Tropical Agricultural Research Vol. 8 1996, 30-37

A Method for Handling Overdispersion in Binary Data: Illustrated by Analysis of Survey Data on Pineapple Wilt Disease

S. Samita and R.O. Thattil

Department of Crop Science
Faculty of Agriculture
University of Peradeniya
Peradeniya

ABSTRACT. *Many problems are associated with the analysis of binary data. A standard analysis of variance cannot be used to model binary data for many reasons. These problems are discussed along with the problem of overdispersion. Although the linear logistic model is commonly used to analyse binomial data, it is not appropriate in the presence of overdispersion.*

It is shown here how overdispersion can be handled using the logistic normal binomial distribution and defining a parameter for aggregation. The method is illustrated using survey data collected on pineapple wilt disease from four sites in Gampaha district (Kattota, Kalagedihena, Navadiga and Urapola).

INTRODUCTION

In an experiment in which the variable of interest is an incidence or occurrence, the observation made on each of the experimental units will take one of two possible values. For example, a seed may or may not germinate under certain experimental conditions; a patient in a clinical trial to compare alternative forms of treatment may or may not experience relief; an insect in an insecticidal trial may survive or die when exposed to a particular dose; a plant may be observed as either diseased or healthy in an epidemiological experiment. Such data are called binary data. The terms incidence data or quantal data are used alternatively. The two possible forms for each of the observations are often described generically by the terms 'success' and 'failure'.

In most circumstances, interest centres not only on the response of one particular experimental unit (seed, plant or insect) but on a group of units that have all been treated alike. For instance, a batch of seed may be exposed to certain set of conditions and the proportion of seeds germinated in the batch is

used as the observation. The resulting data are then referred to as group binary data or binomial data.

Review of analysis of binomial data

For binomial data the response from the i^{th} unit, $I = 1,2$ .....n is the proportion y_i / n_i. The approximate distribution for the i^{th} observation y_i is usually the binomial distribution (Cox and Snell, 1989; McCullagh and Nelder, 1989) with parameters n_i and p_i. $B(n_i, p_i)$ where p_i is termed success probability and n_i is the sample size. For the binomial distribution, the means and variances are given by $\mu_{yi} = n_i p_i$ and $Var(y_i) = n_i p_i (1- p_i)$.

Following the regression model for continuous data, one can easily adopt the model given below for binomial data.

$$E(y_i) = p_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots\ldots\ldots \beta_k x_{ki} \qquad (1)$$

and apply the method of least squares to obtain estimates for, $\hat{\beta}_0, \hat{\beta}_1 \ldots\ldots\ldots \hat{\beta}_k$ (McCullagh and Nelder, 1989).

There are number of drawbacks for fitting a standard regression model to binomial data. They are:

1. Non-constant variance for proportions (Snedecor and Cochran, 1989). The method of weighted least squares (Aitken *et al.*, 1989) using iterative schemes has been proposed as a solution to this problem.

2. Normality assumption cannot be made.

3. Since the range of β_j is $(-\infty, \infty)$ the estimated values for p_i (see equation 1) may not lie in the interval (0, 1). Transformation techniques can be used to solve this problem and the most commonly used transformation is the logistic transformation.

The linear logistic model may be written as :

$$logit\,(p_i) = \log\left[\frac{p_i}{1-p_i}\right] = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_k x_{ki} \qquad (2)$$

and p_i can be obtained as $p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$.

In order to fit a linear logistic model to a given set of data, k+1 unknown parameters, β_0, β_1 ..... β_k have first to be estimated. These parameters are readily estimated using the method of maximum likelihood (McCullagh and Nelder, 1989).

The log likelihood function for the binomial distribution can be written as:

$$\log L(\beta) = \sum_{i=1}^{n} \{\log \binom{n_i}{y_i} + y_i \eta_i - n_i \log(1 + e^{\eta_i})\} \qquad (3)$$

where η_i is the linear predictor as in the equation (2)

The algorithm is implemented in widely available computer packages such as GLIM (GLIM, 1985) and SAS (SAS, 1990), for fitting models to binary response data.

