Series evaluation of Tweedie exponential dispersion model densities*

Peter K. Dunn
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba, Qld 4350, Australia

Gordon K. Smyth
Genetics and Bioinformatics Division
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3052, Australia
smyth@wehi.edu.au

23 February 2005

Abstract

Exponential dispersion models, which are linear exponential families with a dispersion parameter, are the prototype response distributions for generalized linear models. The Tweedie families are those exponential dispersion models with power mean-variance relationships. The normal, Poisson, gamma and inverse Gaussian distributions are Tweedie families. Apart from these special cases, Tweedie distributions do not have density functions which can be written in closed form. Instead, the densities can be represented as infinite summations derived from series expansions. This article describes how the series expansions can be summed in an numerically efficient fashion. The usefulness of the approach is demonstrated, but full machine accuracy is shown not to be obtainable using the series expansion method for all parameter values. Derivatives of the density with respect to the dispersion parameter are also derived to facilitate maximum likelihood estimation. The methods are demonstrated on two data examples and compared with with Box-Cox transformations and extended quasi-likelihoood.

*This is the final preprint version of the article published in *Statistics and Computing*, Volume 15 (2005), pages 267–280.

1 Introduction

An exponential dispersion model (EDM) is a two-parameter family of distributions consisting of a linear exponential family with an additional dispersion parameter. EDMs are important in statistics because they are the response distributions for generalized linear models (McCullagh and Nelder, 1989). EDMs were established as a field of study in their own right by Jørgensen (1987, 1997), who undertook a detailed study of their properties.

Any EDM can be characterized by its variance function $V()$, which describes the mean-variance relationship of the distribution when the dispersion is held constant. If Y follows an EDM distribution with mean μ, variance function $V()$ and dispersion ϕ, then the variance of Y can be written

$$\text{var}(Y) = \phi V(\mu).$$

Of special interest are the class of EDMs with power mean-variance relationships for which $V(\mu) = \mu^p$ for some p. Following Jørgensen (1987, 1997), we call these Tweedie models. The class of Tweedie models includes most of the important distributions commonly associated with generalized linear models including the normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$) and the inverse Gaussian ($p = 3$) distributions. Although the other Tweedie model distributions are less well known, Tweedie models exist for all values of p outside the interval $(0, 1)$. Apart from the four well-known distributions already mentioned, none of the Tweedie models have density functions which have explicit analytic forms. The purpose of this article is to provide fast, accurate computation of these densities.

The Tweedie models for $p > 2$ are generated by stable distributions and have support on the positive reals. The Tweedie model distributions for $1 < p < 2$ can be represented as Poisson mixtures of gamma distributions and are mixed distibutions with mass at zero and with support on the non-negative reals. These distributions have been called "compound Poisson" by Bar-Lev and Stramer (1987), Feller (1968, Section XII.2), Jørgensen and Paes de Souza (1994) and Smyth and Jørgensen (2002) and "compound gamma" by Johnson and Kotz (1970). In this article we call them Poisson-gamma distributions as in Smyth (1996) in recognition of their relationship to both distributions. All Tweedie distributions with $p > 1$ have strictly positive means, $\mu > 0$. Jørgensen (1987) showed that Tweedie model distributions with $p < 0$ have $\mu > 0$ but support for y on the whole real line. We do not give attention to these distributions in this article as they seem to have limited potential application.

Apart from applications of the four special Tweedie distributions listed above, Tweedie distributions have been used in such diverse fields as actuarial studies (Haberman and Renshaw, 1996; Renshaw, 1994); Jørgensen and Paes de Souza, 1994; Haberman and Renshaw (1996, 1998); Millenhall (1999); Murphy, Brockman and Lee, 2000; Smyth and Jørgensen, 2002), assay analysis (Davidian, 1990; Davidian, Carroll and Smith, 1988), survival analysis (Aalen, 1992; Hougaard, Harvald and Holm, 1992; Hougaard, 1986), time spent splicing telephone cables (Nelder, 1994), money spent on hiring outside labour (Jørgensen, 1987), and ecology (Perry, 1981). Generalized linear

models with Tweedie model responses have also been used by Gilchrist and Drinkwater (1999) to analysis alcohol consumption in British teenagers and by Smyth (1996) to analyse medical and metereological data.

The densities of Tweedie distributions for $p = 0$, 1, 2 and 3 can be written in closed form. For other values of p, evaluation of the density requires some numerical process such as the inversion of the cumulant generating function, involving evaluation of an infinite oscillating integral, or evaluating an infinite summation. Since there are generally no closed forms for the Tweedie densities, full likelihood analysis is very difficult. This does not prevent the use of Tweedie distributions in generalized linear models since the fitting algorithm requires knowledge only of the first two moments of the response distribution. The likelihood function is very useful however because it enables efficient estimation of the parameters p and ϕ as well as diagnostic checking of the response distribution using techniques such as the quantile residuals of Dunn and Smyth (1996).

Dunn (2001) considers two broad strategies for evaluating Tweedie densities, one based on numerical inversion of the characteristic function and the other based on series expansions obtained from an analytic approach to the inversion integral. Both strategies have advantages and the two strategies are to some extent complementary; they each work best in different regions of the parameter space. The numerical inversion approach requires lengthy technical development and will be published elsewhere. This article focuses on the series expansion method for the case $p > 1$. We describe how the series expansions can be implemented in an numerically efficient fashion. The usefulness of the series expansions is demonstrated, but full machine accuracy is shown not to be obtainable using the series expansion method for all parameter values. It would be possible to take a similar approach to develop expressions for the densities for the case $p < 0$, but these EDMS are likely to be of far less practical importance.

Previously, Smyth (1996) and Gilchrist and Drinkwater (1999) have examined the series expansion of the Tweedie densities for $1 < p < 2$ using a simple summation of terms in the series. Seigel (1979, 1985) discusses series evaluation of a special case of the Tweedie distributions with $p = 1.5$. Jørgensen (1997) discusses the series themselves in detail but not the actual evaluation of the series.

In the next section, the Tweedie densities are discussed and their properties introduced. Section 3 gives the infinite series expansions. Section 4 addresses the issue of selecting which terms in the summation must be summed for a fast and accurate answer. Section 5 examines the computation complexity and accuracy of the numerical summation. Section 6 considers some issues of maximum likelihood estimation. In Section 7, the problem of evaluating derivatives of the density with respect to ϕ is considered. Section 8 discusses a particular problem that emerges with evaluating the densities for $1 < p < 2$ as $\phi \to 0$, while Section 9 considers two data examples. Brief conclusions follow in Section 10.

2 Tweedie densities

EDMs have density functions or probability mass functions of the form

$$f(y;\theta,\phi) = a(y,\phi)\exp\left[\frac{1}{\phi}\{y\theta - \kappa(\theta)\}\right], \tag{1}$$

for suitable known functions $\kappa()$ and $a()$ (Jørgensen, 1997). The domain of the canonical parameter θ is an open interval satisfying $\kappa(\theta) < \infty$ and the dispersion parameter ϕ is positive. The function $\kappa()$ is called the cumulant function of the EDM because, if $\phi = 1$, the derivatives of κ give the successive cumulants of the distribution. In particular, the mean of the distribution is $\mu = \dot{\kappa}(\theta)$ and the variance is $\phi\ddot{\kappa}(\theta)$.

