The Canadian Journal of Statistics
Vol. 15, No. 3, 1987, Pages 209-225
La Revue Canadienne de Statistique

Negative binomial and mixed Poisson regression

Jerald F. LAWLESS

University of Waterloo

Key words and phrases: Count data, efficiency, overdispersion, quasilikelihood, robustness.
AMS 1980 subject classifications: 62J02, 62'712.

ABSTRACT

A number of methods have been proposed for dealing with extra-Poisson variation when doing regression analysis of count data. This paper studies negative-binomial regression models and examines efficiency and robustness properties of inference procedures based on them. The methods are compared with quasilikelihood methods.

RÉSUMÉ

Plusieurs méthodes ont été proposées en vue de traiter le problème de la variation extra-poissonnienne dans une analyse de régression pour données de dénombrement. Cet article a pour objet l'étude de modèles de régression binomiale négative et se penche sur les propriétés d'efficacité et de robustesse des méthodes inférentielles découlant des modèles. Ces dernières sont comparées aux méthodes de quasi-vraisemblancce.

1. INTRODUCTION

Poisson models are widely used in the regression analysis of count data (e.g. Frome, Kutner, and Beauchamp 1973; Frome 1983; Haberman 1974; Holford 1983). At the same time it is recognized that counts often display substantial extra-Poisson variation, or overdispersion, relative to a Poisson model. Consequently there have been both studies of the effect of overdispersion on inferences made under a Poisson model (e.g. Paul and Plackett 1978; Cox 1983), and models proposed for accommodating overdispersion in statistical analysis. In the latter vein, certain types of negative-binomial regression models are perhaps the most convenient to deal with, and have been used by various authors (e.g. Engel 1984; Lawless 1987; Manton, Woodbury, and Stallard 1981). Other parametric models have also been suggested (e.g. Hinde 1982), and in addition, analysis based on weighted least squares or quasilikelihood has been advocated (e.g. Breslow 1984, McCullagh and Nelder 1983).

Although negative-binomial regression methods have been employed in analyzing data, their properties have not been investigated in any detail. The purpose of this paper is to study negative-binomial regression models, to examine their properties, and to fill in some gaps in existing methodology. Section 2 introduces the model and reviews maximum-likelihood and moment estimation procedures. In Section 3 the asymptotic covariance matrix for weighted least-squares–moment estimators is obtained, and efficiency and robustness properties of them and of maximum-likelihood estimators are studied. Section 4 presents some evidence on the adequacy of large-sample approximations. Section 5 discusses significance tests for Poisson assumptions, and Section 6 looks at some illustrative data sets. Section 7 concludes the paper with some additional remarks on the regression analysis of count data.

2. ESTIMATION FOR A NEGATIVE-BINOMIAL MODEL

Let Y be the response variable, which is a count, and $\boldsymbol{x}$ a $p \times 1$ vector of explanatory variables. A Poisson model would stipulate that the distribution of Y given $\boldsymbol{x}$ is Poisson with mean equal to $\mu(\boldsymbol{x}) = Tg(\boldsymbol{x}; \boldsymbol{\beta})$, where $g(\boldsymbol{x}; \boldsymbol{\beta})$ is a positive-valued function of $\boldsymbol{x}$ and a vector $\boldsymbol{\beta}$ of regression parameters, and T is a measure of exposure. We abbreviate this as $Y \sim \text{Poisson}(\mu(\boldsymbol{x}))$. The log-linear specification $g(\boldsymbol{x}; \boldsymbol{\beta}) = \exp(\boldsymbol{x}^{\mathrm{T}}\beta)$ is widely used. Our objective is to study analogous models which can handle extra-Poisson variation. For count data with no covariates, the negative-binomial distribution is popular for this purpose, and its relationship to the Poisson is well known (e.g. Anscombe 1950). The corresponding negative-binomial regression model considered here is

$$\Pr(Y = y \mid \boldsymbol{x}) = \frac{\Gamma(y + a^{-1})}{y!\Gamma(a^{-1})}\left(\frac{a\mu(\boldsymbol{x})}{1 + a\mu(\boldsymbol{x})}\right)^{y}\left(\frac{1}{1 + a\mu(\boldsymbol{x})}\right)^{a^{-1}}, \qquad y = 0, 1, \ldots, \tag{2.1}$$

where $a \geq 0$ is often referred to as the index or dispersion parameter. (We remark that a more common parametrization uses $k = a^{-1}$, and k is also referred to as an index parameter.) The mean and variance of Y are

$$\mathscr{E}(Y \mid \boldsymbol{x}) = \mu(\boldsymbol{x}) \quad \text{and} \quad \mathscr{Var}\,(Y \mid \boldsymbol{x}) = \mu(\boldsymbol{x}) + a\mu(\boldsymbol{x})^2. \tag{2.2}$$

I will write $Y \sim \text{NB}(\mu(\boldsymbol{x}),a)$ for this model. In the limit as a goes to 0, (2.1) yields the model $\text{Poisson}(\mu(\boldsymbol{x}))$, and $a = 0$ denotes the Poisson case.

The variance-mean relationship embodied in (2.2) often describes data well. More formally, (2.2) arises from a mixed, or random-effects, Poisson model: if v is a positive-valued random variable with mean 1 and variance a, and if the distribution of Y, given v and $\boldsymbol{x}$, is $\text{Poisson}(v\mu(\boldsymbol{x}))$, then the marginal mean and variance of Y given $\boldsymbol{x}$ are as in (2.2). Furthermore, when the distribution of v is gamma, the marginal distribution of Y is $\text{NB}(\mu(\boldsymbol{x}), a)$.

Maximum likelihood for (2.1) is implicit in the work of some earlier authors, but I have not found a full discussion of it. Because of this and the desire to compare maximum likelihood and moment estimates below, I review it briefly here. For simplicity and because of its importance I will work with the log-linear model where $Y_i \sim \text{NB}(\mu_i,a)$, $i = 1, \ldots, n$, are independent, with $\mu_i = T_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Results for other regression specifications are qualitatively similar. The likelihood function is proportional to

$$L(\boldsymbol{\beta}, a) = \prod_{i=1}^{n} \frac{\Gamma(y_i + a^{-1})}{\Gamma(a^{-1})}\left(\frac{a\mu_i}{1 + a\mu_i}\right)^{y_i}\left(\frac{1}{1 + a\mu_i}\right)^{a^{-1}},$$

and noting that for any $c > 0$, $\Gamma(y + c)/\Gamma(c) = c(c + 1)\cdots(c + y - 1)$ if y is an integer ≥ 1, we can write $\log L(\boldsymbol{\beta}, a)$ as

$$l(\boldsymbol{\beta}, a) = \sum_{i=1}^{n}\left(\sum_{j=0}^{y_i^*} \log(1 + aj) + y_i \log \mu_i - (y_i + a^{-1}) \log(1 + a\mu_i)\right),$$

where $y_i^* = y_1 - 1$ and $\sum_{j=0}^{y_i^*}$ is zero when $y_1^* < 0$. The first and second derivatives of l are

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} \frac{x_{ir}(y_i - \mu_i)}{1 + a\mu_i}, \qquad r = 1, \ldots, p, \tag{2.3}$$

$$\frac{\partial l}{\partial a} = \sum_{i=1}^{n} \left\{ \sum_{j=0}^{y_i^*} \left(\frac{j}{1 + aj} \right) + a^{-2} \log(1 + a\mu_i) - \frac{(y_i + a^{-1})\mu_i}{1 + a\mu_i} \right\} \tag{2.4}$$

$$\frac{-\partial^2 l}{\partial \beta_r \, \partial \beta_s} = \sum_{i=1}^{n} \frac{(1 + ay_i)\mu_i x_{ir} x_{is}}{(1 + a\mu_i)^2}, \qquad r,s = 1, \ldots, p, \tag{2.5a}$$

$$\frac{-\partial^2 l}{\partial \beta_r \, \partial a} = \sum_{i=1}^{n} \frac{\mu_i(y_i - \mu_i)x_{ir}}{(1 + a\mu_i)^2}, \qquad r = 1, \ldots, p, \tag{2.5b}$$

$$\frac{-\partial^2 l}{\partial a^2} = \sum_{i=1}^{n} \left\{ \sum_{j=0}^{y_i^*} \left(\frac{j}{1 + aj} \right)^2 + 2a^{-3} \log(1 + a\mu_i) - \frac{2a^{-2}\mu_i}{1 + a\mu_i} - \frac{(y_i + a^{-1})\mu_i^2}{(1 + a\mu_i)^2} \right\}. \tag{2.6}$$

