ORIGINAL ARTICLE

Linear mean-variance negative binomial models for analysis of orange tissue-culture data

Naratip Jansakul[1] and John P. Hinde[2]

Abstract

Jansakul, N. and Hinde, J.P.
Linear mean-variance negative binomial models for analysis of orange tissue-culture data
Songklanakarin J. Sci. Technol., 2004, 26(5) : 683-696

Negative binomial maximum likelihood regression models are commonly used to analyze overdispersed Poisson data. There are various forms of the negative binomial model with different mean-variance relationships, however, the most generally used are those with linear, denoted by NB1 and quadratic relationships, represented by NB2. In literature, NB1 model is commonly approximated by quasi-likelihood approach. This paper discusses the possible use of the Newton-Raphson algorithm to obtain maximum likelihood estimates of the linear mean-variance negative binomial (NB1) regression model and of the overdispersion parameter. Description of constructing a half-normal plot with a simulated envelope for checking the adequacy of a selected NB1 model is also discussed. These procedures are applied to analyze data of a number of embryos from an orange tissue culture experiment. The experimental design is a completely randomized block design with 3 sugars: maltose, lactose and galactose at dose levels of 18, 37, 75, 110 and 150 µM. The analysis shows that the NB1 regression model with a cubic response function over the dose levels is consistent with the data.

Key words : count data, overdispersion, negative binomial regression

[1]Ph.D.(Statistics), Department of Mathematics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand. [2]M.Sc.(Statistics), Prof., Department of Mathematics, National University of Ireland, Galway, Ireland.
Corresponding e-mail: jnaratip@ratree.psu.ac.th
Received, 20 January 2004 Accepted, 18 June 2004

บทคัดย่อ

นราทิพย์ จั่นสกุล[1] และ John P. Hinde[2]

ตัวแบบทวินามนิเสธที่ความแปรปรวนและค่าเฉลี่ยมีความสัมพันธ์เชิงเส้นตรง กับการวิเคราะห์ข้อมูลการเพาะเลี้ยงเนื้อเยื่อส้ม

ว. สงขลานครินทร์ วทท. 2547 26(5) : 683-696

Negative binomial (NB) regression model เป็นตัวแบบที่นิยมใช้วิเคราะห์ข้อมูลที่มีลักษณะแบบ Poisson ที่มีค่าความแปรปรวนสูงกว่าค่าเฉลี่ย (Overdispersed Poisson data) Negative binomial model มีหลายรูปแบบตามลักษณะของความสัมพันธ์ระหว่างค่าเฉลี่ยและความแปรปรวน แต่รูปแบบที่ใช้กันทั่ว ๆ ไปมากที่สุดสำหรับวิเคราะห์ข้อมูลดังกล่าว ได้แก่ NB models ที่มีความแปรปรวนเป็นฟังก์ชันเชิงเส้น (Linear mean-variance negative binomial models: NB1) หรือเป็นฟังก์ชันพหุนามกำลังสองของค่าเฉลี่ย (Quadratic mean-variance negative binomial models: NB2) โดยทั่วไปการวิเคราะห์ข้อมูลโดยใช้ NB1 model ค่อนข้างซับซ้อนจึงมักจะถูกประมาณด้วยวิธี Quasi-likelihood การวิจัยครั้งนี้เน้นที่การใช้ Newton-Raphson algorithm ในการหาค่าประมาณความควรจะเป็นสูงสุดของสัมประสิทธิ์การถดถอย และของ Overdispersion parameter ของ NB1 model, ตรวจสอบความถูกต้องของตัวแบบโดยใช้ Half-normal plot with a simulated envelope และประยุกต์ใช้กระบวนการเหล่านี้วิเคราะห์ข้อมูลเกี่ยวกับจำนวนตัวอ่อนที่เจริญจากการเพาะเลี้ยงเนื้อเยื่อส้ม (Orange tissue culture data) ข้อมูลที่ใช้ในการวิเคราะห์รวบรวมจากการแผนการทดลองแบบ Completely randomized block ที่มีน้ำตาล 3 ชนิด ประกอบด้วย Maltose, Lactose และ Galactose เป็นบล็อก และระดับความเข้มข้นของน้ำตาลเหล่านี้ที่ 18, 37, 75, 110 และ 150 μM เป็น treatment ผลการวิเคราะห์พบว่าตัวแบบที่เหมาะสมสำหรับข้อมูลชุดนี้คือ NB1 regression ที่เป็นฟังก์ชันพหุนามกำลังสามของระดับน้ำตาล ซึ่งได้ผลตรงกันกับข้อมูลที่ได้จากการทดลอง

[1]ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ อำเภอหาดใหญ่ จังหวัดสงขลา 90112 [2]Department of Mathematics, National University of Ireland, Galway, Ireland.

Negative binomial (NB) models are very widely used for analyzing overdispersed Poisson counts as all important statistical inferences can be carried out more easily and conveniently than for other types of compound Poisson models (Lawless, 1987). Applications using the NB distribution can be found in many areas, for instance, economics (Hausman *et al*., 1984), political science (King, 1988 and King, 1989), psychology (Gardner *et al*., 1995) and biostatistics (Alexander *et al*., 2000). The NB model can be considered as arising from a two-stage model assuming the counts to come from a Poisson distribution with varying mean. Taking the Poisson mean as a gamma distributed random variable leads to the NB model and we can obtain various forms of mean-variance relationship, in particular both linear and quadratic, depending on assumptions about the gamma mixing distribution parameters. The linear mean-variance NB model is obtained by allowing the gamma shape parameter to vary across observations and keeping the scale parameter constant, whereas the quadratic form arises from taking the shape parameter as constant and letting the scale vary. These two variance function models can lead to different models for the mean and also different forms of some associated statistics. Here we will denote the NB model with the linear variance by NB1 and the quadratic variance one by NB2. The NB2 model is a generalized linear model (glm) (Hinde and Demetrio, 1998) when the shape parameter is known. The parameter estimates for the NB2 model can be easily obtained using a full Newton-Raphson method, for example as is in Lawless (1987), or an iterative glm fitting procedure as in Hinde (1996).