The mapping from θ to μ is invertible, so we may write $\ddot{\kappa}(\theta) = V(\mu)$ for a suitable function $V(\mu)$, called the variance function of the EDM. In this article we are interested in EDMs with variance functions of the form $V(\mu) = \mu^p$ for some p. These families are called Tweedie models because the underlying linear exponential families were first studied systematically by Tweedie (1984). Jørgensen (1987) showed that Tweedie EDMs exist for all values of p outside the interval $(0, 1)$. The notation $Y \sim \mathrm{ED}_p(\mu, \phi)$ is used to indicate that Y is distributed as a Tweedie EDM with mean μ, dispersion ϕ and variance function $V(\mu) = \mu^p$. Tweedie models are the only EDMs which are closed under re-scaling of the response variable: if $Y \sim \mathrm{ED}_p(\mu, \phi)$ then $cY \sim \mathrm{ED}_p(c\mu, c^{2-p}\phi)$ (Jørgensen, 1997, Section 4.1.1). This makes Tweedie EDMs an obvious choice for modeling data when the unit of measurement is arbitrary.

The cumulant function and mean can be be found for Tweedie EDMs by equating $\ddot{\kappa}(\theta) = d\mu/d\theta = \mu^p$ and solving for μ and κ. Setting the arbitrary constants of integration to zero gives

$$\theta = \begin{cases} \dfrac{\mu^{1-p}}{1-p} & p \neq 1 \\ \log\mu & p = 1 \end{cases}$$

and

$$\kappa(\theta) = \begin{cases} \dfrac{\mu^{2-p}}{2-p} & p \neq 2 \\ \log\mu & p = 2 \end{cases}$$

The remaining factor in the density, $a(y, \phi)$, is more difficult to derive. The numerical evaluation of $a(y, \phi)$ is the focus of the remainder of this article.

3 Series expansions

If $Y \sim \mathrm{ED}_p(\mu, \phi)$ with $1 < p < 2$, then Y can be represented as

$$Y = X_1 + X_2 + \cdots + X_N$$

where N has a Poisson distribution and the X_i are independent gamma random variables. Let λ be the mean of N and let $-\alpha$ and γ be the shape and scale parameters

of the X_i, with $-\alpha\gamma$ and $-\alpha\gamma^2$ the mean and variance of X_i respectively. Note that α is chosen negative so that the notation agrees with that used elsewhere in this paper. Then the parameters are related by

$$\begin{aligned}\lambda &= \frac{\mu^{2-p}}{\phi(2-p)} \\ \alpha &= (2-p)/(1-p) \\ \gamma &= \phi(p-1)\mu^{p-1}.\end{aligned}$$

From this or otherwise it can be shown that

$$P(Y=0) = \exp\left\{-\frac{\mu^{2-p}}{\phi(2-p)}\right\}$$

and for $y > 0$ that

$$a(y,\phi) = \frac{1}{y}W(y,\phi,p)$$

with $W(y,\phi,p) = \sum_{j=1}^{\infty} W_j$ and

$$W_j = \frac{y^{-j\alpha}(p-1)^{\alpha j}}{\phi^{j(1-\alpha)}(2-p)^j j!\Gamma(-j\alpha)}. \tag{2}$$

See Jørgensen (1997, p. 141) and Smyth (1996) for details. Tweedie (1984) has identified $W(y,\phi,p)$ as an example of Wright's (1933) generalized Bessel function. It cannot be expressed in terms of more common Bessel functions. For $p = 1.5$ the distribution is the non-central χ^2 distribution of zero degrees of freedom, studied by Seigel (1979, 1985).

A similar series expansion exists for $p > 2$ and is given in various forms by Aalen (1992), Bar-Lev and Enis (1986), Hougaard (1986) and Jørgensen (1997, p. 141). For $p > 2$ we have

$$a(y;\phi) = \frac{1}{\pi y}V(y,\phi,p) \tag{3}$$

with $V = \sum_{k=1}^{\infty} V_k$ and

$$V_k = \frac{\Gamma(1+\alpha k)\phi^{k(\alpha-1)}(p-1)^{\alpha k}}{\Gamma(1+k)(p-2)^k y^{\alpha k}}(-1)^k \sin(-k\pi\alpha). \tag{4}$$

Note that $0 < \alpha < 1$ for $p > 2$. Jørgensen (1997, p. 137, 141) shows that the series expansions with terms (2) and (4) are related though a reflection formula.

Note that terms W_j are all positive while the V_k are both positive and negative. This will limit the numerical accuracy that is obtainable in summing the second series.

4 Which terms to include?

The functions $W(y,\phi,p)$ and $V(y,\phi,p)$ are evaluated in this paper by directly summing the infinite series. A method is needed for determining which terms need to be included

in the summation to achieve acceptable accuracy. The naïve approach of starting at index 1 and adding more terms can result in a very large number of unnecessary terms. The approach of summing a fixed number of terms, even if this number is large, may sum negligible terms and miss those terms that make important contributions. The number of necessary terms can be arbitrarily large and these terms can occur very far from an index of 1.

The strategy used here is to establish where the terms that contribute to the sum of the series are located in terms of the index and sum the necessary terms of the series in that region. Stirling's approximation is used to approximate the gamma functions and then the index of summation is treated as continuous. This enables the approximate terms in the series to be differentiated and the location (in terms of the index of summation) of the maximum value of terms in the summation to be found. To then find the lower and upper limits of the index necessary for accurate evaluation, the approximate terms either side of the maximum are evaluated until the contributions are negligible. The exact terms are then summed over these values of the index.

4.1 Size of summands for $1 < p < 2$

To evaluate the infinite summation for $W(y, \phi, p)$, the value of j is determined for which W_j reaches a maximum. To do this, j is treated as continuous and W_j is differentiated with respect to j and the derivative set to zero.

Write

$$\log W_j = j \log z - \log \Gamma(1 + j) - \log \Gamma(-\alpha j)$$

where

$$z = \frac{y^{-\alpha}(p-1)^{\alpha}}{\phi^{1-\alpha}(2-p)}.$$

Replacing the gamma functions with Stirling's approximation (Abramowitz and Stegun, 1965) and approximating $1 - \alpha j$ with $-\alpha j$ gives

$$\begin{aligned} \log W_j &\approx j\{\log z + (1-\alpha) + \alpha \log(-\alpha) - (1-\alpha)\log j\} \\ &\quad - \log(2\pi) - \frac{1}{2}\log(-\alpha) - \log j. \end{aligned} \tag{5}$$

This approximate is asymptotically accurate for j large and, as it turns out, not bad for j small either. Note that $\alpha < 0$ for $1 < p < 2$ so the logarithms have positive arguments. Differentiating with respect to j gives

$$\begin{aligned} \frac{\partial \log W_j}{\partial j} &\approx \log z - \frac{1}{j} - \log j + \alpha \log(-\alpha j) \\ &\approx \log z - \log j + \alpha \log(-\alpha j), \end{aligned} \tag{6}$$

since the term $1/j$ can be ignored for j large. Note that this derivative is monotonically decreasing in j for $j \geq 0$, so the sequence $\log W_j$ is unimodal in j. Solving $\partial W_j / \partial j = 0$ for j gives the unique solution

$$j_{\max} = \frac{y^{2-p}}{(2-p)\phi}. \tag{7}$$

This approximation is surprisingly accurate. It is easy to confirm numerically that $j_{\max}$ is always within one of the exact index j of the maximum W_j, i.e., the approximation is good for small as well as large values of j. The approximate maximum value of W_j can be found by substituting $j_{\max}$ from (7) into (5) giving

$$\log W_{\max} = j_{\max}(\alpha - 1) - \log(2\pi) - \log j_{\max} - \frac{1}{2}\log(-\alpha). \tag{8}$$