Expectations of minus the second derivatives yield the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\beta}$, a), with entries

$$\boldsymbol{I}_{rs}(\boldsymbol{\beta}, a) = \sum_{i=1}^{n} \frac{\mu_i x_{ir} x_{is}}{1 + a\mu_i}, \qquad r,s = 1, \ldots, p \tag{2.7a}$$

$$\boldsymbol{I}_{r,p+1}(\boldsymbol{\beta}, a) = 0, \qquad r = 1, \ldots, p \tag{2.7b}$$

$$\boldsymbol{I}_{p+1,p+1}(\boldsymbol{\beta}, a) = a^{-4} \sum_{i=1}^{n} \left\{ \mathscr{E} \sum_{j=0}^{y_i^*} (a^{-1} + j)^{-2} - \frac{a\mu_i}{\mu_i + a^{-1}} \right\} = i(\boldsymbol{\beta}, a). \tag{2.8}$$

The expression for $\boldsymbol{I}_{p+1,p+1}(\boldsymbol{\beta}, a)$ is most easily obtained by rewriting l in terms of $\boldsymbol{\beta}$ and $k = a^{-1}$, calculating $\partial^2 l/\partial k^2$ and then noting that $\mathscr{E}(-\partial^2 l/\partial a^2) = a^{-4}\mathscr{E}(-\partial^2 l/\partial k^2)$. The ith term of the expectation in $\boldsymbol{I}_{p+1,p+1}(\boldsymbol{\beta}, a)$ is equal to

$$a^{-4}\left(\sum_{j=0}^{\infty} (a^{-1} + j)^{-2} \Pr(Y_i \geq j) - \frac{a\mu_i}{\mu_i + a^{-1}} \right),$$

by which it is easily calculated.

The simplest way to obtain $(\hat{\boldsymbol{\beta}}, \hat{a})$ is to maximize $l(\boldsymbol{\beta}, a)$ with respect to $\boldsymbol{\beta}$, for selected values of a. This gives estimates $\tilde{\boldsymbol{\beta}}(a)$ and the profile likelihood $l(\tilde{\boldsymbol{\beta}}(a), a)$, from which it is easy to determine $\hat{a}$; it is of course possible to have $\hat{a} = 0$. Maximization of $l(\boldsymbol{\beta}, a)$ with respect to $\boldsymbol{\beta}$ is easy via Newton-Raphson iteration or the scoring algorithm. Alternatively, generalized linear model software such as GLIM, or least-squares software, can be exploited (see McCullagh and Nelder 1983, p. 170, and Stirling 1984, respectively).

Assuming $a > 0$ and mild conditions on the $\boldsymbol{x}_i$'s to ensure that $n^{-1}\boldsymbol{I}(\boldsymbol{\beta}, a)$ approaches a positive definite limit as $n \to \infty$, we can for large n obtain tests or confidence intervals by treating $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{a} - a)$ as normally distributed with mean $\mathbf{o}$ and covariance matrix

$$n\boldsymbol{I}(\hat{\boldsymbol{\beta}}, \hat{a})^{-1} = n\begin{bmatrix} \boldsymbol{I}_1(\hat{\boldsymbol{\beta}}, \hat{a})^{-1} & \boldsymbol{O} \\ \boldsymbol{O}^{\mathrm{T}} & i(\hat{\boldsymbol{\beta}}, \hat{a})^{-1} \end{bmatrix}, \tag{2.9}$$

where $\boldsymbol{I}_1(\boldsymbol{\beta}, a)$ and $i(\boldsymbol{\beta}, a)$ are given by (2.7) and (2.8). Observed information-matrix entries given by (2.5) and (2.6) evaluated at $(\hat{\boldsymbol{\beta}}, \hat{a})$ can also be used instead of $\boldsymbol{I}_1$ and i in (2.9). A preferable alternative when n is not really large is to use likelihood-ratio statistics, assumed to be distributed as chi-squareds. It is a substantial convenience that $\hat{\boldsymbol{\beta}}$ and $\hat{a}$ are asymptotically independent; some ramifications of this appear below. Asymptotic approximations for estimates' distributions improve as both the μ_i's and n increase. Section 4 discusses the adequacy of these approximations.

To test that $a = 0$ (i.e. that a Poisson model is adequate), slight modifications of standard large-sample theory apply: this is discussed in Section 5.

MOMENT ESTIMATION OF *A*

When a is known, the NB (μ_i, a) distribution is a generalized linear model, and furthermore, the maximum-likelihood equations $\partial l/\partial\beta_r = 0$ $(r = 1, \ldots, p)$ are both quasilikelihood and weighted least-squares equations (McCullagh and Nelder 1983, Breslow 1984, Stirling 1984). Quasilikelihood or weighted least squares is indeed often suggested for dealing with extra-Poisson variation, but to do so properly one should have a way of estimating the variance or dispersion parameter a. The most common approach is to use moment estimation: Breslow (1984) suggests that, given estimates $\tilde{\mu}_i$, we estimate a by solving the equation

$$\sum_{i=1}^{n} \frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\mu}_i(1 + a\tilde{\mu}_i)} = n - p. \tag{2.10}$$

[There is either one or no solutions $a > 0$ to (2.10); if there is none, take $\tilde{a} = 0$.] Breslow recommends first fitting the Poisson model ($a = 0$) to obtain initial μ_i's, then solving (2.10), for $\tilde{a}$. If $\tilde{a} > 0$, this value is used to obtain a new $\tilde{\boldsymbol{\beta}}$ from $\partial l/\partial\beta_r = 0$ with $a = \tilde{a}$ $(r = 1, \ldots, p)$; see (2.3). This process can be iterated, until convergence if desired. Use of a deviance statistic instead of the Pearson statistic on the left-hand side of (2.10) yields a similar procedure (McCullagh and Nelder 1983).

This is a sensible procedure whose applicability is broader than just negative binomial regression. Earlier, practice was to ignore sampling variability in $\tilde{a}$, but Moore (1986) and Section 3 below provide distributional results which overcome this, and also enable us to examine the asymptotic efficiency of weighted-least-squares–moment estimation under the negative-binomial and other models. We turn to this now.

3. EFFICIENCY AND ROBUSTNESS QUESTIONS

Moment estimation for a is likely to be somewhat more robust than maximum likelihood, but is less efficient when the negative-binomial model is correct; we now examine this. I first determine the asymptotic distribution for the estimator $(\tilde{\boldsymbol{\beta}}, \tilde{a})$ obtained by solving the equations

$$\frac{\partial l(\boldsymbol{\beta}, a)}{\partial\beta_r} = 0 \qquad r = 1, \ldots, p, \tag{3.1}$$

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mu_i(1 + a\mu_i)} - (n - p) = 0. \tag{3.2}$$

These are the estimates obtained by using the approach of Breslow (1984) and others, wherein quasilikelihood or weighted least-squares is used to estimate $\boldsymbol{\beta}$ for given a, and moment estimation is used for a. The results concerning the distribution of $\tilde{a}$ have also recently been obtained by Moore (1986), in a more general setting where μ_i is not necessarily of log-linear form; those for $\tilde{\boldsymbol{\beta}}$ are already known in connection with quasilikelihood (e.g. McCullagh 1983; McCullagh and Nelder 1983, Appendix C) or weighted least-squares estimation. I give both results, since to get results for $\tilde{a}$ it's necessary to consider $\tilde{\boldsymbol{\beta}}$ simultaneously.

