Biometrika (2012), **99**, 1, *pp.* 230–237
doi: 10.1093/biomet/asr083
Advance Access publication 29 December 2011

Estimating overdispersion when fitting a generalized linear model to sparse data

By D. J. FLETCHER

Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin, New Zealand
dfletcher@maths.otago.ac.nz

Summary

We consider the problem of fitting a generalized linear model to overdispersed data, focussing on a quasilikelihood approach in which the variance is assumed to be proportional to that specified by the model, and the constant of proportionality, ϕ, is used to obtain appropriate standard errors and model comparisons. It is common practice to base an estimate of ϕ on Pearson's lack-of-fit statistic, with or without Farrington's modification. We propose a new estimator that has a smaller variance, subject to a condition on the third moment of the response variable. We conjecture that this condition is likely to be achieved for the important special cases of count and binomial data. We illustrate the benefits of the new estimator using simulations for both count and binomial data.

Some key words: Generalized linear model; Lack-of-fit; Overdispersion; Pearson's statistic; Quadratic estimating equation; Sparse data.

1. Introduction

When fitting a generalized linear model to data, it is common to allow for the possibility that the variance of the response variable exceeds that specified by the model. This phenomenon of overdispersion has been discussed extensively in the literature (Breslow, 1984; Lawless, 1987; McCullagh & Nelder, 1989; Hinde & Demétrio, 1998; Lindsey, 1999; Browne et al., 2005; Haining et al., 2009). Failure to allow for it can lead to underestimation of standard errors and potentially misleading model comparisons.

One method for dealing with overdispersion is to adopt a quasilikelihood approach, specifying only the mean and variance of the response variable. In particular, it is common to assume that the variance is proportional to that specified by the generalized linear model. The constant of proportionality, ϕ, provides a measure of the amount of overdispersion, and can be used to provide more reliable standard errors and model comparisons. One could model the dispersion directly (Smyth, 1989), but for sparse data this will be difficult and is likely to be less robust than quasilikelihood.

It is common practice to estimate ϕ by dividing Pearson's lack-of-fit statistic, P, by the residual degrees of freedom, as suggested by Wedderburn (1974). In the related context of testing lack-of-fit, it is well known that the use of P may not be reliable for sparse data. Thus, Farrington (1996) proposed a modification of Pearson's statistic, P_F, that has a smaller asymptotic variance than P, particularly for sparse data. Farrington (1995) proposed estimating ϕ using P_F divided by the residual degrees of freedom.

In developing his estimator, Farrington (1995) needed an assumption about the third moment of the response variable. The purpose of this paper is to propose a new estimator of ϕ which, under a less-restrictive third-moment assumption, has a smaller asymptotic variance than either of the existing estimators, where the asymptotic limit is one in which the number of observations is large but the data may be sparse. We conjecture that this assumption will often be met for the important special cases of count and binomial data.

2. Notation and current methods

Let Y_i be independent random variables with mean μ_i and variance ϕV_i, where V_i is the variance that Y_i $(i = 1, \ldots, n)$ is assumed to have under the generalized linear model. In addition, suppose that $\mu_i = h(\eta_i)$

and $\eta = X\beta$, where $h(\cdot)$ is the inverse link function, $X = [x_{ir}]$ is the $n \times p$ model matrix and β is the $p \times 1$ vector of model parameters.

Wedderburn (1974) suggested estimating ϕ using $\hat{\phi}_P = P/(n-p)$, where $P = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2/\hat{V}_i$ and $\hat{\mu}_i$, $\hat{V}_i$ are obtained by setting β equal to its quasilikelihood estimate, $\hat{\beta}$. Throughout the paper we use an asterisk to denote an estimator of ϕ that uses the sample size as the denominator rather than the residual degree of freedom, for example $\hat{\phi}_P^* = P/n$.