This paper concentrates on the maximum likelihood fitting of NB1 models and their application to a real data set. The paper begins in Section 1 with a short review of Poisson regression and overdispersion. Section 2 describes NB1 models and parameter estimation using a Newton-Raphson procedure. Methods of selecting an appropriate model are described in Section 3. To check the adequacy of a selected model, we propose the use of a half-normal plot with a simulated envelope. Details of the construction of this plot are given in Section 4. In Section 5, we consider the application of the NB1 model to a set of orange tissue-culture data. The paper concludes with a brief discussion.

1. Poisson regression and Overdispersion

The random variables Y_i, $i = 1, 2, ..., n$, represent counts with means μ_i, and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ is an associated vector of covariates, with x_{i1} typically equal 1 to include the usual constant term in the model. The standard Poisson regression model assumes that $Y_i \sim \text{Pois}(\mu_i)$, and is a generalized linear model with variance function

$$\text{Var}(Y_i) = \text{Var}(\mu_i) = \mu_i. \qquad (1)$$

The μ_i are typically modelled through the canonical log link function by

$$\eta_i = \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a p vector of unknown parameters. The maximum likelihood estimate of $\boldsymbol{\beta}$ is easily obtained using iteratively reweighted least squares (IRLS) and the asymptotic covariance matrix $\text{Cov}(\beta)$ is $(X^T\hat{W}X)^{-1}$, where $\hat{W}$ is an $n \times n$ diagonal matrix with i^{th} diagonal element $\hat{w}_i = \hat{\mu}_i$, the iterative weight used in the IRLS procedure, see Dobson (1990).

For an appropriate well fitting model, we would expect that the residual deviance (minus twice the difference between the log-likelihood of the maximal model and an estimated model) and the Pearson chi-square defined by $X^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$ would be approximately equal to the degrees of freedom (df). If the residual deviance and X^2 statistic exceed the df, the Poisson regression model may not be adequate, either through some systematic lack of fit, or because the strong assumption from the Poisson model that $\text{Var}(\mu_i) = \mu_i$ is inappropriate; in this case the data are described as overdispersed. If the residual deviance is less than its df, it implies that there is underdispersion in the counts, i.e. the observed variance is less than the nominal Poisson variance. However, in practice, underdispersion is less common, (McCullagh and Nelder, 1989).

In general, when there is overdispersion and we fail to take it into account, it can lead to misinterpretation of the fitted model (Cox, 1983) since the overdispersion produces:

1) smaller standard errors of the parameter estimates than the true values. Therefore we may incorrectly choose explanatory variables for the model that are not required;

2) too large a reduction of deviance associated with model selection tests. This again leads to selecting overly complex models.

To take account of overdispersion, there are a number of different models and associated parameter estimation methods that have been proposed in literature, for example, Lawless (1987), Piegorsch (1990); Hinde and Demetrio (1998), Adamidis (1999) and Thurston *et al*. (2000). These models are extensions of the standard Poisson regression and give a more general form of Poisson variance function. However, associated parameter estimation methods are discussed mostly for NB2 regression since this model is related to a glm.

2. Linear Mean-Variance NB Models

If Y_i, $i = 1, 2, ..., n$, are now negative binomial distributed counts with mean μ_i, and dispersion parameter α a general form of the probability mass function (p.m.f.) of $Y_i \sim \text{NB}(\mu_i, \alpha)$ is given by

$$f(y_i,\mu_i,\alpha)=\begin{cases}\dfrac{\Gamma(y_i+\alpha^{-1}\mu_i^{1-k})}{y_i!\Gamma(\alpha^{-1}\mu_i^{1-k})}\dfrac{\alpha^{y_i}\mu_i^{ky_i}}{(1+\alpha\mu_i^{k})^{y_i+a^{-1}\mu_i^{1-k}}}, & y_i=0,1,...,\alpha>0\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

with $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i (1 + \alpha\mu_i^k)$. Here α is assumed to be a constant. The index k identifies various forms of the NB distribution, but two well-known models are given by $k = 0$ and 1. For $k = 0$ we have a linear-variance NB regression, or NB1 model, with $Var(Y_i) = \mu_i (1+\alpha)$ (this is often approximated by fitting the constant overdispersion quasi-likelihood (QL) model with $Var(Y_i) = \phi\mu_i$, where ϕ is a constant). Taking $k = 1$ gives the more usual quadratic-variance NB, or NB2 model, with $Var(Y_i) = \mu_i(1+\alpha\mu_i)$. As $\alpha \to 0$, the NB model reduces to the Poisson model. For both models, we assume some specific regression model for the mean, i.e. $\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

2.1 Maximum Likelihood Estimation for the NB1 Distribution

Taking $k = 0$ in (2), the p.m.f. of $Y_i \sim$ NB1 (μ_i, α) is specified by

$$f(y_i,\mu_i,\alpha)=\begin{cases}\dfrac{\Gamma(y_i+\alpha^{-1}\mu_i)}{y_i!\Gamma(\alpha^{-1}\mu_i)}\dfrac{\alpha^{y_i}}{(1+\alpha)^{y_i+a^{-1}\mu_i}}, & y_i=0,1,...,\alpha>0\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

For observed values $y_1, y_2, ..., y_n$, the NB1 log-likeli-hood, $\ell = \ell(\boldsymbol{\mu}, \alpha)$, is given by

$$\ell=\sum_{i=1}^{n}\{y_i\log a-\left(y_i+\frac{\mu_i}{\alpha}\right)\log(1+a)+\mathrm{dlg}(y_i,\ a^{-1}\mu_i)-\log y_i!\}, \tag{4}$$

where $\mathrm{dlg}(y, a) = \log \Gamma(y+a) - \log \Gamma(a)$.