Using results of Inagaki (1973), it can be shown that under the same conditions as for the maximum-likelihood asymptotics, $\tilde{\boldsymbol{\beta}}$ and $\tilde{a}$ are asymptotically independent and normally distributed, with

$$\operatorname{asvar}\{\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta})\} = \boldsymbol{I}_1^*(\beta, a)^{-1} \tag{3.3}$$

$$\operatorname{asvar}\{\sqrt{n}(\tilde{a} - a)\} = b_{p+1}^{-2}\{C_{p+1} - \boldsymbol{b}'\boldsymbol{I}_1^*(\boldsymbol{\beta}, a)^{-1}\boldsymbol{b}\}, \tag{3.4}$$

where $\boldsymbol{I}_1^*(\beta, a) = \lim_{n\to\infty} (1/n)\boldsymbol{I}_1(\beta, a)$ [see (2.9)] and $\boldsymbol{b} = (b_1, \ldots, b_p)$, with

$$b_r = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1 + 2a\mu_i}{1 + a\mu_i}\right) x_{ir}, \qquad r = 1, \ldots, p,$$

$$b_{p+1} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\mu_i}{1 + a\mu_i},$$

$$C_{p+1} = 2 + 6a + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\mu_i(1 + a\mu_i)}.$$

The derivation of these results is outlined in Appendix A. In the case where there are no explanatory variables (i.e. $p = 1, x_{i1} \equiv 1$) the expression for $\operatorname{asvar}\{\sqrt{n}(\tilde{a} - a)\}$ is $2(1 + a)(1 + a\mu)^2/\mu^2$, which was given by Anscombe (1950).

The results above indicate that if we proceed as described by Breslow (1984), then the estimator $\tilde{\boldsymbol{\beta}}$ obtained is asymptotically equivalent to the maximum-likelihood estimator $\hat{\boldsymbol{\beta}}$, and its asymptotic covariance matrix is consistently estimated by $\boldsymbol{I}_1(\tilde{\boldsymbol{\beta}}, \tilde{a})^{-1}$. There is, however, some loss of efficiency in the estimation of a by $\tilde{a}$, compared to maximum likelihood. The asymptotic relative efficiency of $\tilde{a}$ to $\hat{a}$ is given by

$$\mathrm{RE} = \frac{\operatorname{asvar}\{\sqrt{n}(\hat{a} - a)\}}{\operatorname{asvar}\{\sqrt{n}(\tilde{a} - a)\}} = \frac{i^*(\boldsymbol{\beta}, a)^{-1}}{b_{p+1}^{-2}\{C_{p+1} - \boldsymbol{b}'\boldsymbol{I}_1^*(\boldsymbol{\beta}, a)^{-1}\boldsymbol{b}\}}, \tag{3.5}$$

where $i^*(\boldsymbol{\beta}, a) = \lim_{n\to\infty} (1/n)i(\boldsymbol{\beta}, a)$ [see (2.8)].

Anscombe (1950) studied the efficiency of $\tilde{a}$ relative to $\hat{a}$ in the case where there are no covariates, that is, when $\mu_i = \mu (i = 1, \ldots, n)$. He found that for any μ, RE of (3.5) approaches 1 as $a \to 0$, and that in most realistic situations, RE is over 0.90. For cases involving covariates this is no longer true: RE does not approach one as $a \to 0$, and lower efficiencies than 0.90 are encountered in many situations. In general, the efficiency tends to be lower in cases where the μ_i's are not too large and where they vary a good deal (i.e. there is a strong regression effect). Figures 1 and 2 show curves of RE vs. a for several situations, as follows: Figure 1 shows results for three cases: (a) $p = 1, \mu = 10$, (b) $p = 1, \mu = 40$, (c) $p = 2$, $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, with $\exp(\beta_0) = 10$, $\beta_1 = 1$ and 0.20 of the x_i's each of $-1, -0.5, 0, 0.5, 1$. Figure 2 shows results for $p = 2$, $\mu_i = \exp(\beta_0 + \beta_1 x_i)$ and one third of the x_i's each of $-1, 0, 1$: (a) $\exp(\beta_0) = 10$, $\beta_1 = 1$, (b) $\exp(\beta_0) = 10$, $\beta_1 = 0.5$, (c) $\exp(\beta_0) = 50$, $\beta_1 = 0.5$.

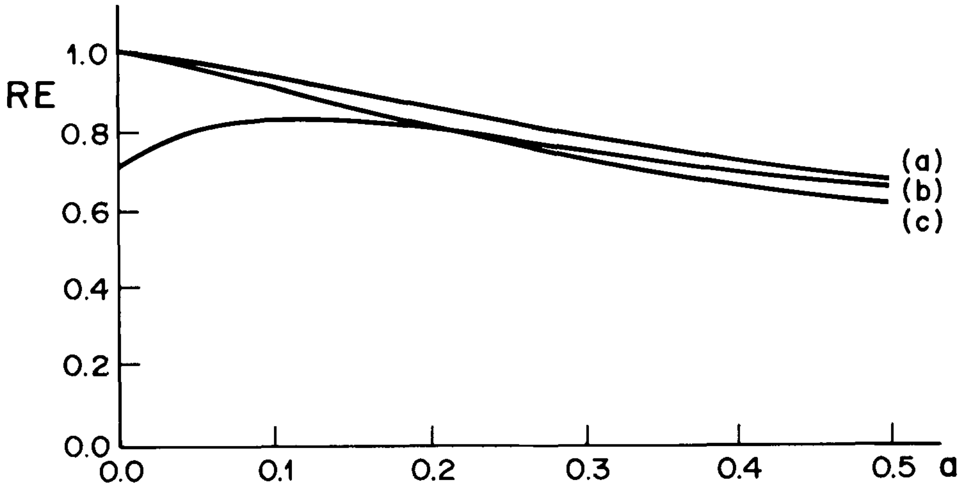

FIGURE 1: Relative efficiency of $\tilde{a}(I)$.

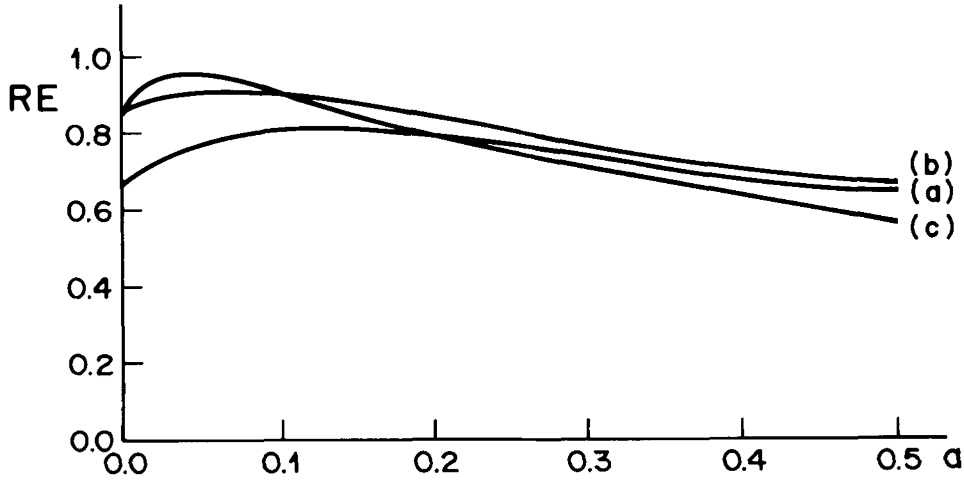

FIGURE 2: Relative efficiency of $\tilde{a}(II)$.

I turn now to the question of robustness: how do the negative-binomial maximum-likelihood or weighted least-squares and moment methods stand up when the negative binomial model is wrong? To study this I consider situations where the regression specification $\mu_i = \mathscr{E}(Y_i \mid \boldsymbol{x}_i) = T_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ is still correct, but the distribution of Y_i given $\boldsymbol{x}_i$ is not negative-binomial.