The maximum achievable log likelihood is attained at the point y_i / n_i which does not occur in the model space under H_0. The residual deviance (D) is defined as twice the difference between the maximum achievable log likelihood and that attained under the fitted model. The deviance function is written as

$$D = 2 \sum_i y_i \log \frac{y_i}{\hat{y}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \hat{y}_i} \qquad (4)$$

where,

$$\hat{y}_i = n_i \hat{p}_i .$$

Since D is asymptotically distributed as X^2_{n-k-1} where k+1 (as in equation 1) is the number of fitted parameters under H_0, D is used as a goodness-of-fit statitic for testing the adequacy of the fitted model. In addition, the significance of a particular term is tested by comparing deviance change (between two nested models), which has an asymptotic X^2 distribution, with the corresponding

models), which has an asymptotic X^2 distribution, with the corresponding change of degrees of freedom (df).

Overdispersion

When a fitted linear logistic model to n binomial proporties is satisfactory, the residual deviance has an approximate X^2 distribution with n-k-1, it follows that the residual deviance for the best-fitting model should be approximately equal to its df. If the fitted model does not adequately describe the observed proportions the residual deviance is likely to be greater than n-k-1. If this happens even after fitting the saturated model the data is said to exhibit overdispersion, a phenomenon that is also known as extra binomial variation.

Solution to overcome overdispersion

A solution proposed in this study for the above problem is to fit a conditional probability model, the logistic-normal binomial model, rather than the linear logistic model. The logistic-normal binomial model may be formulated as:

$$logit\ (p_i) = \eta_i + \gamma z_i \qquad (5)$$

where p_i is the true response probability (Collett, 1991), *i.e.* the expected response probability and the variability in the response probability together, η_i the linear predictor, γ is the coefficient of the random term z_i. The z_i is the standardized variate with zero mean and unit variance and thus γ is the standard deviation of the random effect variable. Thus, $E(logit\ p_i) = \eta_i$ and $Var\ (logit\ p_i) = \gamma^2$.

The software EGRET (EGRET, 1991) can be used to obtain the 'best' fitting logistic normal binomial model.

Illustration

The above analysis is illustrated using data collected from four sites in Gampaha district (Kattota, Kalagedihena, Navadiga and Urapola), in a survey conducted on the incidence of Pineapple wilt disease.

The data comprised of a two factor factorial (2 x 2) structure, where one factor was cultivars (Murici and Kew) and the other factor age of the crop (two years and four years). For each combination, an array of 12 rows x 30 plants per row were investigated. Each of these arrays were divided into quadrats of size 10 so that for each factorial combination there were 36 quadrats. Orientation of the quadrats were taken along the plant rows because disease spread was more prominant along the rows rather than across the rows.

Table 1. Possible models for the data collected.

Fitted deviance model	Fixed effect component		Random effect component	df
A	μ	-	143	458.11
B	$\mu + \alpha_i$	-	142	452.05
C	$\mu + \beta_j$	-	142	423.54
D	$\mu + \alpha_i + \beta_j$	-	141	417.33
E	$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	-	140	368.54
F	μ	γ	142	332.98
G	$\mu + \alpha_i$	γ	141	330.71
H	$\mu + \beta_j$	γ	141	318.32
I	$\mu + \alpha_i + \beta_j$	γ	140	315.08
J	$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	γ	139	292.59
K	$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	γ_{ij}	136	276.98

Where μ = overall mean α_i = effect of i^{th} cultivar, i = 1 (Murici), 2 (Kew), β_j = effect of j^{th} Year of stand j = 1 (two years), 2 (four years), $(\alpha\beta)_{ij}$ = interaction effect of i^{th} level of cultivar and j^{th} level of age.

Table 1 shows possible models that can be fitted for this data, with corresponding df and deviance values.

The differences between deviances are calculated for specified pairs where the difference between two models in the pair is only one term. The

deviance difference and the corresponding tests with the associated probabilities are given in Table 2.

It is evident that the saturated model (fit K in Table 1) is the most appropriate model to describe the data. That is, all main effects (cultivar and age) and the interaction between cultivar and age is present with respect to desease incidence. In addition, varying random effects among different factorial combinations is also present.

Parameter estimates for the best fit are given in Table 3.

Table 2. Possible likelihood ratio tests for the models.