In the context of assessing lack-of-fit of a generalized linear model to sparse data, Farrington (1996) considered estimation of ϕ solely as a means of testing the hypothesis that $\phi = 1$, and proposed the estimator

$$\hat{\phi}_F = \frac{P_F}{n-p} = \hat{\phi}_P - \frac{n\bar{s}}{n-p}, \tag{1}$$

where

$$P_F = P - \sum_{i=1}^n s_i, \quad s_i = \frac{\hat{V}_i'}{\hat{V}_i}(y_i - \hat{\mu}_i),$$

$\hat{V}_i' = \partial \hat{V}_i/\partial \hat{\mu}_i$ and $\bar{s} = \sum_{i=1}^n s_i/n$. He derived this estimator by solving $g(\hat{\beta}, \hat{\phi}) = 0$, where

$$g(\beta, \phi) = \sum_{i=1}^n a_i(y_i - \hat{\mu}_i) + \sum_{i=1}^n \left\{ \frac{(y_i - \hat{\mu}_i)^2}{\hat{V}_i} - \phi \right\}, \tag{2}$$

and where a_i can depend on β but not ϕ. Farrington (1996) showed that the choice $a_i = -V_i'/V_i$ leads to an estimator with desirable properties, including minimizing the variance within the class of estimators defined by (2). The resulting estimator is $\hat{\phi}_F^* = P_F/n$. The solution of (2) for $a_i = 0$ is $\hat{\phi}_P^*$.

Farrington (1995) proposed that $\hat{\phi}_F$ also be used for the more general case in which ϕ is unknown, provided the third cumulant of Y_i is $\kappa_{3i} = \phi^2\kappa_{3i}^*$, where $\kappa_{3i}^* = V_i V_i'$ is the third cumulant of Y_i assumed under the generalized linear model and $V_i' = \partial V_i/\partial \mu_i$.

For count and binomial data, the adjustment to $\hat{\phi}_P$ in (1) will be influenced by the sparseness of the data, as the s_i terms will be larger when the data are sparser. Thus, for count data, $V_i'/V_i = 1/\mu_i$ is large when μ_i is small; for binomial data, in the usual notation, the absolute value of $V_i'/V_i = (1-2\pi_i)/\{n_i\pi_i(1-\pi_i)\}$ is large when π_i is close to 0 or 1 and n_i is small.

3. New method

We derive a new estimator of ϕ by allowing the a_i in (2) to depend on both β and ϕ. Let $\hat{\phi}(a_i)$ denote the resulting estimator of ϕ.

Theorem 1. *The mean and variance of $\hat{\phi}(a_i)$ are*

$$\begin{aligned} E\{\hat{\phi}(a_i)\} = \phi\left(1 - \frac{p}{n}\right) &- \frac{\phi}{2n}\sum_{i=1}^n \left[\{\phi V_i'' + 2(d_i - c_i)V_i'\}\frac{h_i'^2}{V_i} + c_i h_i''\right] Q_{ii} \\ &+ \frac{\phi}{2n}\sum_{i=1}^n\sum_{j=1}^n c_i \frac{h_j''}{V_j} h_i' h_j' Q_{ij} Q_{jj} + \frac{\phi}{n^2}\sum_{i=1}^n \frac{\partial a_i}{\partial \phi} d_i V_i \\ &- \frac{\phi}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{\partial a_i}{\partial \phi} d_j h_i' h_j' Q_{ij} + O(n^{-2}) \end{aligned} \tag{3}$$

and

$$\begin{aligned} \mathrm{var}\{\hat{\phi}(a_i)\} = \frac{\phi^2}{n^2}\sum_{i=1}^n \gamma_i &+ \frac{\phi}{n^2}\sum_{i=1}^n d_i^2 V_i \\ &- \frac{\phi}{n^2}\sum_{i=1}^n\sum_{j=1}^n c_i(2d_j - c_j)h_i' h_j' Q_{ij} + O(n^{-2}), \end{aligned} \tag{4}$$

where $c_i = a_i + \phi V_i'/V_i$, $d_i = a_i + \kappa_{3i}/(\phi V_i^2)$, $h_i' = \partial h(\eta_i)/\partial\eta_i$, Q_{ij} *is the* (i, j)*th element of the covariance matrix* Q *for* $\hat\eta$ *when* $\phi = 1$, *with* $Q = X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}$, $W = \mathrm{diag}(w_i)$, $w_i = h_i'^2/V_i$, $\gamma_i = \kappa_{4i}/(\phi^2 V_i^2) - \kappa_{3i}^2/(\phi^3 V_i^3) + 2$ *and* κ_{3i}, κ_{4i} *are the third and fourth cumulants of* Y_i*, respectively.*

Proof. The proof is given in the Appendix. □

The results (3) and (4) generalize those in Farrington (1996) and in Friedl (1997). Those in Friedl (1997) are obtained by setting $\partial a_i/\partial\phi = 0$ and $\kappa_{3i} = \phi^2 V_i V_i'$, while those in Farrington (1996) are obtained by additionally setting $\phi = 1$. In Farrington (1996), the expression given for $\Delta_{11,rs}$ incorrectly omits the minus sign and the term h_i''/V_j in his (6) should be h_j''/V_j.