Suppose $\mathscr{V}\!\mathit{ar}(Y_i \mid \boldsymbol{x}_i) = \sigma_i^2$, which may of course depend on $\boldsymbol{x}_i$. Then, using results of Cox (1961), Inagaki (1973), and White (1982), we find (see Appendix B) under mild conditions on the distributions of the Y_i's that as $n \to \infty$,

(1) $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are both consistent estimators of $\boldsymbol{\beta}$;

(2) $\hat{a} \xrightarrow{P} a_1^*$ and $\tilde{a} \xrightarrow{P} a_2^*$, where a_1^* and a_2^* are nonnegative constants;

(3) we have

$$\mathscr{asvar}\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}=I_1^{-1}B_1I_1^{-1}, \tag{3.6}$$

where

$$(I_1)_{r,s}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\mu_i x_{ir}x_{is}}{1+a_1^*\mu_i},\qquad r,s+1,\ldots,p,$$

$$(B_1)_{r,s}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_i^2 x_{ir}x_{is}}{(1+a_1^*\mu_i)^2},\qquad r,s+1,\ldots,p;$$

(4) we have

$$\mathscr{asvar}\{\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\}=I_2^{-1}B_2I_2^{-1}, \tag{3.7}$$

where I_2 and B_2 differ from I_1 and B_1 only in that a_2^* replaces a_1^*.

If we use the negative-binomial model for estimation, we thus will get consistent estimates of $\boldsymbol{\beta}$, but will have the covariance matrix wrong in general. (The estimates are of course also less efficient than maximum-likelihood estimates based on the correct model, but that is not our present concern.) The asymptotic covariance matrix we *use* for $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is $I_1^*(\hat{\boldsymbol{\beta}},\hat{a})^{-1}$, which converges in probability to $I_1(\boldsymbol{\beta},a_1^*)^{-1}$. The *correct* asymptotic covariance matrix for $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is $I_1^{-1}B_1I_1^{-1}$, as given by (3.6). Since a_1^* and a_2^* are complicated functions of the negative-binomial estimating equations and the true distribution for Y, it is not obvious how far wrong the variance estimates for $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$ will be, but this can be evaluated numerically in any given situation. I provide an illustration below. I note also that if the form of the variance function is correct, so that

$$\mathscr{Var}(Y_i\,|\,\boldsymbol{x}_i)=\mu_i+a\mu_i^2 \tag{3.8}$$

for some $a>0$, then it can be shown that $a_2^*=a$, i.e., that moment estimation gives consistent estimation of a and hence of σ_i^2. In this case $I_1=I_2=B_2$ in (3.7) and so $\mathscr{asvar}\{\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\}=I_1(\boldsymbol{\beta},a)^{-1}$. Thus, weighted least-squares–moment estimation yields an asymptotically correct covariance matrix for $\tilde{\boldsymbol{\beta}}$. This appears to be an advantage of the moment estimator over the maximum-likelihood estimator for a; however, numerical calculations like those below suggest that the asymptotic covariance matrix obtained via $\hat{a}$ is usually only slightly off. Moore (1985) also suggests through some simulation work that the moment procedure is relatively robust to misspecification of the variance function, as far as estimation of $\boldsymbol{\beta}$ in finite samples is concerned.

To study the bias in estimation of $\mathscr{asvar}\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}$ when the NB model is not correct, but (3.8) is, we note, following Cox (1961) or White (1982) that a_1^* of (3.6) is the solution to

$$\mathscr{E}\left\{\frac{\partial l(\boldsymbol{\beta},a_1^*)}{\partial a_1^*}\right\}=0, \tag{3.9}$$

where the expectation is with respect to the true distribution of Y. Furthermore, as indicated above, the correct $\mathscr{asvar}$ $\{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\}$ is $V_{\mathrm{C}}=I_1(\boldsymbol{\beta},a_1^*)^{-1}B_1(\boldsymbol{\beta},a_1^*)I_1(\boldsymbol{\beta},a_1^*)^{-1}$, whereas the one actually used converges in probability to $V_{\mathrm{NB}}=I_1(\boldsymbol{\beta},a_1^*)^{-1}$. It is

TABLE 1: Comparison of V_{NB} and V_C in a case with no covariates.

μ	a	a_1^*	V_{NB}/V_C
5	.5	.432	0.903
5	.1	.098	0.994
20	.1	.096	0.974
20	.01	.0100	1.000
50	.1	.0951	0.959
50	.01	.0100	0.999

TABLE 2: Comparison of V_{NB} and V_C in a case with one covariate.

e^{β_0}	a	a_1^*	$V_{NB}(1,1)/V_C(1,1)$	$V_{NB}(2,2)/V_C(2,2)$
1	.5	.454	0.974	0.970
1	.1	.0989	0.999	0.999
1	.01	.0100	1.000	1.000
10	.5	.413	0.862	0.863
10	.1	.0964	0.984	0.983
10	.01	.0100	1.000	1.000
20	.5	.404	0.831	0.833
20	.1	.0957	0.974	0.974
20	.01	.0100	1.000	1.000

possible to obtain a_1^* of (3.9) numerically in any given situation and thence to compare V_{NB} with V_{C}. I do this below for some situations where the distribution of Y is a Poisson–inverse Gaussian mixture (e.g. Sichel 1982); this distribution has a longer tail than the negative binomial.

Table 1 shows values of a_1^* and ratios $V_{\mathrm{NB}}/V_{\mathrm{C}}$ for some situations with no covariates (i.e. $p = 1$, all $\mu_i = \mu$). Table 2 shows results for situations with $p = 2$, $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, where $\beta_1 = 1$ and one-third of the x_i's are each of -1, 0, 1. It is seen from the two tables that the NB maximum-likelihood procedure underestimates the variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ slightly in large samples. In practical situations quite small values of a tend to occur with larger μ_i's, so that in fact the underestimation of asymptotic variances is for practical purposes rather inconsequential.

When $\mathscr{Var}(Y_i) = \sigma_i^2$ is of a quite different form than (3.8), the asymptotic covariance matrix used under the NB model can be substantially off. In practice, however, diagnostic checks are used to assess σ_i^2 relative to μ_i, and evidence of gross departures from (3.8) would steer us away from the NB model. The results of this section indicate that, when (3.8) is reasonable, either the full NB maximum-likelihood approach of Section 2 or the combined quasilikelihood–method-of-moments approach can be trusted for inferences about $\boldsymbol{\beta}$. In fact, the covariance-matrix estimates based on (3.7) are valid more generally, and apply to quasilikelihood weighted least-squares estimates for any model for which the μ_i specification and (3.8) are correct.

4. ADEQUACY OF LARGE-SAMPLE APPROXIMATIONS

For tests and confidence intervals based on standard large-sample methods, it is of interest to know about the adequacy in finite samples of the various distributional approximations used. In thesis work currently in progress, C. Dean has examined by simulation the accuracy of asymptotic normal approximations to the distributions of $(\hat{\boldsymbol{\beta}}, \hat{a})$ and $(\tilde{\boldsymbol{\beta}}, \tilde{a})$, as given by (2.9), (3.3), and (3.4), and chi-squared approximations for likelihood-ratio statistics. A very brief indication of results, and some rough guidelines, are given here.

Regarding inferences about $\boldsymbol{\beta}$, it appears that likelihood-ratio statistics, with their distributions approximated by χ^2 distributions in the usual way, are satisfactory except possibly in very small samples. If the asymptotic standard normal distributions of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ or $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are instead used, then it appears that the normal approximation for $\tilde{\beta}$ is somewhat better in smaller samples than that for $\hat{\beta}$, and is generally quite good for samples as small as 25 or 30. Even for sample sizes as big as 80 or 90, the distribution of $\sqrt{n}I(\hat{\boldsymbol{\beta}}, \hat{a})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ appears to give slightly too many small negative values, relative to a standard normal distribution. For inferences about the regression coefficients, therefore, the use of likelihood-ratio statistics is preferable for smaller samples, with the use of normal approximations for $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ also being reasonable. None of the methods, however, are so inaccurate as to affect conclusions in a major way except possibly in very small samples.