4.2 Size of summands for $p > 2$

A similar approach can be taken with the series for $p > 2$. It should be noted however that the terms (4) involve factors $(-1)^k$ and $\sin(-k\pi\alpha)$ which are of changing sign. To proceed as in the previous section, we need to work with the envelope of the terms rather than individual terms themselves. The envelope is defined as V_k without the $(-1)^k$ and sin terms, so that,

$$V_{\text{env}}(k) \;=\; \frac{z^k \Gamma(1 + \alpha k)}{\Gamma(1 + k)} \tag{9}$$

where

$$z = \frac{(p-1)^\alpha \phi^{\alpha-1}}{y^\alpha (p-2)}.$$

The definition ensures that $|V_k| \le V_{\text{env}}(k)$ for all k. The procedure is then the same as for $1 < p < 2$. Stirling's approximation is used to approximate the gamma functions giving

$$\log V_{\text{env}}(k) \approx k\left[\log z + (1-\alpha) - \log k + \alpha \log(\alpha k)\right] + \frac{1}{2}\log \alpha. \tag{10}$$

Treating k as continuous,

$$\frac{\partial \log V_{\text{env}}}{\partial k} \approx \log z + \alpha \log \alpha + (\alpha - 1)\log k. \tag{11}$$

Equating to zero and solving for k gives the unique maximum

$$k_{\max} = \frac{y^{2-p}}{\phi(p-2)}. \tag{12}$$

Note that $k_{\max} > 0$ since $p > 2$. The similarity with the solution for the case $1 < p < 2$ in (12) is obvious.

An upper bound for the maximum of $|V_k|$ over k can be found by substituting (12) into (10) to give

$$\log V_{\max} = (1-\alpha)k_{\max} + \frac{1}{2}\log \alpha. \tag{13}$$

4.3 Evaluating the series

Our aim is to approximate $W(y,\phi,p)$ with

$$\tilde{W}(y,\phi,p) = \sum_{j=j_L}^{j_U} W_j$$

and $V(y,\phi,p)$ with

$$\tilde{V}(y,\phi,p) = \sum_{k=k_L}^{k_U} V_k$$

where j_L, j_U and k_L, k_U are suitably chosen limits. The fact that $\partial \log W_j/\partial j$ is monotonic decreasing implies that $\log W_j$ is strictly convex as a function of j and hence that the W_j decay faster than geometrically on either side of $j_{\max}$. The approximation error can therefore be bounded by geometric sums,

$$W(y,\phi,p) - \tilde{W}(y,\phi,p) < W_{j_L-1}\frac{1-r_L^{j_L-1}}{1-r_L} + W_{j_U+1}\frac{1}{1-r_U}$$

where

$$r_L = \exp\left(\frac{\partial \log W_j}{\partial j}\right)\Bigg|_{j=j_L-1}$$

and

$$r_U = \exp\left(\frac{\partial \log W_j}{\partial j}\right)\Bigg|_{j=j_U+1}$$

The same sort of bound can be constructed for $V(y,\phi,p) - \tilde{V}(y,\phi,p)$ in terms of $V_{\text{env}}(k)$. In practice these geometric bounds are very conservative so we simply choose $j_L < j_{\max}$ and $j_U > j_{\max}$ such that W_{j_L} and W_{j_U} are less than $\epsilon W_{\max}$ and $k_L < k_{\max}$ and $k_U > k_{\max}$ such that $V_{\text{env}}(k_L)$ and $V_{\text{env}}(k_U)$ are less than $\epsilon V_{\max}$. Here $\epsilon = 10^{-16}$ would ensure double precision accuracy in 64-bit floating point arithmetic. In practice we use $\epsilon = \exp(-37) \approx 8 \times 10^{-17}$, simply searching away from $j_{\max}$ for $1 < p < 2$ until the saddlepoint approximation (5) is less than $\log W_{\max} - 37$ or away from $k_{\max}$ for $p > 2$ until (10) is less than $\log V_{\max} - 37$. If $j_{\max} < 1$ or if $\log W_1 > W_{\max} - 37$, then we set $j_L = 1$. Similarly if $k_{\max} < 1$ or if $\log V_1 > V_{\max} - 37$, then we set $k_L = 1$.

To avoid the possibility of floating point overflow, we compute $\tilde{W}(y,\phi,p)$ and $\tilde{V}(y,\phi,p)$ on the log-scale and standardize the individual terms to have maximum value unity. Specifically, we compute

$$\log \tilde{W}(y,\phi,p) = \log W_{\max} + \log \sum_{j=j_L}^{j_U} w_j$$

where $w_j = \exp(\log W_j - \log W_{\max})$ and

$$\log \tilde{V}(y,\phi,p) = \log V_{\max} + \log \sum_{k=k_L}^{k_U} v_j$$

Table 1: The number of terms required to reach machine accuracy for a Tweedie density for various ϕ and $1 < p < 2$. The value of μ does not affect the required number of terms.

		y					
p	ϕ	0.001	1	5	10	100	1000
1.01	1.00	3	4	8	10	22	58
1.01	0.10	3	10	16	22	58	170
1.01	0.01	3	22	42	58	174	532
1.5	1.00	7	16	25	29	54	100
1.5	0.10	11	54	84	100	173	310
1.5	0.01	29	173	262	310	547	970
1.9	1.00	34	46	50	53	60	68
1.9	0.10	118	166	180	186	208	232
1.9	0.01	368	520	564	584	652	732
1.999	1.00	546	546	548	548	548	550
1.999	0.10	1718	1724	1726	1726	1728	1730
1.999	0.01	5422	5442	5446	5448	5454	5460

where $v_k = \exp(\log V_k - \log W_{\max})$.

The simple summation strategy used in this section for computing the densities lends itself well to vectorized arithmetic such as that found in S-Plus, R or MATLAB. When densities are computed simultaneously for a vector of response values y_i, we have found it useful for fast computation to choose a common j_L for all the y_i to be the minimum of the j_L for the individual y_i and a common j_U to be the maximum of the j_U for the individual responses. This means that unnecessary terms are summed for some values of y_i but this is more than compensated usually by the ability to undertake the summations in parallel. Similar comments apply to k_L and k_U.

5 Accuracy and limitations

For the case $1 < p < 2$, the terms in the summation are always positive and so a simple summation can compute the density to machine accuracy. The interest therefore is in the number of terms required and the potential limitation is that the number of terms may be prohibitive for some parameter values. Table 1 shows the number of terms necessary to reach machine precision for $1 < p < 2$. It can be seen that the number terms necessary for accurate evalution becomes large as p near 2, y large or ϕ small. In fact the number of required terms increases without bound at these limits. This qualitative behavior was expected from the form for $j_{\max}$ given in (7) which is unbounded for p near 2, y large or ϕ small.

In the case $p = 1.5$, our series expansion gives identical numerical results to the expressions published by Seigel (1979, 1985).

Table 2: Comparing the exact and series expansion densities for the inverse Gaussian distribution, $p = 3$ with $\mu = 1.4$ and $\phi = 0.74$. The comparison is excellent for $y > 0.1$ but fails for y near zero.

y	**Exact density**	**Series density**	**Relative error**	**No. terms in Series**
0.001	1.39037×10^{-289}	2.149×10^{284}	$-\infty$	895
0.002	2.58550×10^{-143}	0	1	633
0.005	7.04450×10^{-56}	0	1	399
0.01	5.49261×10^{-27}	0	1	280
0.05	0.0001447812	0.0001446184	0.001	112
0.10	0.0432617075	0.0432617076	-5×10^{-10}	79
0.50	0.7504127835	0.7504127835	-3×10^{-15}	41
1.00	0.4388738851	0.4388738851	-4×10^{-16}	33
2.00	0.1540992189	0.1540992189	-5×10^{-16}	27
3.00	0.0665051333	0.0665051333	0	24
4.00	0.0323731652	0.0323731652	0	22
5.00	0.0169737124	0.0169737124	-2×10^{-16}	21
6.00	0.0093555852	0.0093555852	-4×10^{-16}	20
7.00	0.0053446845	0.0053446845	-3×10^{-16}	20
8.00	0.0031365974	0.0031365974	0	19
9.00	0.0018797070	0.0018797070	-6×10^{-16}	19
10.00	0.0011455103	0.0011455103	-2×10^{-16}	18
15.00	0.0001137801	0.0001137801	-5×10^{-16}	17
20.00	1.3334×10^{-5}	1.3334×10^{-5}	-2×10^{-15}	16

