Rivista di Statistica Applicata Vol. 21 n. 1 Marzo 1988

Regression analysis with the Beta–Binomial distribution

Antonio Forcina, Dipartimento di Scienze Statistiche – Università di Perugia
Luisa Franconi, Via Bonciario 6 – Perugia

The Beta–Binomial model is examined in the light of the literature concerning the analysis of binary data with overdispersion. It is shown that the likelihood function can be maximazed by using iteratively a Newton–Raphson algorithm; its implementation in *GLIM* is sketched. In addition this method gives a profile likelihood for the overdispersion parameter.

An application to a data set concerning a study on a population of Saissetia Oleae is presented to illustrate the method.

1. Introduction

The analysis of binary data is usually based on the assumption that, conditionally on a set of explanatory variables, the observations follow the binomial distribution (see Nelder and McCullagh, 1983, p. 72). The reason why this model may be inadequate in practice is that the assumption of independence, implicit in the binomial scheme, is often unrealistic.

Regression models for dependent binary data have been considered by Anderson and Aitkin (1985) and Gilmour *et al.* (1985). Both papers are based on the assumption that the binary variable is equal to 1 (or 0) according to whether a continuous (but unobservable) variable does (or does not) exceeds a given threshold; they also assume a regression model with fixed and random effects for the underlying variable, implying that the binary observations are correlated and that their variance has several components, corresponding to each random effect. Unfortunately the likelihood based on such models is very complicated. Anderson and Aitkin (1985) assume a logistic distribution for the underlying variable and, having simplified the likelihood by numerical integration, apply the E–M algorithm to obtain parameter estimates. Gilmour *et al.* (1985) assume instead that the underlying variable has a normal distribution and show that the quasi–likelihood method can be used after suitable approximations.

Applications of the Beta–Binomial model have also been limited by numerical difficulties. Crowder (1978) described an algorithm for one–way analysis of variance with binary data. The extension to the case of a general regression model was considered by Williams (1982) who proposed to combine quasi–likelihood estimation of the regression parameters with a moment estimator of

the overdispersion parameter.

A different model has been proposed by Ochi and Prentice (1984). They replace the assumption that $y \sim$ Binomial (n, p) with the assumption that y components of a n–dimensional equicorrelated normal distribution exceed a given threshold. Numerical approximations are necessary to solve the likelihood equations; the only advantage of this model over the Beta–Binomial seems to be that underdispersion, as well as overdispersion, can be represented.

In section 2 we show that the maximum likelihood equations for the regression parameters of a general Beta–Binomial model can be solved by a Newton–Raphson algorithm, easily implemented in *GLIM*. A data set concerning the distribution of a population of Saissetia Oleae is used in section 3 to exemplify the method.

2. Maximum likelihood estimation

The Beta–Binomial model can be defined as follows. Let y_i, $i = 1, \ldots, m$ be a set of random variables which are independent and such that $y_i | a_i \sim Bin(n_i, a_i)$; let the variables a_i be independent and follow the Beta distribution with $E(a_i) = p_i$ and $Var(a_i) = p_i(1 - p_i)/(r + 1)$. It is well known that the marginal distribution of $\boldsymbol{y} = (y_i, \ldots, y_m)'$ is such that its components are independent with density function

$$P(y_i) = \binom{n_i}{y_i} \frac{\Gamma(r)}{\Gamma(rp_i)\Gamma(r - rp_i)} \frac{\Gamma(rp_i + y_i)\Gamma(r - rp_i + n_i - y_i)}{\Gamma(r + n_i)} \qquad (1)$$

where $E(y_i) = n_i p_i$ and $Var(y_i) = n_i p_i(1 - p_i)(1 + (n_i - 1)/(r + 1))$. It should be noted that $t = 1/(r + 1)$ may be interpreted as a parameter of overdispersion because when $t = 0, Var(y_i)$ has the same form as with the binomial distribution and it increases with t. Using the well known properties of the gamma function, the log–likelihood can be written as

$$-L(\boldsymbol{p}, r; \boldsymbol{y}) = C + \sum_i \Big(\sum_h \ln(rp_i + y_i - h) + \sum_k \ln(r - rp_i + n_i - y_i - k) - \sum_j \ln(r + n_i - j) \Big) \qquad (2)$$

where $\boldsymbol{p} = (p_1, \ldots, p_m)'$, $h = 1, \ldots, y_i$, $k = 1, \ldots, n_i - y_i$, $j = 1, \ldots, n_i$.

Let us assume that the p_i are connected to a vector of regression parameters by a logit model (only minor changes are needed to use a different link function), i. e. $p_i = \exp(l_i)/(1 + \exp(l_i))$ and $\boldsymbol{l} = (l_1, \ldots, l_m)' = \boldsymbol{X}\boldsymbol{b}$, where $\boldsymbol{b}$ is a p by 1 vector of regression parameters and $\boldsymbol{X}$ is a design matrix assumed of full rank.