Inspection of (3) and (4) does not immediately suggest an optimal choice of a_i, in terms of minimizing either the absolute bias or the variance. The choice $a_i = -\phi V_i'/V_i$ is of interest, in that it eliminates several terms in both the bias and the variance. On the other hand, the choices $a_i = 0$ and $a_i = -V_i'/V_i$, which correspond to $\hat\phi_P^*$ and $\hat\phi_F^*$, respectively, both give $\partial a_i/\partial\phi = 0$ and so eliminate the last two bias terms. For the special case, $\kappa_{3i} = \phi^2 V_i V_i'$, these terms are also eliminated when $a_i = -\phi V_i'/V_i$.

Let $\hat\phi^*$ denote the solution of (2) when we set $a_i = -\phi V_i'/V_i$. Then

$$\hat\phi^* = \frac{\sum_{i=1}^n \hat V_i^{-1}(y_i - \hat\mu_i)^2}{n + \sum_{i=1}^n \hat V_i' \hat V_i^{-1}(y_i - \hat\mu_i)}.$$

The first term in (3) suggests the alternative estimator

$$\hat\phi = \frac{n\hat\phi^*}{n-p} = \frac{\hat\phi_P}{1+\bar s}. \tag{5}$$

Comparison of (1) and (5) shows that $\hat\phi_F$ and $\hat\phi$ involve different types of adjustment to $\hat\phi_P$ using $\bar s$, the former being arithmetic and the latter being geometric in nature.

We can obtain a clearer preference for use of $\hat\phi^*$ rather than $\hat\phi_P^*$ or $\hat\phi_F^*$ if we focus on estimators for which $a_i = aV_i'/V_i$, as all three estimators belong to this class, with $a = -\phi$, 0 and -1, respectively. Often, for count and binomial data, the mechanism that leads to overdispersion will be such that $\kappa_{3i} = \alpha V_i V_i'$, where $\alpha \geqslant \phi^2$. To see this, consider first the case of count data, where $V_i = \mu_i$ and $V_i' = 1$. The condition can then be written as $\kappa_{3i} \geqslant \kappa_{2i}^2/\kappa_{1i}$. Suppose Y has a Poisson stopped-sum distribution, i.e., $Y = \sum_{i=1}^N C_i$, where $N \sim \mathrm{Poisson}(\lambda)$ and $C_1, \ldots, C_N$ are independent, identically distributed random variables that are also independent of N. The rth cumulant of Y is $\kappa_r = \lambda\mu_r'$, where $\mu_r' = E(C^r)$ (Johnson et al., 2005, § 9.3). The condition is therefore satisfied if $\mu_3'\mu_1' \geqslant \mu_2'^2$. As C is a count random variable, this inequality can be written as

$$\sum_{k=0}^\infty k^3 p_k \sum_{k=0}^\infty k p_k \geqslant \left(\sum_{k=0}^\infty k^2 p_k\right)^2$$

where $p_k = \mathrm{pr}(C = k)$. By writing $k^3 = (k^{3/2})^2$ and $k^2 = (k^{1/2})^2$, this result follows from the Cauchy–Schwarz inequality. The class of Poisson stopped-sum distributions includes the negative binomial, Neyman Type A, Polya–Aeppli and Hermite distributions. In addition, two Poisson mixture distributions, the Poisson-lognormal and Poisson-inverse Gaussian satisfy the condition, having $\alpha = \phi(2\phi - 1) + u(\phi - 1)^2$, with $u > 1$, and $\alpha = 3\phi(\phi - 1) + 1$, respectively.