The NB1 is not a standard glm-type exponential family distribution, even when the overdispersion parameter α is known, and standard glm fitting methods will not apply. So here we consider a general Newton-Raphson iterative scheme. The first and second derivatives with respect to the underlying parameters are

$$\frac{\partial\ell}{\partial\beta_j}=\sum_i\left\{\alpha^{-1}\mathrm{ddg}(y_i,a^{-1}\mu_i)-\frac{\log(1+\alpha)}{\alpha}\right\}\mu_i x_{ij},\ j=1,\ 2,\ ...,\ p \tag{5}$$

$$\frac{\partial^2\ell}{\partial\beta_j\partial\beta_k}=-\sum_i\left\{\alpha^{-1}\mathrm{ddg}(y_i,a^{-1}\mu_i)-\frac{\log(1+\alpha)}{\alpha}-\alpha^{-2}\mu_i\mathrm{dtg}(y_i,\alpha^{-1}\mu_i)\right\}\mu_i x_{ij}x_{ik},\quad j,\ k=1,\ 2,\ ...,\ p, \tag{6}$$

$$\frac{\partial\ell}{\partial\alpha}=-\alpha^{-2}\sum_i\left\{\frac{\mu_i-y_i}{1+\alpha^{-1}}-\mu_i\log(1+\alpha)+\mu_i\mathrm{ddg}(y_i,\alpha^{-1}\mu_i)\right\}, \tag{7}$$

$$\frac{\partial^2\ell}{\partial\alpha^2}=\sum_i\left\{2\alpha^{-3}\left[\frac{\mu_i-y_i}{1+\alpha^{-1}}-\mu_i\log(1+\alpha)+\mu_i\mathrm{ddg}(y_i,\alpha^{-1}\mu_i)\right]+\alpha^{-4}\left[\frac{y_i-\mu_i}{(1+\alpha^{-1})^2}+\frac{\alpha\mu_i}{1+\alpha^{-1}}-\mu_i^2\mathrm{dtg}(y_i,\alpha^{-1}\mu_i)\right]\right\}, \tag{8}$$

and

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \alpha} = \alpha^{-2} \sum_i \left\{ \left[\log(1+\alpha) - \mathrm{ddg}(y_i, \alpha^{-1}\mu_i) \right] + \alpha^{-1} \left[\mu_i \mathrm{dtg}(y_i, \alpha^{-1}\mu_i) - \frac{\alpha}{1+\alpha} \right] \mu_i x_{ij} \right\}, \quad (9)$$

where ddg(y,a) and dtg(y,a) denote the differences of the di-gamma and tri-gamma functions. These are defined by

$$\mathrm{ddg}(y,a) = \frac{\partial}{\partial a}(\mathrm{dlg}(y,a)) = \vartheta(y+a) - \vartheta(a)$$

$$= \begin{cases} 0, & y = 0 \\ \sum_{t=0}^{y-1} (a+t)^{-1}, & y > 0, \end{cases}$$

where ϑ is the di-gamma function, and

$$\mathrm{dtg}(y,a) = \frac{\partial^2}{\partial a^2}(\mathrm{dlg}(y,a)) = \zeta(y+a) - \zeta(a)$$

$$= \begin{cases} 0, & y = 0 \\ \sum_{t=0}^{y-1} (a+t)^{-2}, & y > 0, \end{cases}$$

where ζ is the tri-gamma function.

Let s(β,α) be the vector of score functions defined by

$$s(\beta,\alpha) = \begin{bmatrix} s_\beta(\beta,\alpha) \\ s_\alpha(\beta,\alpha) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \ell}{\partial \beta} \\ \dfrac{\partial \ell}{\partial \alpha} \end{bmatrix},$$

and let $I(\beta,\alpha)$ be the $(p+1) \times (p+1)$ observed information matrix, which we partition as

$$I(\beta,\alpha) = \begin{bmatrix} I_{\beta\beta}(\beta,\alpha) & I_{\beta\alpha}(\beta,\alpha) \\ I_{\alpha\beta}(\beta,\alpha) & I_{\alpha\alpha}(\beta,\alpha) \end{bmatrix}, \quad (10)$$

where $I_{\beta\beta} = -\dfrac{\partial^2 \ell}{\partial \beta \, \partial \beta^T}$ is the $p \times p$ symmetric matrix,

$I_{\alpha\alpha} = -\dfrac{\partial^2 \ell}{\partial \alpha^2}$ is a scalar and $I_{\alpha\beta} = I_{\beta\alpha}^T = -\dfrac{\partial^2 \ell}{\partial \alpha \, \partial \beta}$, is a $1 \times (p+1)$ matrix.

