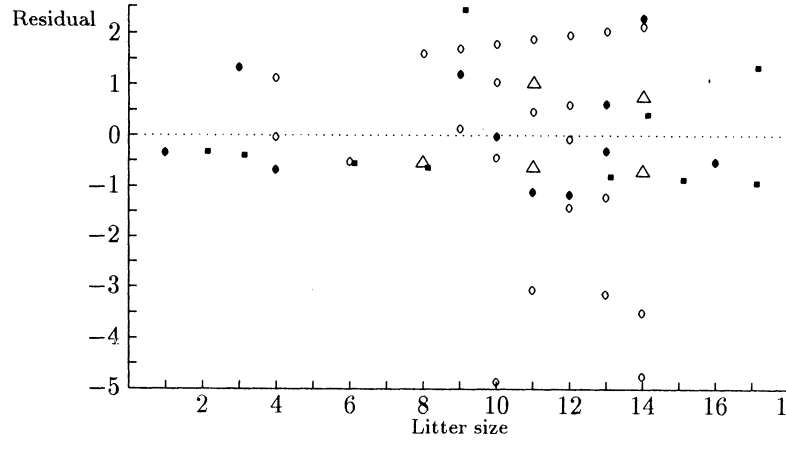

Figure 5. Residuals plotted against litter size for the Shepard–Mackler–Finch data. ◇ denotes a control litter; ◆ denotes injection on day 7 or 10, △ denotes injection on day 0 and day 7, ■ denotes weekly injection.

3. Formal Comparison of Dispersion Models

One way formally to compare the dispersion models (1) and (2) is to take the squared residuals from a well-fitting model as response. Then

$$R_i^2 = \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1 - \hat{\pi}_i)}$$

is approximately distributed as $\sigma_i^2\chi_1^2$ with σ_i^2 satisfying the linear model

$$\sigma_i^2 = \alpha + \beta(m_i - 1). \tag{3}$$

Under (1) we have $\alpha = 1$, whereas under (2) we have $\beta = 0$. If the squared residuals are regarded as independent, we may use gamma regression to fit the two-parameter model (3) together with both one-parameter submodels.

For Weil–Williams' data, the scaled deviances are 26.89 for (1), 26.0 for (2), and 25.17 for (3). Again, neither (1) nor (2) is rejected in favor of (3). There is a statistically insignificant preference of the constant dispersion form (2) over (1).

For Crowder's seed data, the scaled deviances are 32.50 for (1), 33.96 for (2), and 32.32 for (3). Thus neither (1) nor (2) is rejected in favor of the more general form (3). There is a slight, but statistically insignificant, preference for the beta–binomial form (1) over (2).

For the data of Chen et al., the scaled deviances are 26.67 for (1), and 26.47 for (2) and (3). Either (1) or (2) explains the residuals as well as the more general form (3). Judging from the deviances, no preference may be claimed between (1) and (2).

For Soper's data, the scaled deviances are 67.84 for (1) and 54.83 for (2) with 81 degrees of freedom. It is clear that model (2) is favored over the beta–binomial form (1). This preference persists when the dispersion factors are modeled separately for each of four groups. The scaled deviances are 59.27 for (1) and 44.76 for (2) with 78 degrees of freedom. In fact, this additional model fitting suggests that the degree of dispersion varies among the four groups [the likelihood ratio test statistic equals 10.09 on 3 degrees of freedom for model (2)]. The degree of dispersion appears to increase with dose level ($\hat{\sigma}_1^2 = .78$, $\hat{\sigma}_2^2 = .80$, $\hat{\sigma}_3^2 = 3.82$, $\hat{\sigma}_4^2 = 2.98$).

For the data of Shepard et al., we observe as the deviances 69.4, 74.0, and 66.1 for models (1), (2), and (3), respectively. The beta–binomial form (1) is clearly preferable as model (2) is rejected in favor of (3) (the likelihood ratio test equals 3.95 with 1 degree of freedom) whereas no such claim can be made for model (1). This appears to be consistent with the finding from Section 2.5.

The calculations can be done easily, for example, in GLIM (Baker and Nelder, 1978). The relevant commands are given in the Appendix.

4. Discussion

In this paper, we have examined five data sets drawn from the literature to see which variance formula, if either, is more adequate in describing dispersion associated with binary data. In three of the examples, we have failed to refute either form of overdispersion. For the rest, the beta–binomial variance (1) clearly outperformed (2) in one instance, and the reverse holds in the other example. Thus, the answer to the above question is equivocal at best. For these two examples, Table 2 gives the regression estimates and their estimated standard errors when fitting variance formulas (1) and (2). For Soper's data, the estimated effect, on the logit scale, of receiving 4 mg/kg relative to control when using (1) is almost twice as big as when using the more favored variance (2) (.977 versus .503). In addition, the estimated standard errors when using (1) are inflated by as much as 15%. The discrepancies are milder in the Shepard–Mackler–Finch data.