For binomial data, we have $V_i = n_i\pi_i(1 - \pi_i)$ and $V_i' = 1 - 2\pi_i$. When n_i is constant, generating Y_i from a beta binomial distribution, parameterized so that $\mathrm{var}(Y_i) = \phi n_i\pi_i(1 - \pi_i)$, leads to $\alpha = d\phi^2$ with $d \geqslant 1$. It is not clear whether the condition is satisfied by other binomial mixture models; the logistic-normal, for example, has no closed form for the moments.

As many distributions satisfy the condition $\kappa_{3i} = \alpha V_i V_i'$, where $\alpha \geqslant \phi^2$, we develop the following results.

THEOREM 2. *Let $\hat{\phi}(a)$ denote the solution to (2) when $a_i = aV_i'/V_i$. If $\kappa_{3i} = \alpha V_i V_i'$, we have*

$$E\{\hat{\phi}(a)\} = e_0 + e_1(a) + O(n^{-2}), \quad \mathrm{var}\{\hat{\phi}(a)\} = v_0 + v_1(a) + O(n^{-2}), \tag{6}$$

where

$$e_0 = \phi\left(1 - \frac{p}{n}\right) - \frac{\phi}{2n}\sum_{i=1}^{n}\left\{\phi V_i'' + 2\left(\frac{\alpha}{\phi} - \phi\right)\frac{V_i'^2}{V_i}\right\} w_i Q_{ii},$$

$$e_1(a) = -\frac{\phi}{2n}(a+\phi)S_1 + \frac{\phi}{n^2}\frac{\partial a}{\partial \phi}\left(a + \frac{\alpha}{\phi}\right)S_2,$$

$$v_0 = \frac{\phi^2}{n^2}\sum_{i=1}^{n}\gamma_i + \frac{\phi}{n^2}\left(\frac{\alpha}{\phi} - \phi\right)^2\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{V_i'V_j'}{V_iV_j}h_i'h_j'Q_{ij}, \quad v_1(a) = \frac{\phi}{n^2}\left(a + \frac{\alpha}{\phi}\right)^2 S_2,$$

$$S_1 = \sum_{i=1}^{n}\frac{V_i'}{V_i}h_i''Q_{ii} - \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{V_i'h_j''}{V_iV_j}h_i'h_j'Q_{ij}Q_{jj}, \quad S_2 = \sum_{i=1}^{n}\frac{V_i'^2}{V_i} - \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{V_i'V_j'}{V_iV_j}h_i'h_j'Q_{ij}.$$

Proof. This follows directly from substitution of $a_i = aV_i'/V_i$ and $\kappa_{3i} = \alpha V_i V_i'$ in (3) and (4). □

THEOREM 3. *Under the assumption that $\kappa_{3i} = \alpha V_i V_i'$, with $\alpha \geqslant \phi^2$, and ignoring terms of $O(n^{-2})$, we have* $\mathrm{var}(\hat{\phi}^*) \leqslant \mathrm{var}(\hat{\phi}_F^*) < \mathrm{var}(\hat{\phi}_P^*)$. *This implies that* $\mathrm{var}(\hat{\phi}) \leqslant \mathrm{var}(\hat{\phi}_F) < \mathrm{var}(\hat{\phi}_P)$.

Proof. First, we note that $v_1(a)$ is nonnegative, since $S_2 = z^{\mathrm{T}}\{I - U(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}\}z$ is the residual sum of squares for the regression of z on U, where $z^{\mathrm{T}} = (z_1, \ldots, z_n)$, $z_i = V_i'/V_i^{1/2}$ and $U = W^{1/2}X$. Secondly, we also have

$$\left(\frac{\alpha}{\phi} - \phi\right)^2 \leqslant \left(\frac{\alpha}{\phi} - 1\right)^2 < \left(\frac{\alpha}{\phi}\right)^2 \tag{7}$$

whenever $\alpha \geqslant \phi^2$ and $\phi \geqslant 1$. □

The inequalities in (7) imply that the reduction in variance obtained by using $\hat{\phi}^*$ increases with ϕ. Conversely, if $\phi \approx 1$, there will be little difference between $\hat{\phi}^*$ and $\hat{\phi}_F^*$, both of which are consistently better than $\hat{\phi}_P$, even when $\phi = 1$.