For $p > 2$ there are positive and negative terms in the series expression. Machine accuracy is not generally achievable because subtractive cancellation in floating point arithmetic will overcome precision if the summation converges sufficiently slowly. To evaluate the accuracy of the numerical summation in the one case where an exact analytic expression is available, we compare the series summation for $p = 3$ with the density of the inverse Gaussian distribution (Table 2). Arbitrary but typical values $\mu = 1.4$ and $\phi = 0.74$ were used for the comparison. The inverse Gaussian case presents special problems for the algorithm for small y. For $p = 3$ we have $\alpha = 1/2$ so the terms V_k involve the factor $(-1)^k \sin(-k\pi/2)$ which is zero for k even and alternate in sign for k odd. For very small y, the V_k terms for k odd are large and of alternating sign. The numerical difficulties experienced by the algorithm as evidenced in Table 2 are the result of subtractive cancellation of very large but almost equal quantities of opposite sign. On the other hand, the series expansion performs well for $y > 0.1$ where the absolute relative error is less than 10^{-14}.

The number of terms in the series approximation is shown in Table 3 for various $p > 2$. For most parameter values the summation is easily manageable. However for p close to 2 or y or ϕ small, the number of terms required becomes large and the series

Table 3: The number of terms required to reach nominal machine accuracy for a Tweedie density for various ϕ and $p > 2$.

		y					
p	ϕ	0.01	1	5	10	100	1000
2.01	0.01	1774	1732	1718	1714	1694	1674
2.01	0.10	562	550	546	544	538	532
2.01	1.00	180	176	174	172	172	168
2.5	0.01	945	299	200	166	84	49
2.5	0.10	299	84	57	49	32	23
2.5	1.00	84	32	25	23	17	13
3.0	0.01	2435	241	96	70	31	19
3.0	0.10	771	70	38	31	19	13
3.0	1.00	241	31	21	19	13	11
4.0	0.01	21075	190	50	35	17	11
4.0	0.10	6667	67	29	23	13	9
4.0	1.00	2107	35	19	17	11	9

evaluation becomes slow and inaccurate.

The results of this section and the previous are invariant under scale transformation of the distribution. The terms $j_{\max}$ and $k_{\max}$ and other results on accuracy and computational complexity depend on y, μ and p only through p, μ^{2-p}/ϕ and y^{2-p}/ϕ, all of which are invariant under re-scaling of the distribution.

6 Maximum likelihood estimation

This section considers maximum likelihood estimation of the parameters of a Tweedie model, especially ϕ and p. The estimation of ϕ and p has been considered previously by Smyth (1996) and Gilchrist and Drinkwater (1999) for the case $1 < p < 2$.

Our interest is EDMs is motivated by their applications to generalized linear models. A generalized linear model can be defined as follows. Independent responses $Y_1, \ldots, Y_n$ are observed such that

$$Y_i \sim \mathrm{ED}(\mu_i, \phi_i/w_i)$$

where the w_i are known prior weights. The means μ_i are related to linear predictors through a known monotonic link function g,

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of unknown regression parameters. Let q be the dimension of $\boldsymbol{\beta}$. To avoid unnecessary complications, we assume here that the design matrix X with rows $\mathbf{x}^T$ has full column rank.

For p fixed, the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be computed using the well known iteratively reweighted least squares algorithm proposed by Nelder and Wedderburn (1972). This iteration uses working weights given by

$$\frac{w_i}{V(\mu_i)\dot{g}(\mu_i)^2}$$

where in our case $V(\mu_i) = \mu_i^p$. Knowledge of ϕ is not required to compute $\hat{\boldsymbol{\beta}}$.

In the normal and inverse Gaussian cases, the maximum likelihood estimate of ϕ is the mean-deviance estimator

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} w_i d(y_i, \hat{\mu}_i)$$

where $d(\cdot,\cdot)$ is the unit deviance. In the gamma case, $\hat{\phi}$ can be found using the results given by Smyth (1989). In other cases, the unit deviances are not sufficient for ϕ and the maximum likelihood estimator of ϕ must be computed iteratively from the full data. In the next section of this paper we discuss computation of the derivatives $\partial \log f/\partial\phi$ and $\partial^2 \log f/\partial\phi^2$ in order to facilitate estimation of $\hat{\phi}$ using Newton-type methods. In our implementation of maximum likelihood estimation for ϕ we actually use a BFGS-type quasi-Newton optimizer which uses first but not second derivatives of the densities with respect to ϕ (Byrd et al, 1995). Higher order derivatives for ϕ are slightly more difficult to compute accurately than the likelihood function itself and we have not found the second derivatives to result in a worthwhile speeding of the convergence of the algorithm. Nevertheless we give them here for completeness. Gilchrist and Drinkwater (1999) have developed an alternative iterative scheme in which derivatives are evaluated with respect to $1/\phi$ rather than ϕ. Smyth (1996) used the derivative-free Nelder-Mead algorithm to estimate ϕ and p.

Estimation of p is a more difficult problem than estimating $\boldsymbol{\beta}$ or ϕ. Most authors using Tweedie densities have taken p to be specified apriori. Jørgensen (1987), in analysing the amount of money spent by Amazonian peasants hiring outside labour, chooses $p = 1.75$ and explicitly states the choice is somewhat arbitrary. Likewise, Nelder (1994) arbitrarily sets $p = 1.5$ when analysing the time spent splicing cables. Given the ability to compute the maximum likelihood estimator $\hat{\phi}$ conditional on p, the maximum likelihood estimate of p and an approximate confidence interval can obtained by evaluating the profile likelihood for p on a grid of values followed by a univariate optimization. A similar approach was used by Gilchrist and Drinkwater (1999) and Smyth (1996) to estimate p. A special problem that arises in the estimation of p is the possibility that recording of the responses y_i to a limited number of decimal places can cause the profile likelihood for p to be bimodal with a spurious maximum at $p = 1$. This potential problem is discussed further in Section 8.

It is well known that maximum likelihood estimators of variances tend to be biased down if the number of parameters in the mean model is large. This affects the estimation of ϕ and p if q is not very small compared to n. The maximum likelihood estimator of ϕ tends to underestimate ϕ while the maximum likelihood estimator of p may be biased up or down depending on whether the fitted values $\hat{\mu}_i$ are greater or less than one. For

this reason it is of interest to consider also modified profile estimators of ϕ and p. Since the coefficients $\boldsymbol{\beta}$ are orthogonal to the variance parameters ϕ and p, approximately unbiased estimators of ϕ and p can be obtained by maximizing with respect to ϕ and p the adjusted profile likelihood (Cox and Reid, 1987), which in this case is

$$\ell(\mathbf{y}; \hat{\boldsymbol{\beta}}, \phi, p) + \frac{q}{2} \log \phi - \frac{1}{2} |\log X^T D X|.$$

Here ℓ is the log-likelihood function, $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator for $\boldsymbol{\beta}$ conditional on ϕ and p, and D is the diagonal matrix of working weights from the generalized linear model evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Computing modified profile estimators is a straightforward extension of the methods that we develop in this paper.