Writing $\beta^{(m)}$ and $\alpha^{(m)}$ as the estimates at the m^{th} iteration, the standard Newton-Raphson iterative scheme gives

$$\begin{bmatrix} \beta^{(m+1)} \\ \alpha^{(m+1)} \end{bmatrix} = \begin{bmatrix} \beta^{(m)} \\ \alpha^{(m)} \end{bmatrix} + \left[I^{(m)} \right]^{-1} s^{(m)}, \quad (11)$$

where $I^{(m)}$ and $s^{(m)}$ are $I(\beta,\alpha)$ and s(β,α) evaluated at $\beta = \beta^{(m)}$ and $\alpha = \alpha^{(m)}$. The iteration (11) must be carried out until convergence, which can be assessed using a stopping rule such as

$$\left| \alpha^{(m+1)} - \alpha^{(m)} \right| < \in \text{ or } \left| \ell^{(m+1)} - \ell^{(m)} \right| < \in.$$

The procedure requires good initial values, which can be obtained as follows:

- β; fit a standard Poisson regression model to obtain $\hat{\beta}^{(0)}$ and initial estimates of the fitted values $\hat{\mu}_i^{(0)}$.
- α; equate the Pearson X^2 statistic from the Poisson fit to its expected value under the NB1 model, to give

$$\alpha^{(0)} = (n-p)^{-1} \sum_i \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\hat{\mu}_i^{(0)}} - 1,$$

this is simply based on the quasi-likelihood estimate of the overdispersion parameter from the constant overdispersion Poisson model.

The asymptotic variance of $\hat{\beta}$ and $\hat{\alpha}$ are the diagonal elements of $I^{-1}(\hat{\beta}, \hat{\alpha})$, and are automatically provided at the final iteration. This iterative procedure is simply implemented in any computer software that can handle matrices, such as, Splus and the free software **R** (R Development Core Team, 2003)

3. Selecting an Appropriate Model

Testing the Poisson assumption against the NB1 alternative corresponds to testing

$$H_0 : \alpha = 0 \quad \text{against} \quad H_1 : \alpha > 0.$$

The commonly used test statistics, the likelihood ratio test (LRT) or residual deviance, normally used in the context of statistical modeling, defined by $-2\{\ell(\hat{\mu})-\ell(\hat{\mu},\hat{\alpha})\}$, where $\ell(\hat{\mu})$ and $\ell(\hat{\mu},\hat{\alpha})$ are maximized likelihood estimates under the Poisson and NB1 model, respectively, and the Wald test specified by $\frac{\hat{\alpha}^2}{\hat{\text{Var}}(\hat{\alpha})}$, are both applicable here. Some care is required as the null hypothesis is on the boundary of the parameter space (e.g. the null distribution of the LRT is not the usual χ^2_1 distribution), and also the alternative hypothesis is one-sided as we are only testing for overdespersion.

Selecting an appropriate model among all possible NB1 regression models is straight-forward using the standard likelihood criteria, for example, Akaike information criterion (AIC) (Akaike, 1973) or Baysian information criterion (BIC) given in Schwarz (1978). These criteria simply require the maximized log-likelihood value from the NB1 distribution fit and are defined as:

AIC = -2 ℓ +2(number of fitted parameters),

BIC = -2 ℓ +log(n) × (number of fitted parameters).

The fitted model with smallest value of AIC or of BIC is preferable. In addition, both AIC and BIC are also applicable for selecting between non-nested models (Lindsey, 1997, pp.208-209) where we will illustrate the use of these criteria to select between NB1 and NB2 model.

4. Model Checking

A model diagnostic technique that has been found to be useful for checking the adequacy of fitted models is the use of half-normal plots with a simulated envelope. This technique was first proposed by (Atkinson, 1985). He applied the plot to check model adequacy using Pearson residuals or Cook's statistics in normal regression. The technique was further developed for glms using (standardized) Pearson residuals and (standardized) deviance residuals by Williams (1987). Williams claimed that the plot could detect both outliers and overdispersion in both Poisson and binomial regression models.

Even though the NB1 regression model is not a glm, we can define its complete p.m.f. and hence the log-likelihood function. The associated (standardized) Pearson residual, or the standardized studentized residual, for the NB1 model can be obtained by using the general definition, $(y-\hat{\mu})/\sqrt{\hat{\text{Var}}(\text{Y})}$ (Lawless, 1987). Denoting the standardized Pearson residual for an NB1 fit by $\dot{r}_{Pi}$, the i^{th} component is

$$\dot{r}_{Pi}=\frac{Y_i-\hat{\mu}_i}{\sqrt{\hat{\mu}_i(1+\hat{\alpha})}}.$$

NB1 deviance residuals cannot be obtained simply based on the usual deviance expression for glms: -2{ $\ell(\mu, \alpha; y)$ - $\ell(y, \alpha ; y)$}, as some of the individual components can be negative. Nelder (1991) pointed out that the log-likelihood (4) does not have the property that its mode occurs at $\mu = y$ unless $y = 0$. He used $y_i+\frac{1}{2}$ as the approximate mode of ℓ and then approximated the deviance component for y_i by

$$\begin{cases}\dfrac{2\mu_i\log(1+\alpha)}{\alpha}, & y_i=0\\[2ex] -2\times\left\{\left[y_i+\dfrac{1}{2}-\mu_i\right]\dfrac{\log(1+\alpha)}{\alpha}+\text{dlg}(y_i,\alpha^{-1}\mu_i)-\text{dlg}(y_i,\alpha^{-1}(y_i+\dfrac{1}{2})\right\}, & y_i>0.\end{cases}$$