Assume for the moment that r is known, then differentiating (2) with respect to $\boldsymbol{b}$ and equating to 0, we have

$$\partial L/\partial \boldsymbol{b} = \boldsymbol{X}'\boldsymbol{S}\boldsymbol{u} = \boldsymbol{0} \tag{3}$$

where $\boldsymbol{S} = \partial \boldsymbol{p}'/\partial \boldsymbol{l} = \text{diag}(p_1(1-p_1),\ldots,p_m(1-p_m)), \boldsymbol{u} = (u_1,\ldots,u_m)'$ and $u_i = r(\sum_h 1/(rp_i + y_i - h) - \sum_k 1/(r - rp_i + n_i - y_i - k))$. The Newton–Raphson method can be used to solve equation (3) with respect to $\boldsymbol{b}$. The method is based on a first order Taylor series expansion of (3) with the matrix of 2nd derivatives replaced by its expectation (see Wedderburn, 1976). Here we obtain

$$E((\partial/\partial \boldsymbol{b})(\partial L/\partial \boldsymbol{b})') = -\boldsymbol{X}'\boldsymbol{S}E(\boldsymbol{V})\boldsymbol{S}\boldsymbol{X} \tag{4}$$

where $v_i = r^2(\sum_h (rp_i + y_i - h)^{-2} + \sum_k (r - rp_i + n_i - y_i - k)^{-2})$ and where $\boldsymbol{V} = \text{diag}\ (v_1,\ldots,v_m)$. After simple calculations we obtain an equation which, given a starting value $\boldsymbol{b}$ provides an updated estimate $\hat{\boldsymbol{b}}$.

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}'\boldsymbol{S}E(\boldsymbol{V})\boldsymbol{S}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{S}E(\boldsymbol{V})\boldsymbol{S}\boldsymbol{z} \tag{5}$$

where $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{b} + (\boldsymbol{S}E(\boldsymbol{V}))^{-1}\boldsymbol{u}$ and $\boldsymbol{S}, \boldsymbol{V}, \boldsymbol{u}$ are evaluated at $\boldsymbol{b}$. Maximum likelihood estimates of $\boldsymbol{b}$ can be obtained by iterative use of (5) which has the familiar form of a weighted least square equation. This can be easily implemented in *GLIM* by setting $\%yv = 0$, $\%fv = -\boldsymbol{u}$, $\%dr = (\boldsymbol{S}E(\boldsymbol{V}))^{-1}$ and $\%va = E(\boldsymbol{V})$. Because there is no simple way to compute the expectation it is necessary to replace $E(\boldsymbol{V})$ with $\boldsymbol{V}$. This adjustement will usually increase the rate of convergence, though the algorithm may be less stable if the starting point is far away from the maximum (see Jorgensen, 1984 p. 286, for further comments). We tested the algorithm on 2 data sets provided by Crowder (1978) and on the example described in section 3 and convergence was always obtained in few steps (2 to 4).

The estimates of $\boldsymbol{b}$ and $\boldsymbol{p}$ are functions of r, hence, if we replace $\boldsymbol{p}$ with $\hat{\boldsymbol{p}}$ in (2), we obtain the profile log–likelihood with respect to r, say $L(r)$. This function is easy to compute and its plot is very useful to detect unexpected features of the likelihood function. It can be used also to construct likelihood based confidence intervals for $\hat{r}$ (or $\hat{t}$). Simple numerical methods are available to locate the maximum of $L(r)$. Moreover, the variance of $\hat{r}$ and the covariance matrix of $\hat{\boldsymbol{b}}$ can be computed by the normal approximation (see e.g. Forcina, 1986).

$$Var(\hat{r}) = -\partial \hat{r}/\partial L(\hat{r})$$

$$\text{Var}(\hat{\boldsymbol{b}}) = \text{Var}(\hat{\boldsymbol{b}}|\hat{r}) + \text{Var}(\hat{r})\partial \hat{\boldsymbol{b}}(\hat{r})/\partial \hat{r}(\partial \hat{\boldsymbol{b}}(\hat{r})/\partial \hat{r})'$$

where the derivatives can be replaced by numerical approximations.

3. An application

We used a data set described by Bagnoli et alt. (1984) concerning the distribution of a population of Saissetia Oleae. A random sample of 140 branchlets were collected from 5 different olive trees. For each branchlet the number of infested leaves and the total number of leaves present were recorded together with the length of the branchlet, and the number of adults of Saissetia Oleae present outside the leaves. The number of infested leaves per branchlet may be considered a binary variable and there are several questions of interest concerning the probability that a leaf is infested, for example:

(i) is this probability constant across trees?

(ii) does it increases with the length of the branchlet and the number of adults present outside the leaves?

(iii) is the binomial model adequate or, because of dependence between the state of each leaf within the same branchlet, an overdispersion model is necessary?

A binomial logistic regression model was first used. It showed that trees were not homogeneous and that the probability of infestation increased significantly with the number of adults outside the leaves and the square root of the length of the branchlet. However the residual deviance was quite larger than expected, indicating a lack of fit probably due to overdispersion as stated in (iii) above. The Beta–Binomial logistic model showed that this conjecture was probably correct as the t parameter is significantly larger than 0. The main results are displayed in tab. I below.