The most commonly used estimators for ϕ in generalized linear models are the mean deviance estimator

$$\tilde{\phi} = \frac{1}{n-q} \sum_{i=1}^{n} w_i d(y_i, \hat{\mu}_i)$$

and the Pearson estimator

$$\bar{\phi} = \frac{1}{n-q} \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

The Pearson estimator is approximately unbiased but is more variable than estimators based on the deviances or the likelihood. The mean deviance estimator coincides with the modified profile estimator of ϕ in the normal and inverse-Gaussian cases. For other values of p, $\tilde{\phi}$ is an biased estimator of ϕ. The size of this bias has been investigated by Dunn (2001).

7 Derivatives with respect to ϕ

Series expansions were given for evaluating the Tweedie density functions in Section 4. Similar series expressions are now used to evaluate derivatives of the density with respect to ϕ.

7.1 The case $1 < p < 2$

Differentiating the log-density with respect to ϕ gives

$$\frac{\partial \log f}{\partial \phi} = \begin{cases} \dfrac{\mu^{2-p}}{\phi^2(2-p)} & \text{for } Y = 0 \\ \dfrac{y\mu^{1-p}}{\phi^2(p-1)} + \dfrac{\mu^{2-p}}{\phi^2(2-p)} + \dfrac{\partial W/\partial \phi}{W} & \text{for } Y > 0 \end{cases} \tag{14}$$

where $W = \sum_{j=1}^{\infty} W_j$ for W_j given in (2).

The evaluation of the derivative $\partial W/\partial \phi$. is similar to the evaluation of the actual series $W = \sum_{j=1}^{\infty} W_j$ itself. Differentiating the W_j terms shows

$$\frac{\partial W}{\partial \phi} = \left(\frac{\alpha - 1}{\phi}\right) \sum_{j=1}^{\infty} j W_j. \tag{15}$$

It can be deduced from (6) that

$$\frac{\partial \log(jW_j)}{\partial j} \approx \log z - \log j + \alpha \log(-\alpha j).$$

This is exactly the expression used to solve for $j_{\max}$ for the series $\sum_j W_j$, since in (6) the $1/j$ term was ignored for large j. Thus, the value of j at which the series terms in (15) reach a maximum occurs is approximated by

$$j_{\max} = \frac{y^{2-p}}{\phi(2-p)}$$

as in the series $W = \sum_j W_j$ itself. The actual implementation is very similar to that of the density. Simply, search either side of the value of $j_{\max}$ to find the values of j over which to sum, and then perform the summation on the exact terms.

The maximum value of the terms jW_j can also be deduced from (8) to be

$$\log(jW)_{\max} \approx j_{\max}(1-\alpha) - \log(2\pi) - \frac{1}{2}\log(-\alpha). \tag{16}$$

In general, mth order derivatives with respect to ϕ require computation of $\sum_{j=1}^{\infty} j^{\nu} W_j$ for $\nu = 1, \ldots, m$. These can all be computed using the ideas presented above. The number of terms required generally increases slightly with increasing m.

7.2 The case $p > 2$

The details for this case are very similar to the case $1 < p < 2$. It becomes necessary to evaluate $\partial V/\partial\phi = \sum_{k=1}^{\infty} \partial V_k/\partial\phi$. Differentiating the terms V_k with respect to ϕ gives

$$\frac{\partial V}{\partial \phi} = \left(\frac{\alpha - 1}{\phi}\right) \sum_{k=1}^{\infty} k V_k.$$

Proceeding as before, the maximum occurs at

$$k_{\max} \approx \frac{\mu^{2-p}}{\phi(p-2)},$$

as with the series $\sum_k V_k$ itself.

The procedure is then the same: search on either side of this maximum to find where the series terms are negligible relative to the maximum value of the kV_k terms. Then the series can be summed over these values of k that contribute to the sum.

The maximum value of $kV_{\text{env}}(k)$ can be deduced from (13) to be

$$\log(kV_{\max}) \approx (1-\alpha)k_{\max} + \frac{1}{2}\log\alpha + \log k_{\max}. \tag{17}$$

In general, nth order derivatives require computing $\sum_{k=1}^{\infty} k^{\nu} V_k$ for $\nu = 1, \ldots, m$. These can all be computed using the ideas presented above. The number of terms required generally increases slightly with increasing m.

7.3 Selection of terms

In implementing the series expansions of $\sum_j jW_j$ and $\sum_k kV_k$, the numerical aspects of the implementation are similar to that of the density. The upper and lower limits of the index of summation are found by determining when the individual summation terms, jW_j or kV_k, are such that $\log jW_j < \epsilon \log jW_{\max}$ (or $\log kV_k < \epsilon k \log V_{\max}$) where $\epsilon = \exp(-37) \approx 8 \times 10^{-17}$.

To reduce the number of situations where overflow may occur, the log-scale is used in calculations. In practice (where the case $1 < p < 2$ is used as an example, but the same procedures follow for the case $p > 2$), the logarithm of the terms to be summed is determined, $\log jW_j = jz + \log j - \log \Gamma(1 + j) - \log \Gamma(\alpha j)$, and then the largest value of this quantity, say $\log(jW)_{\max}$, is found. Then the jW_j terms less this maximum value are used to avoid overflow, defining $\log jW'_j = \log jW_j - \log(jW_{\max})$. Then, the summation itself is reconstructed using $\sum jW = \sum jW'_j \exp\{\log(jW_{\max})\}$.

7.4 Accuracy and limitations

As with the density itself, it is difficult to make definitive statements about the accuracy of the algorithms since exact values are not generally available. However comparisons can be made with the case $p = 3$, the inverse Gaussian distribution, for which

$$\frac{\partial \log f}{\partial \phi} = -\frac{1}{2\phi} + \frac{(y-\mu)^2}{2\phi^2\mu^2 y}. \tag{18}$$

The comparison is made in Table 4. The relative accuracy of the series expansion is excellent again for $y > 0.1$ while the effects of subtractive cancellation become apparent for small y. The number of terms necessary for evaluating $\sum_{j=1}^{\infty} W_j$ and $\sum_{j=1}^{\infty} jW_j$ in this example are almost always the same. As with the density evaluation, the accuracy is poor and the number of terms needed increases without bound for y near 0. The accuracy is excellent for $y \geq 0.50$.

8 Multimodal densities

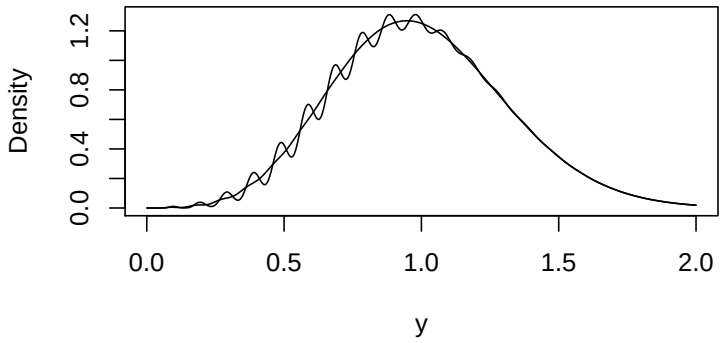
In this section we derive a condition for the Tweedie density function to be unimodal. We also address a technical complication which arises from the fact that responses y_i are generally recorded to only a limited precision, for example to a limited number of decimal places.