For obtaining tests or confidence intervals for a, the likelihood ratio statistic is preferable. When a is close to zero, even it may not be well approximated by its asymptotic $\chi^2_{(1)}$ distribution unless n or the values of $a\mu_i$ are fairly large. When a is quite close to zero, the normal approximations to the distributions of $\sqrt{n}(\hat{a} - a)$ and $\sqrt{n}(\tilde{a} - a)$ are also poor unless the $a\mu_i$'s or n is large. The actual distributions of these quantities have shorter left tails, and longer right tails, than the approximating normal distributions. For testing the hypothesis that $a = 0$, the limiting distributions of $\sqrt{n}\hat{a}$, $\sqrt{n}\tilde{a}$ or the likelihood-ratio statistic have a probability mass of 0.5 at zero: see Section 5. In this case, unless n or the μ_i's are quite large, any of the three statistics used with their asymptotic distributions will give significance levels that are too big. This is in large part due to the fact that the proportion of samples in which $\hat{a}$ or $\tilde{a}$ equals zero appears to be well over 0.5 unless n or the μ_i's are fairly large.

Table 3 presents a few simulation results for one of the scenarios represented in Table 2: that with one covariate, for which $p = 2$, $e^{\beta_0} = 10$, and one-third of the μ_i's are equal to each of 3.7, 10, and 27. The table shows, for selected values of a and n, the proportion of times in 500 simulations that

$$Z(\hat{\beta}_j) = \sqrt{n}(\hat{\beta}_j - \beta_j)/\widehat{asvar}(\hat{\beta}_j)^{\frac{1}{2}} \text{ and } Z(\tilde{\beta}_j), \qquad j = 0, 1,$$

fell outside the 0.05 and 0.01 standard normal limits ±1.96 and ±2.58, respectively. Similar figures are shown for

$$Z(\hat{a}) = \sqrt{n}(\hat{a} - a)/\widehat{asvar}(\hat{a})^{\frac{1}{2}} \quad \text{and} \quad Z(\tilde{a}).$$

Asymptotic variances were based on (2.9) and on (3.3), (3.4) with $\tilde{\boldsymbol{\beta}}$, $\tilde{a}$ used to estimate $\boldsymbol{\beta}$, a. Results are also shown for the likelihood-ratio statistic $\Lambda(\hat{a}) = 2l(\hat{\boldsymbol{\beta}}, \hat{a}) - 2l(\hat{\boldsymbol{\beta}}(a), a)$, the figure given being the proportion of times that it exceeded the upper 0.05 and 0.01 percentage points of $\chi^2_{(1)}$. For the cases with $a = 0$, the modified (half-normal and half-$\chi^2_{(1)}$) limiting distributions discussed in Section 5 provided the nominal percentage points. The table displays the general qualitative features mentioned above, and provides some feel for the adequacy of the various approximations. I remark that it is rather uncommon to see an a-value as large as 0.5

TABLE 3: Proportion of time in 500 samples that statistics exceeded nominal normal or chi-squared percentage points.

a	n	%Pt	$Z(\hat{\beta}_0)$	$Z(\tilde{\beta}_0)$	$Z(\hat{\beta}_1)$	$Z(\tilde{\beta}_1)$	$Z(\hat{a})$	$Z(\tilde{a})$	$\Lambda(\hat{a})$
0.5	30	.05	.096	.088	.094	.084	.126	.094	.068
		.01	.028	.024	.030	.028	.076	.046	.014
	90	.05	.064	.056	.068	.062	.076	.050	.052
		.01	.016	.018	.024	.020	.022	.012	.008
0.1	30	.05	.112	.064	.084	.046	.172	.102	.062
		.01	.046	.018	.024	.010	.118	.058	.010
	90	.05	.068	.054	.096	.058	.122	.096	.076
		.01	.024	.012	.022	.006	.054	.034	.010
0.01	30	.05	.088	.050	.096	.044	.000	.002	.008
		.01	.026	.008	.022	.004	.000	.000	.002
	90	.05	.120	.054	.108	.052	.000	.006	.024
		.01	.036	.018	.044	.014	.000	.000	.000
0	30	.05	.058	.054	.048	.040	.042	.006	.020
		.01	.014	.014	.006	.006	.012	.000	.000
	90	.05	.040	.038	.046	.042	.028	.020	.018
		.01	.006	.006	.008	.008	.008	.002	.008

when the μ_i's are in the range represented here, and that with proportions in the table based on 500 samples, standard errors associated with them are about 0.0044 and 0.01 for %Pt. = 0.05 and 0.01, respectively.

The large-sample approximations are reasonably satisfactory; in most situations inferences about $\boldsymbol{\beta}$ will be of main interest, whereas interval estimates for a will not be as major a concern. Approximate confidence intervals for a which are based on the asymptotic $\chi^2_{(1)}$ distribution of $\Lambda(\hat{a})$ should, however, be reliable except when a is close to zero. As a practical guideline, if the confidence interval includes the value $a = 0$, then we should expect that the right-hand end of the confidence interval is larger than it should be. Finally, tests of the Poisson hypothesis $a = 0$ will often be of interest; we discuss this in the next section.

5. TESTING A POISSON ASSUMPTION

Poisson regression models are very useful, and it is desirable to have a test of the Poisson assumption. One way to do this is to test that $a = 0$ within the negative-binomial model. Using results of Moran (1971) which apply to situations where the parameter is on the boundary of the parameter space, we have that when $a = 0$, the distribution of $Z = \sqrt{n}\hat{a}i(\hat{\boldsymbol{\beta}}, 0)^{\frac{1}{2}}$ [see (2.9)] asymptotically has a (half) normal distribution for $Z > 0$ and a probability mass of $\frac{1}{2}$ at 0. Here, $\hat{\boldsymbol{\beta}}$ is the m.l.e. of $\boldsymbol{\beta}$ obtained under $a = 0$ (i.e. the Poisson model). Alternatively, one can use analogous results of Chernoff (1954), which show that the likelihood-ratio statistic for testing $a = 0$ is, under the null hypothesis, asymptotically like a random variable which has a probability mass of $\frac{1}{2}$ at 0 and a $\frac{1}{2}\chi^2_{(1)}$ distribution above 0.

Investigations described in Section 4 indicate that unless n or the μ_i's are very large, the limiting asymptotic distributions just mentioned substantially overestimate the significance levels associated with these tests. Another way to test the Poisson model is via a partial-score [$\equiv C(\alpha)$ or Lagrange-multiplier] test of $a = 0$. For the situation with no covariates it is well known that this procedure leads to the Fisher dispersion test, based on the statistic

$$D = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\bar{Y}}. \tag{5.1}$$

Collings and Margolin (1985) extend this to the case of a one-way layout, and to the general case of a single regressor variable. Furthermore, as Collings and Margolin and other authors note, this test arises more generally as a partial-score test for the Poisson model against rather arbitrary mixed Poisson alternatives, of which the negative-binomial is a special case. Dean and Lawless (1987) show that for general regression situations of the type discussed in this paper, where $\mathscr{E}(Y_i \mid \boldsymbol{x}_i) = \mu_i = \mu_i(\boldsymbol{x}_i, \boldsymbol{\beta})$, the partial score leads to a test based on the standardized dispersion statistic

$$S = \frac{\sum_{i=1}^{n} \{(Y_i - \hat{\mu}_i)^2 - \bar{Y}\}}{\left(2\sum_{i=1}^{n} \hat{\mu}_i^2\right)^{\frac{1}{2}}}. \tag{5.2}$$

In addition, under the hypothesis that the Y_i's are independent Poisson (μ_i) random variables, S is asymptotically standard normal. Large positive values of S indicate overdispersion relative to a Poisson distribution; as noted in Section 7, large negative values indicate underdispersion.