However, our exploration (will be reported elsewhere) found that $y+\frac{1}{2}$ is not an adequate approximation of the mode. The investigations indicated that there is no simple form for the mode of ℓ, but values such as $y+c$, where $\frac{\alpha}{2+1/y} < c < \frac{\alpha}{2}$ are likely to be close to giving the mode, and for large y, $c \approx \frac{\alpha}{2}$ works well. Using this simple form gives the deviance residuals $r_{D,i} = \text{sgn}(y_i - \mu_i)\sqrt{D_i}$ for the NB1 model, where

$$D_i = \begin{cases} \dfrac{2\hat{\mu}_i \log(1+\hat{\alpha})}{\hat{\alpha}}, & y_i = 0 \\ -2\times\left\{\left[y_i + \dfrac{\hat{\alpha}}{2} - \hat{\mu}_i\right]\dfrac{\log(1+\hat{\alpha})}{\hat{\alpha}} + \text{dlg}(y_i, \hat{\alpha}^{-1}\hat{\mu}_i) - \text{dlg}\left(y_i, \hat{\alpha}^{-1}\left(y_i + \dfrac{\hat{\alpha}}{2}\right)\right)\right\}, & y_i > 0. \end{cases}$$

Following the general procedure for constructing half-normal plots with a simulated envelope given in Vieira *et al*. (2000), a plot for checking a selected NB1 model using (standardized) deviance residuals can be obtained as follows:

- Fit a NB1 model to obtain $\hat{\mu}_i, \hat{\alpha}$ and calculate the ordered absolute values of deviance residuals $r_{D,i}$;
- Simulate nineteen samples for the response variable under the fitted model, by first generating $e_{0j,i}$ where $e_{0j,i} \sim \Gamma(\hat{\alpha}^{-1}\hat{\mu}_i, 1)$, $j = 1,2,...,19$, $i = 1, 2, ..., n$, calculating $e_{ji} = e_{0j,i} \times \hat{\alpha}\hat{\mu}_i^{-1}$ then simulating $Y_{ji} \sim \text{Pois}(\hat{\mu}_i e_{ji})$ to give $Y_{ji} \sim \text{NB1}(\hat{\mu}_i, \hat{\alpha})$, i.e. 19 datasets based on the fitted model.
- Refit the model, using the same explanatory variables, to each sample and calculate the ordered absolute values of the deviance residuals, $r^*_{j(D,i)}$, $j = 1, 2, ..., 19$, $i = 1, 2, ..., n$;
- For each i calculate the minimum, maximum and the mean of the $r^*_{j(D,i)}$;
- Plot these values and the observed $r_{D,i}$ against the half-normal scores (expected order statistics); $\Phi^{-1}\{(i+n-\frac{1}{8})/(2n+\frac{1}{2})\}$, where Φ is the normal cumulative density function (Demetrio and Hinde, 1997).

If the selected model is adequate, the observed $r_{D,i}$ should lie within the simulated envelope.

Demetrio and Hinde (1997) gave a GLIM macro to construct such plots with special emphasis on overdispersed models (i.e. constant overdispersion Poisson and NB2 models for extra Poisson variation). These are easily adapted for the NB1 deviance residuals as all that is required are two functions, one to calculate the NB1 deviance residuals and the other to simulate from an NB1 distribution; these are easily implemented in software such as R.

5. Application: An Orange Tissue-culture Experiment

The orange variety *Valencia* was used in a tissue-culture experiment conducted in Brazil to study the effect of six carbohydrate sources (maltose, glucose, galactose, lactose, sucrose and glycerol) on the stimulation of somatic embryos from callus cultures. The response variable is the number of embryos observed after approximately four weeks. The experiment was a completely randomized block design with the above six sugars at dose levels of 18, 37, 75, 110 and 150 μM for the first five and 6, 12, 24, 36, and 50 μM for the glycerol, and 5 replicates of each treatment, see Tomaz *et al*. (2001), for further details of the experiment and histological analysis. The main interest was in the dose-response relationship for

the sugars (maltose, galactose and lactose) that produced large numbers of embryos. The number of embryos produced is highly variable, see Figure 1, with marked differences between the three sugars. Table 1 presents the mean and variance of the number of embryos, classified by sugars and dose levels (excluding 1 missing value). Most of the sample overdispersion index values (relative to a baseline Poisson distribution) exceed 3, and give strong evidence of overdispersion. In their analysis Tomaz *et al*. (2001) used a quadratic response function over the dose levels and a simple constant overdispersion Poisson model fitted by quasi-likelihood to take account of overdispersion. Here we use this data set to illustrate the use of maximum likelihood estimation for the NB1 distribution.

Writing $\boldsymbol{\mu}$ for the vector of the mean numbers of embryos and taking sugar (S) and DOSE as factors, fitting the full interaction Poisson regression model ($\log(\boldsymbol{\mu})$ = S * DOSE), the residual deviance is 298.04 on 29 df (p-value of 0.00, based on χ^2_{29}, distribution), which as expected shows strong evidence of overdispersion. The half-normal plot, Figure 2(a), also indicates greater variation than in the Poisson model as all the Poisson deviance residuals lie above the upper envelope.