Inspection of (6) does not immediately suggest that using $a = -\phi$ leads to minimizing the absolute bias. However, in general, the squared bias is typically $O(n^{-2})$ and the variance is $O(n^{-1})$, regardless of the choice of a, so minimization of the variance is more important.

As discussed by Farrington (1996), the class of estimators defined by (2) can be viewed within the context of quadratic estimating equations (Crowder, 1987). Using Crowder (1987, (4.1)), it can be shown that, if $\kappa_{3i} = \phi^2\kappa_{3i}^*$ and $\kappa_{4i} = \phi^3\kappa_{4i}^*$, where κ_{3i}^* and κ_{4i}^* are the third and fourth cumulants of Y_i assumed under the generalized linear model, then use of $\hat{\beta}$ and $\hat{\phi}^*$ is optimal, in that their asymptotic covariance matrix is as small as possible, in the sense defined by Crowder (1987). Although these conditions on the third and fourth moments are unlikely to be met exactly for count and binomial data, this result lends further support to the use of $\hat{\phi}$.

4. EXAMPLE AND SIMULATIONS

We illustrate the difference between the three methods using data on the number of deaths of New Zealand sea lions caused by drowning in squid-fishery trawl nets near the Auckland Islands. The data span the period 1990–1996; for simplicity of presentation, we ignore the effect of year and focus on that of season. The data were first summarized by calculating the number of sea lions caught per sample unit, which was defined as a set of consecutive tows on a single vessel during a 6-hour period. Table 1 shows the number of sample units for which there were 0, 1, 2 or 3 sea lions caught, separately for each season. Each sample unit corresponds to a different number of tows; the mean number of tows per sample unit

Table 1. *Data from the sea lion study, showing the frequency of sample units, classified according to the number of sea lions killed and the season. The corresponding mean number of tows per sample unit is shown in parentheses*

	Sea lions killed			
	0	1	2	3
Spring	20 (7·1)	1 (4·0)	0	0
Summer	87 (6·9)	3 (5·3)	0	0
Autumn	125 (9·3)	27 (19·8)	4 (7·5)	2 (14·5)

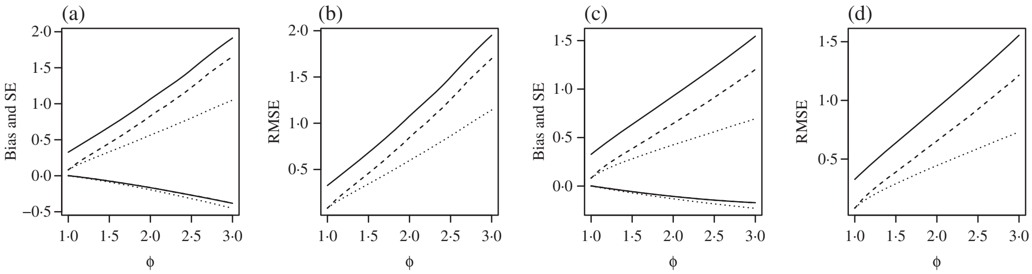

Fig. 1. Poisson simulation results showing the bias, standard error, SE, and the root mean squared error, RMSE, for $\hat{\phi}_P$ (solid), $\hat{\phi}_F$ (dashed) and $\hat{\phi}$ (dotted), when the response variable has a negative binomial distribution, (a) and (b), or a Neyman Type A distribution, (c) and (d). The bias (negative) and standard error (positive) are shown on the same graph for comparison purposes. The simulation standard error of the bias is at most 0·019.

ranges from 4·0 to 19·8. This example is typical of a situation in which it is difficult to see how we might overcome some of the problems with sparseness by pooling the data.

If we fit a Poisson regression model with season as a predictor variable and the logarithm of the number of tows as an offset, the estimates of dispersion are $\hat{\phi}_P = 2{\cdot}17$, $\hat{\phi}_F = 1{\cdot}19$ and $\hat{\phi} = 1{\cdot}10$. Thus $\hat{\phi}_P$ suggests that there is substantial overdispersion, whereas both $\hat{\phi}_F$ and $\hat{\phi}$ suggest that there is relatively little. The estimated means are all less than unity, suggesting that $\hat{\phi}_P$ is unlikely to be reliable.