When p is close to one, the density $f_p(y; \mu, \phi)$ is multimodal, reflecting the fact that the Poisson limit at $p = 1$ is a discrete distribution. For p very near one, the random variables X_i defined in Section 3 have very small standard deviations and the density of Y has multiple modes corresponding to distinct values of the Poisson count N. The modes occur at the values $\mathrm{E}(\sum_{i=1}^{n} X_i) = -n\alpha\gamma = n(2-p)\phi\mu^{p-1}$ for $n = 1, 2, \ldots$. The density will have multiple modes near $y = -n\alpha\gamma$ if the standard deviation of $\sum_{i=1}^{n} X_i$ is much smaller than the spacing $-\alpha\gamma$ between the successive modes. The ratio of the

Table 4: Comparing the exact and series first derivatives for the inverse Gaussian distribution, with $p = 3$, $\mu = 1.4$ and $\phi = 0.74$. The comparison is excellent for $y > 0.1$ but fails for small y.

y	**Exact first derivative**	**Series first derivative**	**Relative error**	**No. Terms: for** $\sum V_k$	**for** $\sum kV_k$
0.001	911.095634	−927.644712	2	895	895
0.002	454.558482	−509.162138	2	633	633
0.005	180.637308	156.319402	0.1	399	399
0.01	89.332113	85.265128	0.05	280	280
0.05	16.304729	16.356116	−0.003	112	112
0.10	7.197269	7.197269	1×10^{-9}	79	79
0.50	0.079009	0.079009	3×10^{-14}	41	41
1.00	−0.601139	−0.601139	-1×10^{-15}	33	33
2.00	−0.591822	−0.591822	-6×10^{-16}	27	27
3.00	−0.278146	−0.278146	-1×10^{-15}	24	24
4.00	0.111619	0.111619	3×10^{-15}	22	22
5.00	0.531820	0.531820	4×10^{-16}	21	21
6.00	0.967239	0.967239	-3×10^{-16}	20	20
7.00	1.41153	1.411353	2×10^{-16}	20	20
8.00	1.860903	1.860903	1×10^{-16}	19	19
9.00	2.314076	2.314076	2×10^{-16}	19	19
10.00	2.769786	2.769786	0	18	18
15.00	5.068624	5.068624	0	17	17
20.00	7.382679	7.382679	4×10^{-16}	16	16

Figure 1: Densities of Tweedie distributions with $\mu = 1$ and $\phi = 0.1$. The density is multimodal with $p = 1.02$ but unimodal with $p = 1.05$.



standard deviation to the spacing is $(-n\alpha\gamma^2)^{1/2}/(-\alpha\gamma) = (-n\alpha)^{1/2}$. Inserting for n the maximum of $\mathrm{E}(N) = \lambda$ and one, a condition for the density to be multimodal is that

$$\left(-\frac{\lambda \vee 1}{\alpha}\right)^{1/2} = \left[\left\{\frac{\mu^{2-p}}{\phi(2-p)} \vee 1\right\}\frac{p-1}{2-p}\right]^{1/2}$$

be less than about 0.5. Figure 1 shows then Tweedie density for $\mu = 1$ and $\phi = 0.1$. For p small the density will have modes at multiples of 0.1. For $p = 1.02$ we have $(-\lambda/\alpha)^{1/2} = 0.46$ and the density is is multimodal. For $p = 1.05$ we have $(-\lambda/\alpha)^{1/2} = 0.74$ and the multiple modes are gone. A conservative condition to ensure that the density be unimodal for most values of μ and ϕ is $(-1/\alpha)^{1/2} > 0.5$, i.e., $p > 1.2$.

The multimodality of the density for small p causes a technical problem for maximum likelihood estimation with rounded data. If the observations y_i are rounded to d decimal places say, then the likelihood is unbounded as $p \downarrow 1$ and $\phi \downarrow 10^{-d}$. This limit models the data as discrete on the lattice $n10^{-d}$ for $n = 0, 1, \ldots$. If the y_i are rounded observations sampled from $\mathrm{ED}_p(\mu, \phi)$ with $p > 1$, then the likelihood will usually have two local maxima, one infinite at $p = 1$ and one finite for $p > 1$, and it is usually the latter which is required. This phenomenon has been noted by Jørgensen and Paes de Souza (1994), Gilchrist and Drinkwater (1999) and by Burridge in the discussion of Jørgensen (1987). Gilchrist and Drinkwater (1999) constrain $p > 1.1$ while Jørgensen and Paes de Souza (1994) similarly try to avoid small values of p. We prefer to avoid the spurious singularity in the likelihood by requiring that the density not have multiple modes corresponding to the rounding accuracy. This can be achieved by constraining

$$-(\lambda \vee 1)\,\alpha\gamma^2 = \left\{\frac{\mu^{(2-p)}}{\phi(2-p)} \vee 1\right\}\frac{(2-p)(p-1)\phi^2}{\mu^{2-2p}} > 10^{-2d}.$$

In most relevant cases we will have $\lambda > 1$ so the above condition is

$$-\lambda\alpha\gamma^2 = (p-1)\phi\mu^p = (p-1)\mathrm{var}(Y) > 10^{-2d},$$

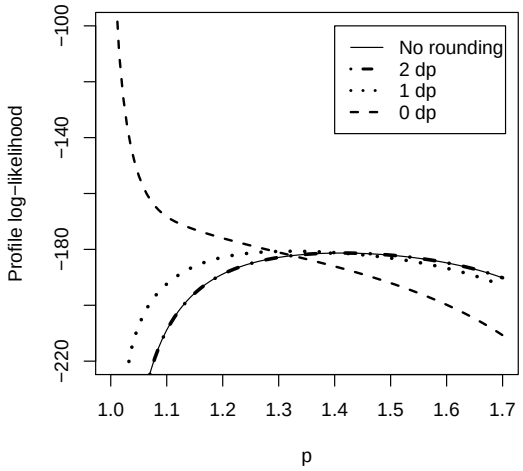
i.e.,

$$p > 1 + \frac{10^{-2d}}{\mathrm{var}(Y)}.$$

In this paper we do not apply such a constraint explicitly, but simply inspect the profile likelihood as a function of p.

As an example, 100 random deviates were generated from the Tweedie distribution with $\mu = 2$, $\phi = 1$ and $p = 1.5$. The maximum likelihood estimate of p is 1.41. Rounding the data to two decimal places introduces an infinite peak in the likelihood at $p = 1$, although the likelihood is effectively unchanged for larger values of p; see Figure 2. Rounding the data to one decimal place decreases the location of local peak to 1.3. Rounding the data to the nearest integer loses the peak of interest altogether — the likelihood is now monotonically decreasing for $p > 1$.

Figure 2: Profile likelihood for 100 random deviates from the Tweedie distribution with $\mu = 2$, $\phi = 1$ and $p = 1.5$. The maximum likelihood estimate is $\hat{p} = 1.41$. Rounding the data decreases the estimate value of p and introduces a spurious likelihood peak at $p = 1$.



9 Examples

Enabling likelihood calculations for Tweedie models with arbitary p opens up many potential applications. Two small data examples are given here to hint at possibilities, the first featuring data with positive support where we expect $p > 2$ and the second featuring mixed data with mass at zero where we expect $1 < p < 2$.

9.1 Sensitivity to poison

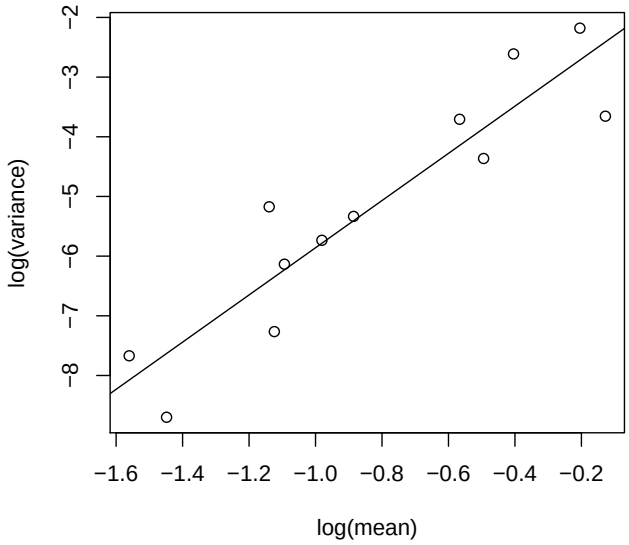
Box and Cox (1964) give the results of a 3×4 factorial experiment in which the survival times of animals were recorded after exposure to poisons. There are two factors: the type of poison with three levels and the type of treatment with four levels. Each of the twelve factor combinations was applied to four randomly allocated animals for a total of 48 observations. The data are available from `http://www.statsci.org/data/general/poison.html`. Box and Cox defined what is now known as the Box-Cox transformation,

$$y_\omega = \begin{cases} (y^\omega - 1)/\omega & \omega \neq 0 \\ \log y & \omega = 0. \end{cases}$$

They analysed the data assuming the reponses to be normal after a reciprocal transformation, i.e., assuming the y_ω to be normal with $\omega = -1$.