Fitting the corresponding NB1 and NB2 model with the full interaction between dose and sugar gives a likelihood ratio test statistic for overdispersion of 166.59 and 138.85 on 1 df, respectively. The models certainly fit the data much better than the Poisson model. This full interaction model is equivalent to fitting a model with an interaction between sugar and a quartic polynomial over the actual dose levels: ($\log(\boldsymbol{\mu})$ = S * (D + D^2 + D^3 + D^4)). This suggests that we might consider

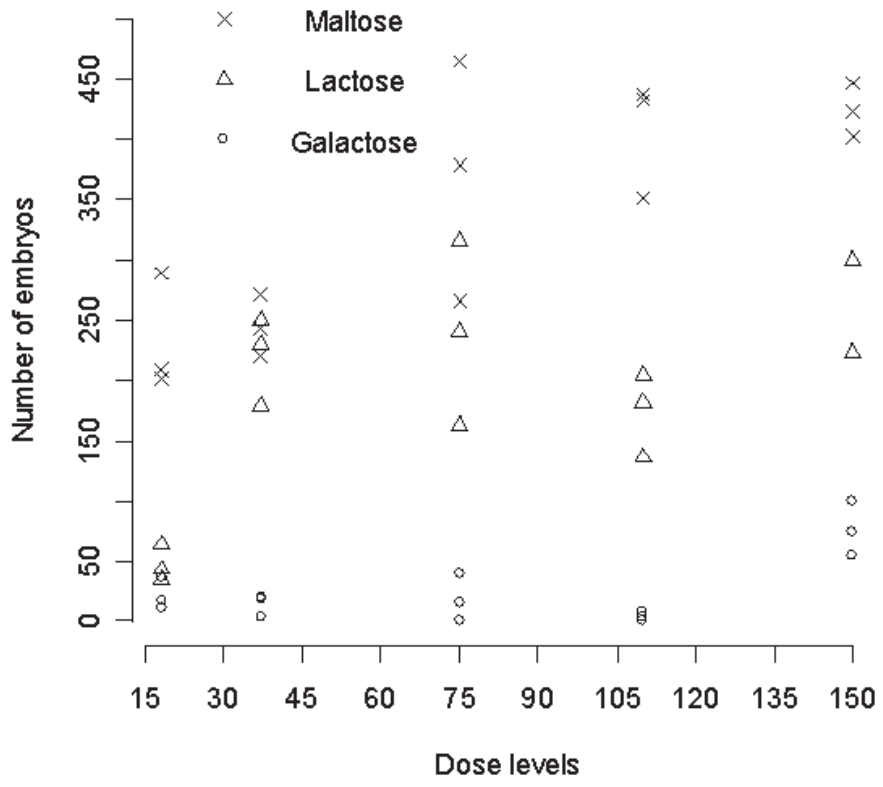

Figure 1. Orange (*Valencia*) tissue culture data: Observed number of embryos classified by types of sugars and dose levels.

Table 1. Orange (*Valencia*) tissue culture data: Mean and variance (Var) of number of embryos classified by sugars and dose levels.

Sugars		Dose levels (μM)				
		18	**37**	**75**	**110**	**150**
Maltose	Mean	233.00	245.33	369.67	407.00	424.33
	Var	2368.00	654.33	9952.33	2356.00	506.33
	o.i.	10.16	2.67	26.92	5.79	1.19
Lactose	Mean	47.33	219.33	239.33	174.33	260.50
	Var	224.33	1310.33	5854.33	1234.33	2964.50
	o.i.	4.74	5.97	24.46	7.08	11.38
Galactose	Mean	21.67	14.00	18.33	4.00	75.67
	Var	185.33	76.00	408.33	13.00	508.30
	o.i.	8.55	5.43	22.28	3.25	6.72

***o.i.* denotes overdispersion index** $= \frac{\text{Var}}{\text{Mean}} - 1$.

simplifying the model by fitting lower order polynomials over the dose levels.

Considering NB regression, the best NB2 model is log (μ) = S * (D + D^2 + D^3 + D^4) with the smallest AIC and BIC values of 466.293 and 494.840, respectively, whereas the best model fitted based on the NB1 regression is the cubic response function over the dose levels: log(μ) = S * (D + D^2 + D^3) with AIC and BIC of 439.000 and 462.194 respectively, see Table 2. Using AIC and BIC to choose between these models suggests that the NB1 log cubic model is preferable. This is also the model chosen using a QL approach with constant overdispersion, but the different model in Tomaz *et al.* (2001). The corresponding half-normal plot, Figure 2(b), indicates that the NB1 model is more consistent with the data than the NB2, shown in Figure 2(c). The selected model gives a separate cubic dose relationship for each of the sugar types.

However, the plot of the predicted mean number of embryos for NB1 of cubic model

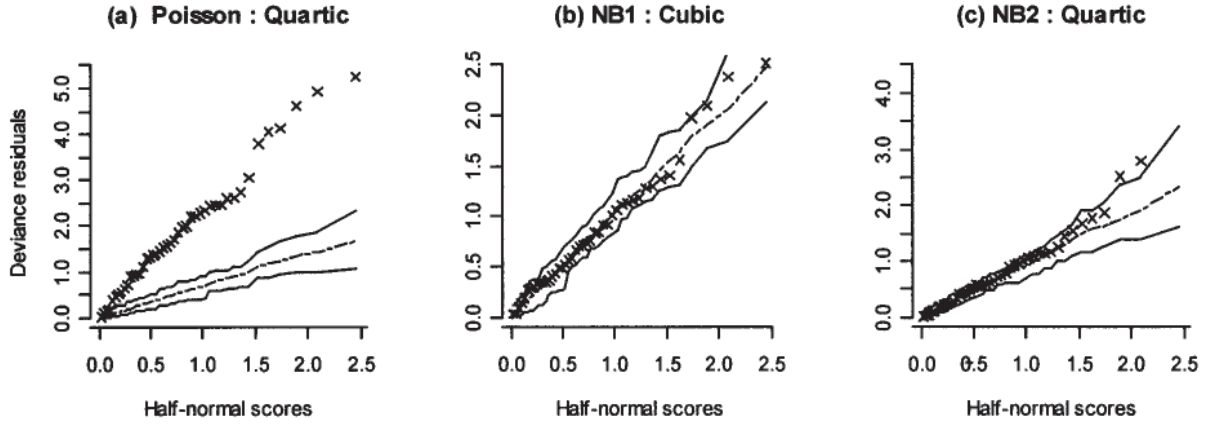

Figure 2. Orange (*Valencia*) tissue culture data: Half-normal plots based on Poisson, NB1 and NB2 model.