On a note of caution, S appears to approach normality rather slowly, and the normal approximation is not recommended unless n is at least 50 or so; work is under way on better approximations. We remark also that S is designed to test for extra-Poisson variation, whereas the familiar Pearson statistic $\Sigma\ (Y_i - \hat{\mu}_i)^2/\hat{\mu}_i$ or deviance statistic $2\Sigma\ Y_i \log (Y_i/\hat{\mu}_i)$ (cf. McCullagh and Nelder 1983, pp. 130-131) are designed to test for inadequacy of the regression specification $\mu_i = \mu_i(\boldsymbol{x}_i, \boldsymbol{\beta})$ within the Poisson framework. It is clear, however, that these and the statistic S are to some extent interchangeable, and a study of their respective abilities to recognize different types of departures from a Poisson regression model is under way. As always, the examination of residuals and influence statistics (e.g. Frome 1983) should supplement formal tests.

6. EXAMPLES

Two sets of data will be examined briefly, to illustrate some of the points discussed earlier.

EXAMPLE 1 (Ship damage incidents). The responses Y_i are the numbers of damage incidents for 35 individual ships over various five year periods, and the exposures T_i ($i = 1, \ldots, 35$) are the total months in service for each ship. The y_i's ranged in size from 0 to 58. There are three qualitative factors: ship type (A, B, C, D, or E), year of construction (1960-64, 1965-69, 1970-74, or 1975-79), and period of operation (1960-74 or 1975-79). For the results reported below, binary indicator covariates were used to represent main effects (four for ship type, three for year of construction, one for period of operation, and one for an intercept), and a log-linear specification $\mathscr{E}(Y_i \mid \boldsymbol{x}_i) = \mu_i = T_i\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ was employed.

These data were analyzed in some detail by McCullagh and Nelder, so I report only on a few points of interest. They fitted a log-linear model with the same μ_i as above by using quasilikelihood with a variance function $\mathscr{V}\!ar(Y_i|\boldsymbol{x}_i) = \sigma^2\mu_i$. Although this is different than the NB-model variance function, both approaches fit the data well; they give the same estimates of regression coefficients, give similar results for inferences about regression effects, and give similar standardized residuals, using the definitions $r_i = (Y_i - \hat{\mu}_i)/(\hat{\mu}_i + \hat{a}\hat{\mu}_i^2)^{\frac{1}{2}}$ for the NB model and $r_i = (y_i - \hat{\mu}_i)/(\hat{\sigma}^2\hat{\mu}_i)^{\frac{1}{2}}$ for the McCullagh-Nelder (MN) model. Briefly, both approaches indicate that main effects are significant, and there is some inconclusive evidence for an interaction of ship type by year of operation.

One point of note concerns the estimates of the variance or dispersion parameters. With the main-effects model, which has $p = 9$ covariates, the NB maximum likelihood gives $\hat{a} = 0$; on the other hand, the method-of-moments estimate is $\tilde{a} = 0.149$, with standard error 0.113. Table 4 shows under (i) and (ii), estimates of the regression coefficients and their standard errors, obtained under the two approaches. The estimates and standard errors obtained by McCullagh and Nelder are shown under (iii).

We remark first that although there is not strong evidence of extra-Poisson variation under either the NB or the MN model (i.e. that $a > 0$ or $\sigma^2 > 1$, respectively), which approach one uses has a fairly strong effect on the standard errors of the estimated regression coefficients. This is clearly seen in Table 4; the two estimates $\hat{a} = 0$, $\tilde{a} = 0.149$ yield quite different standard errors, and the MN standard errors are somewhere in between. In fact, effects under (ii) for ship type or service period do not show up as especially significant, whereas they do under (i) and to a little lesser extent, under the MN analysis, (iii). When $n - p$ is not large (here it is 26), the dispersion parameter may not be estimated very precisely, and different methods can lead to rather different estimates of the standard error.

We remark that there is one fairly large residual, which naturally shows up as more extreme (with a value over 3.5) under analysis (i) than under (ii) or (iii) [e.g., under

TABLE 4: Estimates in the ship-damage example.

	Estimates (standard errors)		
Parameter	(i) $\hat{a} = 0$ (ML)	(ii) $\tilde{a} = 0.149$ (moments)	(iii) $\hat{\sigma}^2 = 1.69$ (MN)
Intercept	-6.41	-6.45	-6.41
Ship type:			
A	0	0	0
B	-.54(.18)	-.49(.30)	-.54(.23)
C	-.69(.33)	-.56(.42)	-.69(.43)
D	-.08(.29)	-.11(.40)	-.08(.38)
E	.33(.24)	.46(.36)	.33(.31)
Year of construction:			
60-64	0	0	0
65-69	.70(.15)	.72(.35)	.70(.19)
70-74	.82(.17)	.91(.34)	.82(.22)
75-79	.45(.23)	.46(.41)	.45(.30)
Service period:			
60-74	0	0	0
75-79	.38(.12)	.34(.24)	.38(.15)

TABLE 5: Number of revertant colonies of salmonella (Y_i).

	Y_i					
Obs.	$x_i = 0^a$	10	33	100	333	1000
1	15	16	16	27	33	20
2	21	18	26	41	38	27
3	29	21	33	60	41	42

[a]Dose of quinoline (μg/plate).

(iii) it is 2.9]. This residual is decreased a good deal if an interaction of ship type by year of construction is included, in which case the results of the three analyses also come a little more into line.

EXAMPLE 2 (Ames salmonella assay). Margolin et al. (1981) present data, shown in Table 5, from an Ames salmonella reverse mutagenicity assay; the data were also analyzed by Breslow (1984). The response variable Y is the number of revertant colonies observed on a plate, and covariates are based on x, the dose level of quinoline on the plate. In the assay in question, three observations were taken at each of six dose levels.

I will work with an approximation to Margolin et al.'s "single hit" model which is considered by Breslow; this has

$$\mathscr{E}(Y_i \mid x_i) = \mu_i = \exp\{\beta_0 + \beta_1 x_i + \beta_2 \log(x_i + 10)\}.$$

Tests of $H : \beta_2 = 0$ are of special interest, with $\beta_2 > 0$ representing a mutagenic effect. Fitting the model $Y_i \sim \mathrm{NB}(\mu_i, a)$ by maximum likelihood and by weighted least-squares-method of moments, respectively, yields the estimates (standard errors)

$$\hat{a} = 0.0488\ (0.0275), \quad \hat{\beta}_0 = 2.198\ (0.321), \quad \hat{\beta}_1 = -0.000980\ (0.000381), \quad \hat{\beta}_2 = 0.313\ (0.0868);$$

$$\tilde{a} = 0.0718\ (0.0303), \quad \tilde{\beta}_0 = 2.203\ (0.359), \quad \tilde{\beta}_1 = -0.000974\ (0.000430), \quad \tilde{\beta}_2 = 0.311\ (0.0974).$$

The latter results agree with Breslow (1984). To test $H : a = 0$ we use the likelihood-ratio statistic $R = 2l(\hat{\boldsymbol{\beta}}, \hat{a}) - 2l(\hat{\boldsymbol{\beta}}(0), 0)$, which asymptotically under H has a probability mass of $\frac{1}{2}$ at 0 and a $\frac{1}{2}\chi^2_{(1)}$ distribution for $R > 0$. From Section 4 we note that although this asymptotic approximation may not be highly accurate for the situation at hand, the tendency is for the significance level to be overestimated. The observed value of R here is 10.4, which provides strong evidence against the Poisson model. The actual estimates of β_0, β_1, β_2 do not change drastically if the Poisson model ($a = 0$) is used, but their standard errors do. For example, the value of $\hat{\beta}_2/\mathrm{se}(\hat{\beta}_2)$ changes from $0.313/0.0868 = 3.6$ under the NB model with $\hat{a} = 0.0488$ to $0.320/0.057 = 5.6$ under the Poisson model. Consequently, the Poisson model overstates the significance of the mutagenic effect.