A plot of the log of sample variances against the log of the sample means for each poison–treatment combination is shown in Figure 3. The plot is almost linear with slope 3.95 showing that a power mean-variance relationship $V(\mu) = \mu^p$ with p close to 4 is appropriate for this data. There is a rough correspondence between Box-Cox

Figure 3: Log-sample variances versus log-sample means for the poison data. The plot is almost linear with slope close to 4.



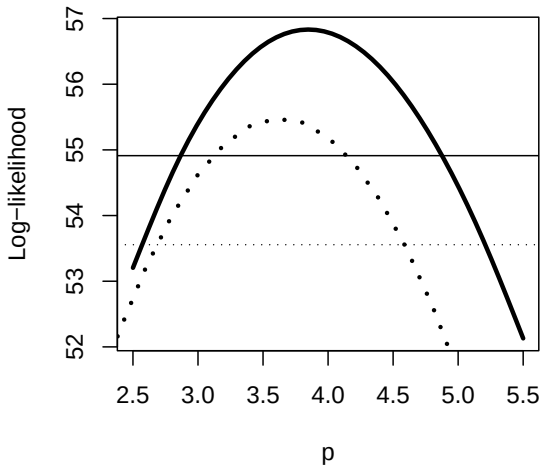
transformations and power variance functions in that the Box-Cox transformation with $\omega = 1 - p/2$ is to first order the variance stabilizing transformation for $V(\mu) = \mu^p$. The reciprocal transformation is the variance-stabilizing transformation for $V(\mu) = \mu^4$, so our observation that $p \approx 4$ for this data matches the transformation chosen by Box and Cox (1964).

Rather than having to transform the response variable, we entertain the possibility of modeling the survival times directly on their original scale using a Tweedie generalized linear model. Comparison of likelihoods gives a means to compare the fit of the Tweedie and Box-Cox models for this data. For each value of p we fit a generalized linear model with variance function $V(\mu) = \mu^p$ allowing for interactions between the poison and treatment explanatory factors. Similarly for each p we consider a normal linear interaction model for the transformed responses y_ω with $\omega = 1 - p/2$. Figure 4 plots Tweedie and Box-Cox profile likelihoods on the same scale and shows that the Tweedie likelihood is somewhat higher for any given value of p. The Tweedie log-likelihood peaks at 56.8 while the maximum Box-Cox log-likelihood is 55.5. The Tweedie and Box-Cox models are not nested hypotheses so comparing the log-likelihoods in this way does not constitute a formal hypothesis test comparing the two models without further work (Cox 1961, 1962; Pereira, 1977), Nevertheless, this does show that the Tweedie model is at least as acceptable as the transformation model. The Tweedie log-likelihood is defined here as

$$\ell_{\mathrm{Tw}}(\mathbf{y}; \boldsymbol{\beta}, \phi, p) = \sum_{i=1}^{n} \log f_p(y; \mu, \phi),$$

where $f_p(y; \mu, \phi)$ is the Tweedie distribution density function, evaluated using the al-

Figure 4: Poison data. Profile log-likelihoods for p under the Tweedie (solid lines) and Box-Cox models (dotted lines). Horizontal lines indicate 95% likelihood regions. The peaks occur at similar values of p for the Tweedie and Box-Cox models but the Tweedie likelihood is higher.



gorithms in this paper. Under the Box-Cox model, the log-likelihood function is that assuming $y_{\omega,i} \sim N(\mu, \sigma^2)$ for $\omega = 1 - p/2$, i.e.,

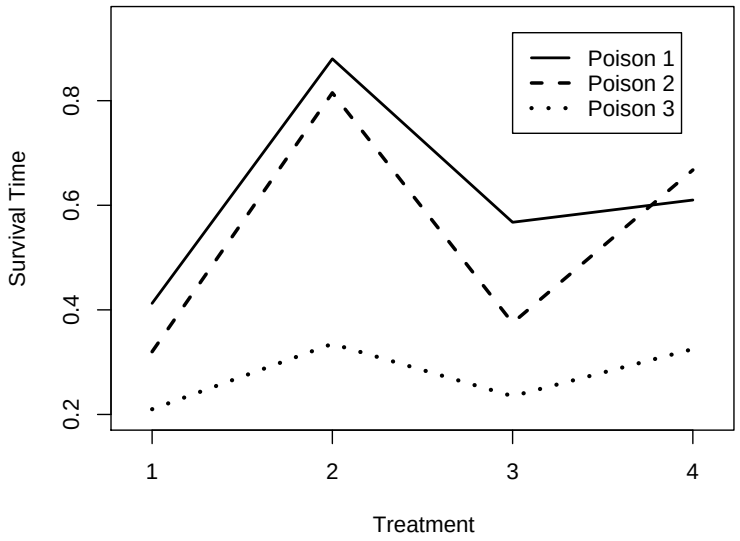
$$\ell_{\mathrm{BC}}(\mathbf{y}; \boldsymbol{\beta}, \sigma^2, p) = -\frac{1}{2} \sum_{i=1}^{n} \left\{ \log(2\pi\sigma^2) + \left(\frac{y_{\omega,i} - \mu_i}{\sigma} \right)^2 + p \log y_i \right\}.$$

In Figure 4 the log-likelihoods have been maximised over $\boldsymbol{\beta}$, ϕ and σ^2 for given values of p. The maximum likelihood estimates for the Tweedie parameters are $\hat{p} = 3.85$ and $\hat{\phi} = 0.151$. The 95% confidence interval for p computed from the profile likelihood is $(2.87, 4.88)$. For the Box-Cox model, the maximum likelihood estimate of ω is equivalent to $\hat{p} = 3.64$.

Now consider the effects of the explanatory factors on survival time. An appealing advantage of the Tweedie model over the Box-Cox transformation is that it makes sense to summarize responses using sample means. Under the Tweedie model, sample means follow the same distribution as the individual observations except that the dispersion parameter is divided by the sample size, i.e., $Y_i \sim \mathrm{ED}_p(\mu, \phi)$ implies $\bar{Y} \sim \mathrm{ED}_p(\mu, \phi/n)$ where n is the sample size. (This generalizes the corresponding result for the normal distribution.) Figure 5 plots the mean survival time for each poison–treatment group. The effects of poison and treatment appear approximately multiplicative, except that the relative changes between the treatments are somewhat less for poison 3, which also has the shortest survival times.

Finally we consider link functions for the generalized linear model, and it is natural to consider link functions in the power family, $g(\mu) = \mu^r$. Another advantage of the Tweedie model approach is that the variance and link functions can be chosen separately. We assume a power variance function. For convenience, we fix $p = 3.85$ from this point

Figure 5: Poison data. Mean survival time by treatment and poison.