Tab. I – Parameter estimates and standard errors

Param.	*Estimate*	*S.E.*	*Param.*	*Estimate*	*S.E.*
Intercept	− 1.966	0.219	*pol*	0.132	0.029
D	0.484	0.252	$t = 1/(r+1)$	0.086	0.026

The parameter D corresponds to the difference between trees 2, 4 and 5 with respect to trees 1 and 3 which are less infested; *pol* is the regression coefficient for the number of insects present outside the leaves, it is positive and highly significant, because the leaves are more likely to be infested if a larger population is present.

The fact that with this model the probability of infestation does not depend on the length of the branchlet may be explained by noting that observations with a large n (number of leaves) are weighted less than with the binomial model because they have a relatively larger variance.

The profile log–likelihood of t, the overdispersion parameter, is plotted in Fig. 1: it is only slightly asymmetric on the left. Because twice the difference

in the log–likelihood has an asymptotic chi–square distribution with 1 d.f., t is greater than 0 at a significance level smaller than 0.001. The likelihood based 0.95 confidence interval (0.036, 0.152) is slightly different from the one based on the normal approximation (0.035, 0.137).

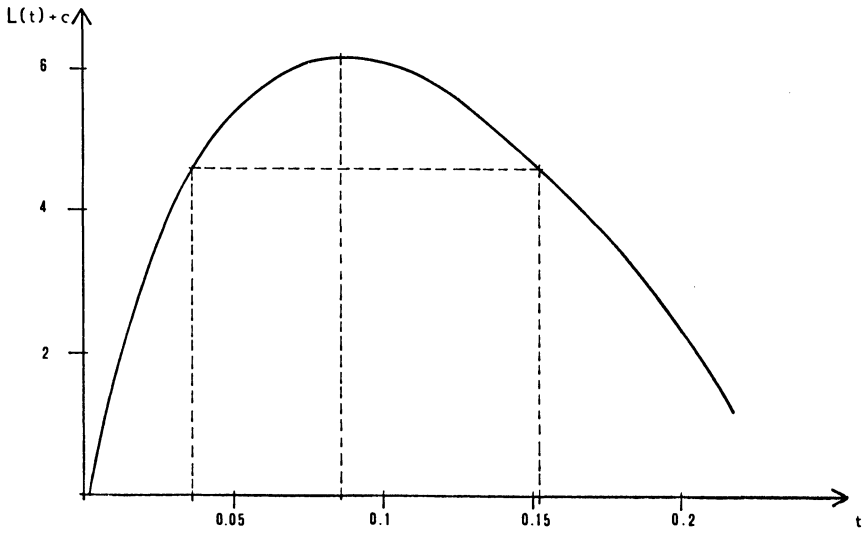

Fig. 1 – Profile log–likelihood for the parameter of overdispersion

References

Anderson D.A., Aitkin M., 1985, Variance component models with binary response: interviewer variability, *Journ. Royal Statist. Soc. B*, **47**, 203–210.

Bagnoli B., Forcina A., Pucci C., 1984, Studio della distribuzione degli adulti di Saissetia Oleae (oliv.), *Redia* **67**, 527–537.

Crowder M.J., 1978, Beta–binomial ANOVA for proportions. *Appl. Statist.*, **27**, 34–37.

Forcina A., 1986, Correlated observations with normal error, *GLIM Newsletter*, **12**, 31–32.

Gilmour A.R., Anderson R.D., Rae A.L., 1985, The analysis of binomial data by a generalized linear mixed model, *Biometrika*, **72**, 3, 593–599.

Jorgensen B., 1984, The delta algorithm and *GLIM. Int. Stat. Review* **3**, 283–300.

McCullagh P., Nelder J.A., 1983, *Generalized linear models*, Chapman and Hall, London.

Ochi Y., Prentice R.L., 1984, Likelihood inference in a correlated probit regression model, *Biometrika*, **71**, 3, 531–543.

Wedderburn R.W.M., 1976, On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models, *Biometrika*, **63**, 27–32.

Williams D.A., 1982, Extra binomial variation in logistic linear models, *Appl. Statist*, **31**, 144–148.

A set of GLIM 3.77 Macros that implement the above algorithm is available from Antonio Forcina.

Riassunto

Dopo una breve rassegna della letteratura concernente i modelli per l'analisi di dati binari in presenza di iperdispersione, si riprende in esame il modello Beta–Binomiale. Si dimostra che la funzione di verosimiglianza può essere massimizzata usando interattivamente un algoritmo di tipo Newton–Raphson, che è di facile implementazione in *GLIM*. Il metodo proposto fornisce inoltre il profilo di verosimiglianza rispetto al parametro di iperdispersione.

Un'applicazione ad un insieme di dati relativi ad uno studio su una popolazione di Saissetia Oleae viene presentato per illustrare le caratteristiche del metodo.