7. CONCLUDING REMARKS

It is important to have methods of dealing with extra-Poisson variation in regression situations. Mixed Poisson models for which the mean-variance relationship is of the form (2.2) provide one way of doing this, and the negative-binomial model discussed in this paper is particularly convenient. It allows the use of standard maximum-likelihood methods, and has good properties. The alternative use of quasilikelihood or weighted least-squares, along with moment estimation of the dispersion parameter a, is also recommended, although it does not allow the easy use of likelihood-ratio statistics that maximum likelihood does. Several authors have investigated maximum likelihood for other mixed Poisson models, for example log-normal (Hinde 1982) and log-student-t (Gaver and O'Muircheartaigh 1987) mixtures. These are considerably more difficult to handle than the negative-binomial model.

Sometimes the variance-mean relationship $\sigma_i^2 = \mathscr{Var}(Y_i \mid \boldsymbol{x}) = \mu_i + a\mu_i^2$ implied by a mixed Poisson model may clearly be inadequate. Indeed, many authors in fitting count data have noted that relationships such as $\sigma_i^2 = a\mu_i^b$ (e.g., Armitage 1957; Finney 1976) and $\sigma_i^2 = a\mu_i$ (e.g., McCullagh and Nelder 1983) often appear plausible. The evidence for such relationships is usually empirical, and in many cases a number of them will provide similar fits to the data. Usually when another variance function than (3.8) is employed, a semiparametric approach to estimation is used, which does not involve the specification of the full distribution for Y given $\boldsymbol{x}$. Well-known approaches of this kind are quasilikelihood estimation (cf. McCullagh and Nelder 1983, Ch. 8) or weighted least squares, based on solving equations of the form

$$\sum_{i=1}^{n} \tilde{w}_i(y_i - \mu_i)\frac{\partial \mu_i}{\partial \beta_r} = 0, \qquad r = 1, \ldots, p, \tag{6.1}$$

where the $\tilde{w}_i$'s are weights, preferably equal to consistent estimates of σ_i^{-2}. Some of these approaches do not naturally allow for the estimation of nuisance parameters in σ_i^2, but approaches such as extended quasilikelihood (Nelder and Pregibon 1987), pseudo-Gaussian estimation (Whittle 1961; Crowder 1985, the double-exponential family models of Efron (1986), and quadratic estimating functions (Crowder 1987; Godambe and Thompson 1987) can handle this. Some such methods have robustness advantages over maximum likelihood based on a fully specified model, but can be less efficient, and may be harder to adapt to specific tasks, such as prediction or assessing departures from a base model. Burridge (1986) and Firth (1987) examine some aspects of the gap between semiparametric and fully parametric estimation. Sometimes models with components of dispersion are required. Quasilikelihood or weighted least-squares methods can be adapted to this (e.g., McCullagh and Nelder 1983, p. 225; Crowder 1985), but fully parametric approaches based on NB or other models are also attractive.

Finally, a referee raised a question concerning the models (2.1) when $a < 0$. It is well known (e.g. Olkin *et al.* 1981) that by allowing $a < 0$ one obtains binomial or generalized binomial models, which are underdispersed relative to a Poisson distribution. Although the focus of this paper was overdispersion, or extra Poisson variation, I note that some of the discussion in this paper can be extended to the case $a < 0$. For example, the statistic (5.2) can detect underdispersion, large negative values of S indicating this.

APPENDIX A. THE ASYMPTOTIC DISTRIBUTION OF $(\tilde{\boldsymbol{\beta}}, \tilde{\alpha})$ UNDER THE NB MODEL

We suppose that $Y_i \sim \mathrm{NB}(\mu_i, a)$ with $\mu_i = T_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, and consider the estimating equations

$$u_r(\boldsymbol{\beta}, a) = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ir}}{1 + a\mu_i} = 0, \qquad r = 1, \ldots, p, \tag{A1}$$

$$u_{p+1}(\boldsymbol{\beta}, a) = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{\mu_i(1 + a\mu_i)} - n = 0, \tag{A2}$$

which when p is fixed are asymptotically equivalent to (2.3), (2.10) as $n \to \infty$. Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, a)$. Then, from results of Inagaki (1973), the estimator $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{a})$ obtained by solving the equations above is, under conditions similar to those for which standard maximum-likelihood asymptotics hold, consistent and asymptotically normal with covariance matrix

$$\operatorname{asvar}\{\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\} = A(\boldsymbol{\theta})^{-1}B(\boldsymbol{\theta})\{A(\boldsymbol{\theta})^{-1}\}^{\mathrm{T}}, \tag{A3}$$

where $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ are $(p + 1) \times (p + 1)$ matrices with respective entries

$$A(\boldsymbol{\theta})_{r,s} = \lim_{n\to\infty} \left\{\frac{1}{n}\mathscr{E}\left(\frac{-\partial u_r}{\partial \theta_s}\right)\right\}, \quad B(\boldsymbol{\theta})_{r,s} = \lim_{n\to\infty} \left\{\frac{1}{n}\mathscr{E}(u_r u_s)\right\}, \qquad r, s = 1, \ldots, p + 1.$$

Noting that when $Y_i \sim \mathrm{NB}\,(\mu_i,a)$ we have $\mathscr{E}\,\{Y_i - \mu_i\} = 0$, $\mathscr{E}\,\{(Y_i - \mu_i)^2\} = \mu_i(1 + a\mu_i)$, $\mathscr{E}\,\{(Y_i - \mu_i)^3\} = \mu_i(1 + a\mu_i)(1 + 2a\mu_i)$, $\mathscr{E}\,\{(Y_i - \mu_i)^4\} = \mu_i(1 + a\mu_i)(1 + 3\mu_i + 6a\mu_i + 3a\mu_i^2 + 6a^2\mu_i^2)$, it follows after some algebra that $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ are the limits, respectively, of

$$A_n(\boldsymbol{\theta}) = \begin{bmatrix} n^{-1}\boldsymbol{I}_1(\boldsymbol{\beta}, a) & \boldsymbol{O} \\ \boldsymbol{b}^{\mathrm{T}} & b_{p+1} \end{bmatrix}, \qquad B_n(\boldsymbol{\theta}) = \begin{bmatrix} n^{-1}\boldsymbol{I}_1(\boldsymbol{\beta}, a) & \boldsymbol{c} \\ \boldsymbol{c}^{\mathrm{T}} & c_{p+1} \end{bmatrix}$$

where $\boldsymbol{I}_1(\boldsymbol{\beta}, a)$ is as in (2.7) and where $\boldsymbol{b} = (b_1, \ldots, b_p)^{\mathrm{T}}$ and $\boldsymbol{c} = (c_1, \ldots, c_p)^{\mathrm{T}}$, with

$$b_r = c_r = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1 + 2a\mu_i}{1 + a\mu_i}\right) x_{ir} \qquad r = 1, \ldots, p,$$

$$b_{p+1} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mu_i}{1 + a\mu_i}\right),$$

$$c_{p+1} = 2 + 6a + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\mu_i(1 + a\mu_i)}.$$

Inverting $A(\boldsymbol{\theta})$ and using (A3), we get

$$\operatorname{asvar}\{\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}, \tilde{a} - a)\} = \begin{bmatrix} \boldsymbol{I}_1^{*}(\boldsymbol{\beta}, a)^{-1} & \boldsymbol{O} \\ \boldsymbol{O}^{\mathrm{T}} & \dfrac{1}{b_{p+1}^2}\{c_{p+1} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{I}_1^{*}.(\boldsymbol{\beta}, a)^{-1}\boldsymbol{b}\} \end{bmatrix}.$$

where $\boldsymbol{I}_1^{*}(\boldsymbol{\beta}, a) = \lim_{n\to\infty} (1/n)\, \boldsymbol{I}_1(\boldsymbol{\beta}, a)$.