The methods presented here are most useful when the number of responses is small or modest. This is the situation where choice of the variance formulas may play an important role for regression inference. These two methods, residual analysis and formal comparison introduced in Section 3, are easy to employ. We also note that for all the examples considered, conclusions regarding choice of the variance formula are consistent between the two approaches. It is our recommendation that both methods be adopted to strengthen the conclusion when examining the adequacy of the variance expression for overdispersion. It is, however, worthy of note that residuals are inevitably skewed when the litter sizes are small and caution should be exercised in this case to avoid overinterpretation of the results.

Finally, when the number of responses is sufficiently large, the "robust" variance estimate of regression coefficients by Liang and Zeger (1986) may be employed. It is robust in the sense that it remains consistent even when the wrong variance structure is chosen. However, the question

Table 2

Estimates and standard errors from logistic regression analyses using quasi-likelihood for two data sets. Note that the regression coefficients for the last three variables in each data set represent the difference, in logit scale, between each of three treated groups with the control reference group.

Variable	Var(Y)	
	$m\pi(1-\pi)(1+(m-1)\tau^2)$	$m\pi(1-\pi)\sigma^2$
Soper's data		
Intercept	$-4.480 \pm .857$	$-4.454 \pm .747$
2 mg/kg	$.370 \pm 1.141$	$.311 \pm 1.005$
4 mg/kg	$.977 \pm 1.022$	$.503 \pm .968$
8 mg/kg	$1.404 \pm .964$	$1.422 \pm .839$
	$\hat{\tau}^2 = .119$	$\hat{\sigma}^2 = 2.333$
Shepard–Mackler–Finch's data		
Intercept	$1.212 \pm .223$	$1.144 \pm .218$
Group 2[a]	$-3.370 \pm .562$	$-3.323 \pm .560$
Group 3	-4.585 ± 1.302	-4.476 ± 1.237
Group 4	$-4.250 \pm .848$	$-4.130 \pm .806$
	$\hat{\tau}^2 = .192$	$\hat{\sigma}^2 = 2.863$

[a] Group 2: injections on day 7 or day 10 only; group 3: injections on day 0 and day 7; group 4: injections weekly.

concerning the number of responses that is "sufficiently large" remains open and will be addressed through simulations in a separate report.

Acknowledgements

This work was supported in part by Grants GM-39261 from the National Institutes of Health and DSM-91-01333 from the National Science Foundation.

Résumé

Il est courant, dans les études biomédicales réalisées avec des données dichotomiques, que la variabilité du nombre d'événements observé soit plus grande que la variabilité d'une loi binomiale: c'est ce qu'on appelle une "overdispersion" (ou, en français, une surdispersion). Lorsque cela se produit, le fait de ne pas ajuster la précision des paramètres en fonction de ce phénomène peut conduire, notamment dans les analyses de régression, à des inférences peu faibles. Le présent article examine une série d'exemples tirés de la littérature, afin de déterminer laquelle des deux formules communément employées pour modéliser la variance dans ce cas de surdispersion s'avérerait la plus adéquate. L'outil utilisé pour la comparaison consiste en deux méthodes—l'"analyse résiduelle" et la "comparaison formelle"—particulièrement indiquées pour évaluer la qualité de la modélisation de la variance dans le cas des réponses dichotomiques. Des cinq exemples abordés ici—qui présentent tous des surdispersions marquées—l'un amène à préférer une modélisation bêta-binomiale pour la surdispersion, alors qu'un autre encourage plutôt la simple prise en compte d'un facteur constant (pour cette même surdispersion). Dans les trois autres exemples, il est indifférent de choisir l'un ou l'autre modèle.

References

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System, Release 3, Generalized Linear Interactive Modelling*. Oxford: Numerical Algorithms Group.

Chen, J. J., Kodell, R. L., Howe, R. B., and Gaylor, D. W. (1991). Analysis of trinomial responses from reproductive and developmental experiments. *Biometrics* **47,** 1049–1058.

Cox, D. R. and Snell, E. J. (1989). *Binary Data*. London: Chapman and Hall.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics* **27,** 34–37.

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

Griffiths, D. A. (1973). Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases in a disease. *Biometrics* **29,** 637–648.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

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

Moore, D. F. and Tsiatis, A. (1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics* **37,** 383–401.

Shepard, T. H., Mackler, B., and Finch, C. A. (1980). Reproductive studies in the iron-deficient rat. *Teratology* **22,** 329–334.

Weil, C. S. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology* **8,** 177–182.

Williams, D. A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31,** 949–952.

Received October 1991; *revised March and June* 1992; *accepted July* 1992.

Appendix

The following are the commands in GLIM needed to perform the formal comparison introduced in Section 3. The variable R denotes the squared residual.

```
$YVAR R
$ERR G
$SCALE 2
$LINK L (or SQRT)
$FIT —
$RECYCLE
$LINK I
$FIT.
```

Note that the intermediate steps (commands 4–6) are needed only if one encounters negative fitted values. Our experience in analyzing these five data sets has been that this is not an uncommon phenomenon and the recommended steps are very helpful in avoiding this difficulty.