In order to compare the performance of the three methods in this setting, we carried out a simulation study. The number of sample units per season and the number of tows per sample unit were set equal to those observed, and β was set equal to its quasilikelihood estimate. We specified a true value for ϕ and then generated Y_i using either a negative binomial or a Neyman Type A distribution. We considered values of ϕ ranging from 1 to 3 and used 10^6 simulations. For the special case $\phi = 1$, we generated the data using a Poisson model. For each value of ϕ, we calculated the bias, standard error and square root of the mean squared error of each estimator. Occasionally, all the values of Y within a season were zero; in these cases, we ignored the data for that season when calculating the estimate.

Figure 1 summarizes the full set of simulation results. As expected from the asymptotic results, for both mechanisms of generating the overdispersion, $\hat{\phi}$ has the lowest variance. All three estimators are negatively biased; the bias increases with ϕ, and $\hat{\phi}$ is slightly more biased than the other two estimators. This extra bias is more than offset by the reduction in variance, as shown by the square root of the mean squared error. The benefits of using $\hat{\phi}$ in this setting are most apparent when ϕ is larger. When $\phi \approx 1$, there is little to choose between $\hat{\phi}$ and $\hat{\phi}_F$, which both outperform $\hat{\phi}_P$, again as suggested by the asymptotic results.

Figure 2 gives some insight into the behaviour of the three estimators when $\phi = 2$, for a subset of 10^4 simulations. For both types of overdispersion mechanism, $\hat{\phi}_F$ tends to be consistently smaller than $\hat{\phi}_P$, while $\hat{\phi}$ is in turn generally smaller than $\hat{\phi}_F$, the effect increasing with $\hat{\phi}_F$. These results reflect the form of the expressions in (1) and (5), with $\bar{s}$ ranging from $-0{\cdot}55$ to $2{\cdot}40$ for the negative binomial case and

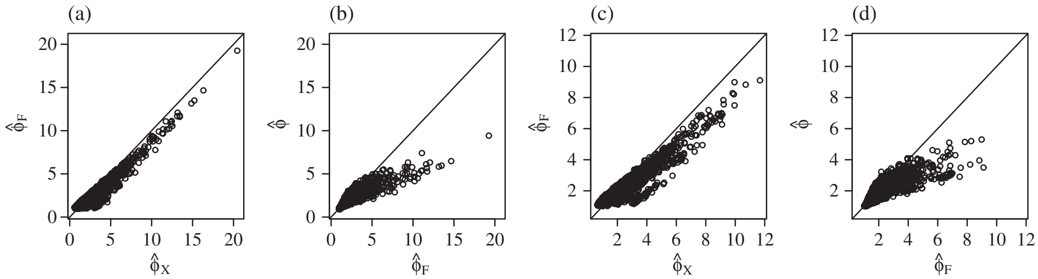

Fig. 2. The relationship between $\hat{\phi}_F$ and $\hat{\phi}_P$ and between $\hat{\phi}$ and $\hat{\phi}_F$ from the Poisson simulations, when the response variable has a negative binomial distribution, (a) and (b), or a Neyman Type A distribution, (c) and (d), and when $\phi = 2$.

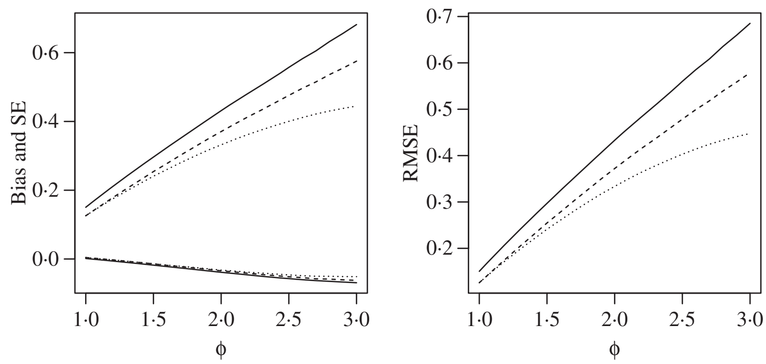

Fig. 3. Binomial simulation results showing the bias, standard error, SE, and the root mean squared error, RMSE, for $\hat{\phi}_P$ (solid), $\hat{\phi}_F$ (dashed) and $\hat{\phi}$ (dotted), when the response variable has a beta binomial distribution. The bias, negative, and standard error, positive, are shown on the same graph for comparison purposes. The simulation standard error of the bias is at most 0·007.

from −0·52 to 2·67 for the Neyman Type A case. Overall, the main benefit of using $\hat{\phi}$ is the geometric manner in which it shrinks $\hat{\phi}_P$.