APPENDIX B. EFFECTS OF MODEL MISSPECIFICATION

We consider maximum-likelihood estimation under the model $Y_i \sim \mathrm{NB}(\mu_i, a)$, when in fact the true distribution of Y_i is something else with, however $\mathscr{E}(Y_i) = \mu_i = T_i e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}$ correct. Denote $\mathscr{V}\!ar\,(Y_i) = \sigma_i^2$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, a)$, and let $l(\boldsymbol{\theta}) = l(\boldsymbol{\beta}, a)$ be the negative-binomial log likelihood. Define $(p+1)\times(p+1)$ matrices $A(\boldsymbol{\theta})$, $B(\boldsymbol{\theta})$ with entries

$$A(\boldsymbol{\theta})_{r,s} = \lim_{n\to\infty}\left\{\frac{1}{n}E\left(\frac{-\partial^2 l}{\partial\theta_r\,\partial\theta_s}\right)\right\}, \qquad B(\boldsymbol{\theta})_{r,s} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial l_i}{\partial\theta_r}\,\frac{\partial l_i}{\partial\theta_s}\right)\right\}, \tag{B1}$$

where l_i is the contribution to $\mathrm{l}(\boldsymbol{\theta})$ from the ith observation $(Y_i; \boldsymbol{x}_i)$ and where expectations here and below are taken with respect to the true distribution of the Y_i's.

The results of Cox (1961), Inagaki (1973), and White (1982) can be used to show that under mild conditions the NB maximum likelihood equations $\partial l/\partial\theta_r = 0$ $(r = 1, \ldots, p + 1)$ have solutions $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{a})$ which converge in probability as $n \to \infty$ to a vector $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, a_1^*)$, and that $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is asymptotically normal with covariance matrix

$$\mathit{ascov}\{\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} = A(\boldsymbol{\theta}^*)^{-1}B(\boldsymbol{\theta}^*)A(\boldsymbol{\theta}^*)^{-1}. \tag{B2}$$

Furthermore, it can be shown that $\boldsymbol{\beta}^* = \boldsymbol{\beta}$, that is, that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$.

Straightforward calculations similar to those in Appendix A then yield

$$\mathit{asvar}\{\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\} = I_1(\boldsymbol{\beta}, a_1^*)^{-1}B_1(\boldsymbol{\beta}, a_1^*)I_1(\boldsymbol{\beta}, a_1^*))^{-1}, \tag{B3}$$

where

$$I_1(\boldsymbol{\beta}, a_1^*)_{r,s} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\mu_i x_{ir} x_{is}}{1 + a_1^*\mu_i}\right\},$$

$$B_1(\boldsymbol{\beta}, a_1^*)_{r,s} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\sigma_i^2 x_{ir} x_{is}}{(1 + a_1^*\mu_i)^2}\right\}.$$

For the case of moment estimation of a, the estimating equations (A1), (A2) are used. In this case results are similar to those for the maximum-likelihood equations above, except that the estimating equation $\partial l/\partial\theta_{p+1} \equiv \partial l/\partial a$ is replaced by (A2). Thus $\tilde{\boldsymbol{\beta}}$, $\tilde{a}$ converge in probability to $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_2^*, a_2^*)$, and $\mathit{asvar}\ \{\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_2^*, \tilde{a} - a_2^*)\}$ is given by (B2). Furthermore, $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}$, so that $\tilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$, and $\mathit{asvar}\ \{\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$ is given by (B3), with a_2^* replacing a_1^*. Finally, if $\sigma_i^2 = \mu_i + a\mu_i^2$ for some $a > 0$, then (A2) is an unbiased estimating equation and $a_2^* = a$. Thus moment estimation yields consistent estimation of σ_i^2 in this case, $I_1(\boldsymbol{\beta}, a) = B_1(\boldsymbol{\beta}, a)$, and so the NB variance estimate $I_1(\tilde{\boldsymbol{\beta}}, \tilde{a})^{-1}$ for $\tilde{\boldsymbol{\beta}}$ is asymptotically correct.

ACKNOWLEDGMENTS

Thanks go to Charmaine Dean for computational assistance and to Jim Burridge, Martin Crowder, and Charmaine Dean for helpful discussions. This work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada, and part of it was completed while the author was a visitor at Imperial and University Colleges, London.

REFERENCES

Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37, 358–382.

Armitage, P. (1957). Studies in the variability of pock counts. *J. Hygiene Camb.*, 55, 564–581.

Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Appl. Statist.*, 33, 38–44.

Burridge, J. (1986). Mean-variance relationships, generalized linear models and GLIM. Unpublished.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.*, 25, 573–578.

Collings, B.J., and Margolin, B.H. (1985). Testing goodness of fit for the Poisson assumptions when observations are not identically distributed. *J. Amer. Statist. Assoc.*, 80, 411-418.

Cox, D.R. (1961). Tests of separate families of hypotheses. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, 105–123.

Cox, D.R. (1983). Some remarks on over-dispersion. *Biometrika*, 70, 269–274.

Crowder, M. (1985). Gaussian estimation for correlated binomial data. *J. Roy. Statist. Soc. Ser. B*, 47, 229-237.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika*, 74, 591-598.

Dean, C., and Lawless, J.F. (1987). Testing for overdispersion in Poisson regression models. Unpublished.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.*, 81, 709-721.

Engel, J. (1984). Models for response data showing extra-Poisson variation. *Statist. Neerlandica*, 38, 159-167.

Finney, D.J. Radioligand assay. *Biometrics*, 32, 721-740.

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

Frome, E.L. (1983). The analysis of rates using Poisson regression models. *Biometrics*, 39, 665-674.

Frome, E.L.; Kutner, M.H., and Beauchamp, J.J. (1973). Regression analysis of Poisson-distributed data. *J. Amer. Statist. Assoc.*, 68, 935-940.

Gaver, D.P. and O'Muircheartaigh, I.G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics*, 29, 1-15.

Godambe, V.P., and Thompson, M.E. (1987). An extension of quasi-likelihood estimation. *J. Statist. Plann. Inference*, to appear.

Haberman, S.H. (1974). *The Analysis of Frequency Data*. Univ. of Chicago Press, Chicago.

Hinde, J. (1982). Compound Poisson regression models. *GLIM 82: Proc. Internat. Conf. Generalized Linear Models* (R. Gilchrist, *ed.*), Springer, Berlin, 109-121.

Holford, T.R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39, 311-324.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. *Ann. Inst. Statist. Math.*, 265, 1-26.

Lawless, J.F. (1987). Regression methods for Poisson process data. *J. Amer. Statist. Assoc.*, Sept. 1987.

Manton, K.G.; Woodbury, M.A., and Stallard, E. (1981). A variance components approach to categorical data models with heterogeneous cell populations: Analysis of spatial gradients in lung cancer mortality rates in North Carolina counties. *Biometrics*, 37, 259-269.

Margolin, B.H.; Kaplan, N., and Zeiger, E. (1981). Statistical anaylsis of the Ames salmonella/microsome test. *Proc. Nat. Acad. Sci. U.S.A.*, 76, 3779-3783.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, 11, 59-67.

McCullagh, P., and Nelder, J.A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

Moore, D.F. (1985). Ph.D. Thesis, Univ. of Washington, Seattle.

Moore, D.F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika*, 23, 583-588.

Moran, P.A.P. (1971). Maximum likelihood estimation in non-standard conditions. *Proc. Cambridge Philos. Soc.*, 70, 441-450.

Nelder, J.A., and Pregibon, D. (1987). Quasi-likelihood and generalized linear models. *Biometrika*, to appear.

Olkin, I.; Petkau, A.J., and Zidek, J.V. (1981). A comparison of m estimators for the bionomial distribution. *J. Amer. Statist. Assoc.* 76, 637-642.

Paul, S.R., and Plackett, R.L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika*, 65, 591-602.

Sichel, H. (1982). Asymptotic efficiencies of three methods of estimation for the inverse Gaussian-Poisson distribution. *Biometrika*, 69, 467-472.

Stirling, W.D. (1984). Iteratively reweighted least squares for models with a linear part. *Appl. Statist.*, 33, 1-17.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

Whittle, P. (1961). Gaussian estimation in stationary time series. *Bull. Internat. Statist. Inst.*, 39, 1-26.

Received 9 October 1986
Revised 18 June 1987
Accepted 18 June 1987

Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario N2L 3G1