Table 2. Orange (*Valencia*) tissue-culture data: Statistics for Poisson and overdispersed models.

S is a three-level factor for sugar
DOSE is a five-level factor for the dose levels
D is a variate for the dose level

Description	Models		-2 ℓ	df	AIC	BIC
	log (μ)	α				
Poisson	S * (D + D^2 + D^3 + D^4)†	0	573.143	29	**603.143**	**629.906**
	S * (D + D^2 + D^3)	0	648.715	32	672.715	694.125
	S * (D + D^2)	0	959.656	35	977.656	993.414
	S * D	0	1182.815	38	1194.815	1205.520
NB1	S * (D + D^2 + D^3 + D^4)	6.331	406.552	28	438.552	467.099
	S * (D + D^2 + D^3)	7.772	413.000	31	**439.000**	**462.194**
	S * (D + D^2)	15.360	438.385	34	458.385	476.227
	S * D	24.410	457.234	37	471.234	483.724
NB2	S * (D + D^2 + D^3 + D^4)	0.060	434.293	28	**466.293**	**494.840**
	S * (D + D^2 + D^3)	0.114	451.576	31	477.576	500.771
	S * (D + D^2)	0.244	473.058	34	493.058	510.900
	S * D	0.373	487.576	37	501.576	514.066
		$\hat{\phi}$	Deviances	df		
Constant QL	S * (D + D^2 + D^3 + D^4)	9.717	30.671	28	-	-
	S * (D + D^2 + D^3)	9.717	38.448	31	-	-
	S * (D + D^2)	9.717	70.446	34	-	-
	S * D	9.717	93.411	37	-	-

S * (D + D^2 + D^3 + D^4)† is equivalent to S*DOSE

shown in Figure 3 suggests that the dose response relationship for maltose and galactose may be approximately linear or quadratic.

In order to investigate this, we fitted NB1 regression models with cubic, quadratic and linear functions over the dose levels using the constant overdispersion estimate ($\hat{\alpha}$ = 7.77), see Table 3. The dose levels here are transformed to standardized values, denoted by $\tilde{D}$, to avoid convergence problems in the maximum likelihood estimation procedure. The best NB1 model suggested by AIC and BIC for each sugar is different; a log-quadratic model in the dose levels for maltose and galactose and a log-cubic model for lactose. In addition, lactose requires a medium level of dose for the optimum production of embryos, while maltose and galactose need a high dose level, the corresponding parameter estimates and standard errors are presented in Table 4.

Conclusion and Discussion

The paper gives a framework for NB1 regression including estimation, model selection and use of an approximate deviance residual function as a model diagnostic. Fitting NB1 models using a Newton-Raphson iterative procedure is conveniently performed in any computer software that can deal with matrices, in particular, R or SPlus, as the commands for calculating di-gamma and tri-gamma functions are also available. The correct asymptotic covariance matrix of the

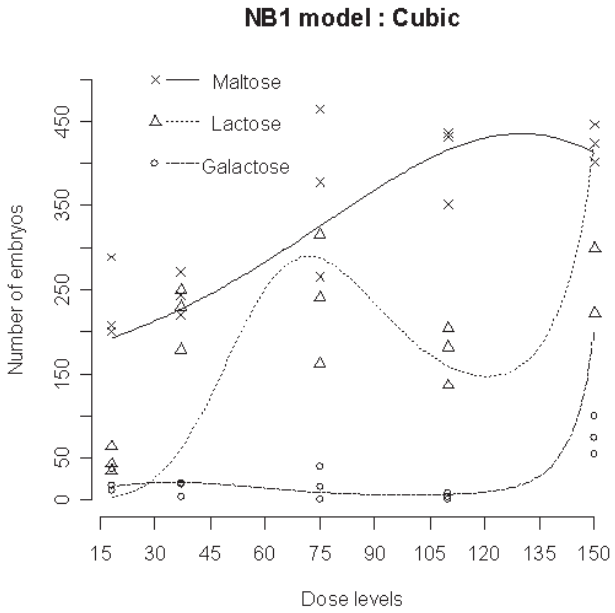

Figure 3. Orange (*Valencia*) tissue culture data: Observed (*symbols*) and estimated (*lines*) values of embryogenic responses

Table 3. Orange (*Valencia*) tissue-culture data: Statistics for NB1 models with fixed $\hat{\alpha}$ = 7.77, classified by sugar.

Sugars	Models (log(μ)	-2 ℓ	df	AIC	BIC
Maltose	$\tilde{D}+\tilde{D}^2+\tilde{D}^3$	157.30	11	166.30	169.13
	$\tilde{D}+\tilde{D}^2$	158.44	12	**164.44**	**166.57**
	$\tilde{D}$	165.74	13	**165.74**	**167.15**
Lactose	$\tilde{D}+\tilde{D}^2+\tilde{D}^3$	142.30	10	**150.30**	**152.85**
	$\tilde{D}+\tilde{D}^2$	174.97	11	180.97	182.88
	$\tilde{D}$	184.92	12	188.92	190.19
Galactose	$\tilde{D}+\tilde{D}^2+\tilde{D}^3$	112.41	11	120.41	123.24
	$\tilde{D}+\tilde{D}^2$	114.70	12	**120.70**	**122.82**
	$\tilde{D}$	140.42	13	144.42	145.80

$\tilde{D}$ **denotes a vector of standardized** $\tilde{D}_i$; $\tilde{D}_i = \frac{D_i - \overline{D}}{\sqrt{Var(D)}}$, $i = 1, 2, ..., n$,

where $\overline{D} = n^{-1} \sum_i D_i$.

parameter estimates Cov($\hat{\beta}, \hat{\alpha}$) is automatically provided at the final iteration. Moreover, the robust and empirical covariance matrices can be easily implemented.