	Test	Compare fits	Likelihood ratio statistics	df	*P* value
1.	Test for cultivar effect adjusted for the age	C Vs D	6.207	1	0.01
2.	Test for age adjusted for the cultivar effect	B Vs D	34.72	1	<0.01
3.	Test for interaction adjusted for both factors	D Vs F	48.79	1	<0.01
4.	Test for the random effect	E Vs J	75.95	1	<0.01
5.	Test for the cultivar effect (adjusted) in the presence of random effect	H Vs I	3.24	1	<0.01
6.	Test for age effect (adjusted) in the presence of random effect	G Vs J	15.64	1	<0.01
7.	Test for interaction (adjusted) in the presence of random effect	I Vs J	22.49	1	<0.01
8.	Test of difference levels of random effect	J Vs K	24.61	1	<0.01

Models A, B, C, D, E, F, G, H, I and K as defined in Table 1

The parameterization adopted in GLIM and EGRET is called 'corner point' parameterization in which first level of each factor is set to zero. For

instance second row estimate of Table 3 is the difference between cultivar Kew and Murici for the disease incidence. Since the disease incidence is -1.953 and standard error is 0.481, t = -1.953/0.481 = 4.06 of which P<0.05. Therefore, disease incidence is higher with Murici compared to Kew. Other comparisons can be made similarly.

In comparison with linear logistic model, standard errors with logistic normal binomial model are much larger, in this case at least by 50%. This increase takes into account variability in the response probability due to non random spatial patterns of disease incidence. Consequently the quantities derived from these estimates, such as fitted probabilities, will have larger

Table 3. Estimates and standard errors of estimates of the bet fit (K) with corresponding linear logistic model parameters an their standard errors.

Parameter	Under logistic-normal binomial model		Under liner logistic model	
	Estimate (η)	S.E.	Estimate (η)	S.E.
μ	- 0.9266	0.211	- 0.7691	0.113
Cv. (Kew)	- 1.953	0.953	- 0.0518	0.161
Yr. 4	- 0.1669	0.338	- 1.339	0.204
Cv. (Kew) . Yr 4	2.605	0.567	1.743	0.257

standard errors than they would have had in the absence of overdispersion. The corresponding confidence intervals for these quantities will then be wider than they would have been if no adjustment were made for overdispersion.

It is important to note that the deviance of the saturated model (K) is still much larger that the residual df. The distribution of the residual deviance for the logistic normal binomial model is not well known (Collett, 1991).

Thus, the residual deviance does not necessarily have X^2 distribution. Therefore, deviance of saturated model need not be approximately equal to its degrees of freedom.

CONCLUSION

Epidemiological studies with aggregated spatial patterns of disease incidence are very common in practice (Jeger, 1989). Fitting a binomial distribution (or linear logistic model) does not adequately describe disease incidence when the incidence is aggregated. The use of binomial distribution in such situation gives misleading results.

From the results of the model fitted for pineapple wilt disease it was found that among cultivars used in commercial cultivation, the cultivar Kew is more suceptible to wilt disease than cultivar Murici. In addition the disease incidence increases substantially with time and therefore, the control of the disease at early stages is crucial. The significant interaction component implies that the disease increase over time was more prominent in cultivar Kew when compared to that of cultivar Murici.

REFERENCES

Aitken, M., Anderson, D.A., Francis, B. and Hindle, J.P. (1989). Statistical Modelling in GLIM. Clarendon Press, Oxford.

Cochran, W.G. (1977). Sampling Techniques (3rd Ed.), John Wiley and Sons, New York.

Collett, D. (1991). Modelling Binary Data. Chapman and Hall, London.

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

EGRET (1991). The EGRET System, Revision 3 Manual. Statistics and Epidemiological Research Corporation, Washington.

GLIM (1985). The GLIM System, Release 3.77 Manual. Numerical Algorithems Group: Oxford.

Jeger, M.J. (1989). The spatial component of plant disease epidermics: In SpatialComponent of Plant Disease of Plant Disease Epidermics. pp. 1-13. *In*: Jeger, M.J. (Ed). Spatial Component of Plant Disease Epidermics, Prentice Hall, London.

McCullagh, P. and Nelder, F.A. (1989). Generalized Linear Models (2nd Ed.), Chapman and Hall, London.

SAS. (1990). The Statistical Analysis System, Release 6.04 Manul. SAS Institute Inc. North Carolina, USA.

Snedecor, G.W. and Cochran, W.G. (1989). Statistical Methods (8th Ed.), Iowa State University Press, Ames, Iowa.