The results from an analogous set of simulations involving binomial data are shown in Fig. 3. These involved a logistic link and a single predictor variable $x = (0, 1/(n-1), \ldots, 1)^{\mathrm{T}}$, with $n = 100$, $n_i = 5$ $(i = 1, \ldots, n)$ and $\beta = (-4, 8)^{\mathrm{T}}$. The results are similar to those for count data, except that the three estimators now have essentially the same bias.

5. Discussion

Theorem 3 also suggests a new estimator for the dispersion parameter in exponential family models, such as the gamma and inverse Gaussian models. These two models can be parameterized so that $\mathrm{var}(Y_i) = \phi V_i$, with $V_i = \mu_i^2$ and $V_i = \mu_i^3$, respectively. They then have $\kappa_{3_i} = \phi^2 V_i V_i'$, and $\hat{\phi}^*$ therefore minimizes the variance amongst the class of estimators for which $a_i = a V_i'/V_i$. For the gamma model, this means that $\hat{\phi}$ will have a smaller variance than both the estimator in McCullagh & Nelder (1989, § 8.3.6) and the modification proposed in Farrington (1995).

In principle, one might consider calculating a confidence interval for $\hat{\phi}$ by estimating its variance. In doing so, it can be argued that it is more appropriate to use an estimate of the variance conditional on $\hat{\beta}$ (McCullagh, 1985; Farrington, 1996). Either way, this involves estimation of the third and fourth moments, and is therefore likely to be prone to problems when the data are sparse. For this reason, and for simplicity of presentation, we have omitted any derivation of conditional moments. The issue of constructing a reliable confidence interval for ϕ is certainly a topic worthy of further research.

Acknowledgement

I am grateful to the referees for very helpful comments, and for one referee pointing out the result for Poisson stopped-sum distributions. I am also grateful to Byron Morgan for remarks on a first draft, and for his hospitality at the University of Kent.

Appendix

Proof of Theorem 1

From (4) of Farrington (1996) we have

$$E\{\hat{\phi}(a_i)\} = \phi + \frac{1}{n}E(Z_2^2) + O(n^{-2}) \tag{A1}$$

and

$$\operatorname{var}\{\hat{\phi}(a_i)\} = \frac{1}{n}E(Z_1^2 Z_1^{2\mathrm{T}}) + O(n^{-2}), \tag{A2}$$

where

$$Z_1^2 = -n^{-1/2}(\Delta_{21}\Delta_{11}^{-1}g^1 - g^2),$$

$$Z_2^2 = \Delta_{21}\Delta_{11}^{-1}\left(-\frac{1}{2}c^1 + \frac{1}{n}E_{11}\Delta_{11}^{-1}g^1\right) + \frac{1}{2}c^2 - \frac{1}{n}E_{21}\Delta_{11}^{-1}g^1 - \frac{1}{n}E_{22}(\Delta_{21}\Delta_{11}^{-1}g^1 - g^2),$$

$$\Delta_{11,rs} = -\frac{1}{n}\sum_{i=1}^{n}\frac{h_i'^2}{V_i}x_{ir}x_{is}, \quad \Delta_{21,s} = -\frac{1}{n}\sum_{i=1}^{n}\left(a_i + \phi\frac{V_i'}{V_i}\right)h_i'x_{is}, \quad g^1 = (g_1, \ldots, g_p)^{\mathrm{T}},$$

$$g_r = \sum_{i=1}^{n}\frac{y_i - \mu_i}{V_i}h_i'x_{ir}, \quad g^2 = \sum_{i=1}^{n}a_i(y_i - \mu_i) + \sum_{i=1}^{n}\left\{\frac{(y_i - \mu_i)^2}{V_i} - \phi\right\},$$

$$c^1 = (c_1^1, \ldots, c_p^1)^{\mathrm{T}}, \quad c_r^1 = \frac{1}{n}g^{1\mathrm{T}}\Delta_{11}^{-1}\nu_r^{11}\Delta_{11}^{-1}g^1,$$