The comparison between NB1 and NB2, and even QL, models discussed in the application shows that a more general framework for NB regression modelling with some ability to choose

Table 4. Orange (*Valencia*) tissue-culture data: Parameter estimates and their standard errors (given in the parentheses).

	Maltose	Lactose	Galactose
$\tilde{D}^3$	-	0.677 (0.126)	-
$\tilde{D}^2$	-	-0.639 (0.106)	1.030 (0.226)
$\tilde{D}$	0.226 (0.040)	-0.606 (0.197)	0.050 (0.149)
constant	5.783 (0.043)	5.562 (0.086)	1.793 (0.371)
$\hat{\alpha}$	7.772 (1.941)	7.772 (1.941)	7.772 (1.941)

between different variance functions can be useful. A more detailed study of this is being undertaken and will be reported elsewhere.

Acknowledgements

The authors are grateful to M.L. Tomaz and B. M. J. Mendes for providing orange *Valencia* tissue culture data. We also thank the anonymous referees for kind suggestions and comments.

References

Adamidis, K. 1999. An EM algorithm for estimating negative binomial parameters, Australian and New Zealand J of Statistics, 41: 213-221.

Akaike, H. 1973. Information theory and extension of the maximum likelihood principle. In: B.N. Petrov & F. Csaki (Eds.), Proceedings 2nd International Symposium on Inference Theory. Budapest: Akadmiai Kiado, 1973: 267-281.

Alexander, N., Moyeed, R. and Stander, J. 2000. Spatial modelling of individual-level parasite counts using the negative binomial distribution, Biostatistics, 1: 453-463.

Atkinson, A. 1985. Plots, Transformations and Regression. An introduction to graphical methods of diagnostic regression analysis, Oxford: Clarendon Press.

Cox, D.R. 1983. Some remarks on overdispersion, Biometrika, 70: 269-274.

Demetrio, C.G.B. and Hinde, J.P. 1997. Half-normal plots and overdispersion, GLIM Newsletter, 27: 19-26.

Dobson, A.J. 1990. An Introduction to Generalized Linear Models, Chapman and Hall, London.

Gardner, W., Mulvey, E.P. and Shaw, E.C. 1995. Regression analyses of counts and rates: Poisson, Overdispersed Poisson, and Negative binomial models, Psychological Bulletin, 11: 392-404.

Hausman, J., Hall, B.H. and Griliches, Z. 1984. Econometric models for count data with an application to the patents-R&D relationship, Econometrica, 52: 909-938.

Hinde, J.P. 1996. Macros for fitting overdispersion models, GLIM Mewsletter, 26: 10-19.

Hinde, J.P. and Demetrio, C.G.B. 1998. Overdispersion: Models and Estimation, Computational Statistics and data Analysis, 27: 151-170.

King, G. 1988. Statistical models for Political Science event counts: Bias in conventional procedures and evidence for the experimental Poisson regression model, American J of Political Science, 32: 838-863.

King, G. 1989. Variance specification in event count models: From restrictive assumptions to a generalized estimator, American J of Political Science, 33: 762-784.

Lawless, J.F. 1987. Negative binomial and mixed Poisson regression, The Canadian J of Statistics, 15: 209-225.

Lindsey, J.K. 1997. Applying Generalized Linear Models, Springer-Verlag, New York.

McCullagh, P. and Nelder, J.A. 1989. Generalized Linear Models, 2nd ed., London: Chapman & Hall.

Nelder, J.A. 1991. Generalized linear models with negative binomial or beta-binomial errors, 10 pages, Department of Mathematics, Imperial College of Science, Technology and Medicine.

Piegorsch, W.W. 1990. Maximum likelihood estimation for negative binomial dispersion parameter, Biometrics, 46: 863-867.

R Development Core Team 2003. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, http://www.R-project.org

Schwarz, G. 1978. Estimating the dimension of a model, Annals of Statistics, 6: 461-464.

Thurston, S.W., Wand, M.P. and Wiencke, J.K. 2000., Negative binomial additive models, Biometrics, 56: 139-144.

Tomaz, M.L., Mendes, B.M.J., Mourão Filho, F.A., Demetrio, C.G.B., Jansakul, N. and Rodriguez, A.P.M. 2001. Somatic embryogenesis in *citrus* spp: Carbohydrate stimulation and histo-differentiation. In Vitro Cellular & Developmental Biology-Plant, 37: 446-452.

Vieira, A.M.C., Hinde, J.P. and Demetrio, C.G.B. 2000. Zero-inflated proportion data models applied to a biological control assay, J of Applied Statistics, 27: 373-389.

Williams, D.A. 1987. Generalized linear model diagnostics using the deviance and single case deletions, Applied Statistics, 36: 181-191.

Appendix

Lists of Notations

AIC	:	Akaike information criterion
BIC	:	Baysian information criterion
Cov(x, y)	:	Covariance matrix of x and y
Glms	:	generalized linear models
IRLS	:	Iteratively reweighted least squares
QL	:	Quasi likelihood
NB	:	Negative binomial models
NB1	:	Linear mean-variance negative binomial models
NB2	:	Quadratic mean-variance Negative binomial models