$$c^2 = \frac{1}{n}(g^{1\mathrm{T}}\Delta_{11}^{-1}\nu_q^{11}\Delta_{11}^{-1}g^1 + 2g^{1\mathrm{T}}\Delta_{11}^{-1}\nu_q^{12}\Delta_{21}\Delta_{11}^{-1}g^1 - 2g^{1\mathrm{T}}\Delta_{11}^{-1}\nu_q^{12}g^2),$$

$$E_{11,rs} = \sum_{i=1}^{n}\frac{y_i - \mu_i}{V_i}\left(\frac{h_i''}{h_i'^2} - \frac{V_i'}{V_i}\right)h_i'^2x_{ir}x_{is},$$

$$E_{21,s} = \sum_{i=1}^{n}\left\{\left(\frac{\partial a_i}{\partial\eta_i} - \frac{2}{V_i}\right)(y_i - \mu_i) - \frac{V_i'}{V_i}\left(\frac{(y_i - \mu_i)^2}{V_i} - \phi\right)\right\}h_i'x_{is},$$

$$E_{22} = \sum_{i=1}^{n}\frac{\partial a_i}{\partial\phi}(y_i - \mu_i), \quad \nu_{r,st}^{11} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_i}\left(2\frac{V_i'}{V_i} - 3\frac{h_i''}{h_i'^2}\right)h_i'^3x_{ir}x_{is}x_{it},$$

$$\nu_{q,st}^{11} = -\frac{1}{n}\sum_{i=1}^{n}\left\{2\left(\frac{\partial a_i}{\partial\eta_i} - \frac{1}{V_i} - \phi\frac{V_i'^2}{V_i^2}\right) + \phi\frac{V_i''}{V_i} + \left(a_i + \phi\frac{V_i'}{V_i}\right)\frac{h_i''}{h_i'^2}\right\}h_i'^2x_{is}x_{it},$$

$$\nu_{q,s}^{12} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial a_i}{\partial\phi}h_i'x_{is}.$$

If we make use of the following identities

$$\sum_{i=1}^{n}\sum_{j=1}^{n}x_{ir}\Delta_{11,rs}^{-1}x_{js} = -nQ_{ij}, \quad \sum_{k=1}^{n}\frac{h_k'^2}{V_k}Q_{ik}Q_{kj} = Q_{ij}$$

then evaluation of (A1) and (A2) leads to the expressions given in Theorem 1.

References

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

Browne, W. J., Subramanian, S. V., Jones, K. & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Statist. Soc.* A **168**, 599–613.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591–7.

Farrington, C. P. (1995). Pearson statistics, goodness of fit, and overdispersion in generalised linear models. In *Proc. 10th Int. Workshop Statist. Mod.*, Ed. G. Seeber, B. Francis, R. Hatzinger and G. Steckel-Berger, pp. 109–16. New York: Springer.

Farrington, C. P. (1996). On assessing goodness of fit of generalized linear models to sparse data. *J. R. Statist. Soc. B* **58**, 349–60.

Friedl, H. (1997). On the asymptotic moments of Pearson type statistics based on resampling procedures. *Comp. Statist.* **12**, 265–78.

Haining, R., Law, J. & Griffith, D. (2009). Modelling small area counts in the presence of overdispersion and spatial autocorrelation. *Comp. Statist. Data Anal.* **53**, 2923–37.

Hinde, J. & Demétrio, C. G. B. (1998). Overdispersion: models and estimation. *Comp. Statist. Data Anal.* **27**, 151–70.

Johnson, N. L., Kemp, A. W. & Kotz, S. (2005). *Univariate Discrete Distributions*, 2nd edn. New York: Wiley.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Can. J. Statist.* **15**, 209–25.

Lindsey, J. K. (1999). On the use of corrections for overdispersion. *Appl. Statist.* **48**, 553–61.

McCullagh, P. (1985). On the asymptotic distribution of Pearson's statistic in linear exponential-family models. *Int. Statist. Rev.* **53**, 61–7.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. R. Statist. Soc.* B **51**, 47–60.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–47.

[*Received October* 2011. *Revised December* 2011]