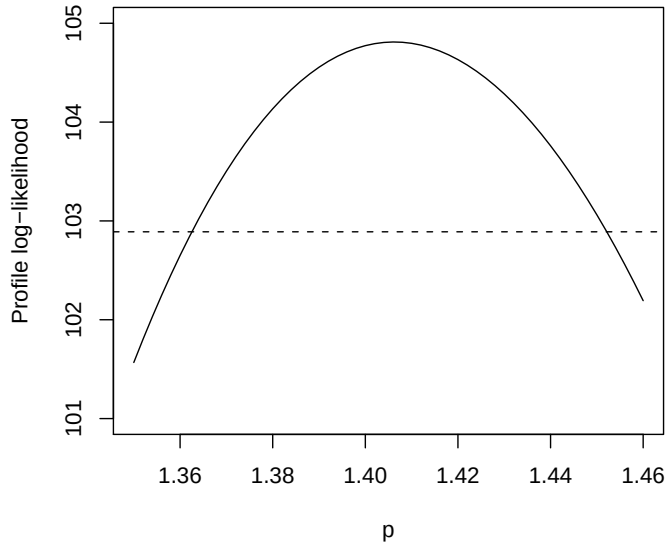
so that covariate and link model selection can be undertaken using standard generalized linear model methods. Although convenient, this is not strictly necessary. In principle, p could be left unknown in which case all model selection would be undertaken comparing likelihoods rather than comparing deviances.

The residual mean deviance allowing for poison–treatment interaction is 0.199 on 36 degrees of freedom, unchanged by the choice of link function. The mean deviance for interaction is on 6 degrees of freedom and is low and non-significant for a range of link functions. It is 0.28 for the reciprocal link $r = -1$ and 0.33 for the log-link $r = 0$, with a minimum of 0.21 at $r = -0.6$. Any of these link functions could be used, as the interaction remains non-significant, but the log-link seems the most intuitively appealing as it treats the poison and treatment effects as multiplicative. The slight deviation from additivity on the log-scale could be explained by an initial survival period before the poison starts to take effect. If a small constant is subtracted from all the survival times, the times become very closely additive on the log-scale. Using the log-link, the mean deviances for the poison and treatment main effects are significant at 14.6 and 5.66 respectively, both highly significant according to the usual generalized linear model methods.

9.2 Root length density of apple trees

The root length density data of de Silva *et al.* (1999), also available from `http://-www.statsci.org/data/oz/fineroot.html`, provides an example of data with exact zeros. The data concerns the underground root system of eight separate apple trees. Three different root stocks are considered (Mark, MM106 and M26) and two plant spacing (4×2 metres and 5×3 metres). Soil core sampling units taken were classified

Figure 6: Root length density data. Profile log-likelihood for p. The peak is at $p = 1.406$ with 95% likelihood region from 1.363 to 1.452.



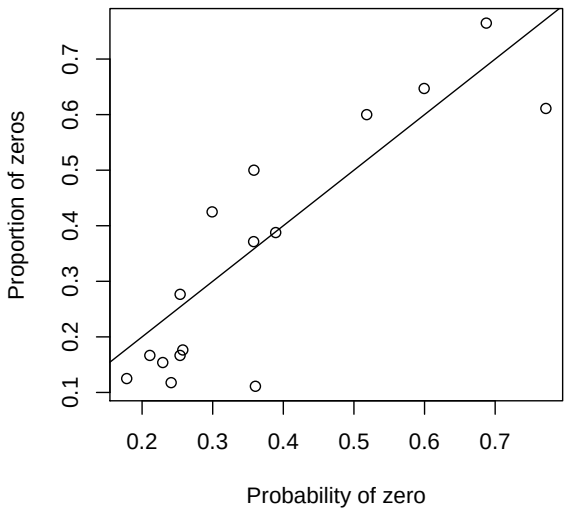
as belonging to an inner or outer zone relative to each plant. The response variable is the density of fine roots, also called the root length density, RLD (in cm/cm^3), which can have zeros as well as continuous positive values. There are 511 observations, of which 193 or 38% have a zero response.

The design is not a full factorial design: plants 1 and 2 are Mark root stock at 5×3 spacing; plants 3 and 4 are Mark root stock at 4×2 spacing; plants 5 and 6 are MM106 root stock at 5×3 spacing; and plants 7 and 8 are M26 root stock at 4×2 spacing. The Mark root stock is therefore tested at both plant spacings but the MM106 only at 5×3 and M26 only at 4×2.

No transformation to normality is likely to be successful for this data because of the large number of exact zeros, so we investigate a Tweedie generalized linear model with $1 < p < 2$. Since zone is the only variable which varies within plant, the mean structure of the responses is fully described by a plant-zone model allowing for plant by zone interactions. For each p a generalized linear model is fitted to the data allowing for plant-zone effects and the profile log-likelihood function is shown in Figure 6. The maximum likelihood estimates for p and ϕ are 1.406 and 0.3118 respectively. The 95% likelihood confidence interval for p is tight, from 1.363 to 1.452.

A primary issue when modeling data with exact zeros is to accurately model the probability of exact zeros. One question is whether the occurence of zeros needs to be modeled separately to the distribution of the positive values, perhaps through a logistic regression model. Figure 7 plots the observed proportion of zeros for each plant and each zone versus the probability of zeros $\exp(-\hat{\lambda})$ derived from the maximum likelihood Tweedie model. Although the proportion of zeros is somewhat less than expected at the

Figure 7: Root length density data. Observed versus predicted proportion of exact zeros in each zone of each plant. The line of equality is also plotted.



lowest probabilities, the model appears to predict well the observed pattern of zeros. The Tweedie approach of modeling the zeros and the positive observations together appears to be perfectly adequate here.

We can now use generalized linear model methods to answer some of the experimental questions of interest. The parameter p is orthogonal to μ and ϕ in the Tweedie model, implying that the estimator of p changes relatively slowly as $\hat{\mu}$ and $\hat{\phi}$ change. As a first approximation therefore, we can hold p fixed in our analyses in which case the Tweedie model reduces to an ordinary generalized linear model with a power variance function. By analogy with normal theory, an approximately unbiased estimator of ϕ is obtained by scaling the maximum likelihood estimator up by the ratio of the number of observations to the residual degrees of freedom,

$$\tilde{\phi} = \frac{n}{n-q}\hat{\phi} = \frac{511}{511-16}0.3118 = 0.3219$$

This is slightly lower than the residual mean deviance for the plant-zone interaction model, which is 0.360 on 495 degrees of freedom.

The mean RLD is greater in the inner than the outer zone for every plant, the difference being greater for some plants than others. Let us first test the significance of these differences. It seems natural to use a log-link to preserve positivity of μ for any linear predictor, although in fact an ordinary linear additive fit is just as good as the log-linear additive fit for this data. The mean deviance for differences between plants is 2.8 on 7 df. The sequential mean deviances for Zone and Plant×Zone interactions are 16.6 on 1 df and 1.4 on 7 df respectively. All of these give highly significant F-statistics when compared to $\tilde{\phi}$ on 495 degrees of freedom. We conclude that the inner zone gives greater RLD than the outer zone, but that this effect differs between plants.

Table 5: Root length density data. Mean RLD by root stock and zone.

	Zone	
Stock	Inner	Outer
M26	0.123	0.079
Mark	0.078	0.015
MM106	0.115	0.076

This analysis can be done in S-Plus or R by

```
fit <- glm(RLD Plant*Zone,family=tweedie(var.power=1.406,link.power=0))
anova(fit,test="F",dispersion=0.3219)
```

using our function `tweedie()`.

Now consider the between-plant factors of Stock and Spacing. The main effects for Stock and Zone are highly significant ($p = 2 \times 10^{-10}$) and ($p = 2 \times 10^{-13}$) as is the Stock by Zone interaction ($p = 2 \times 10^{-5}$). After allowing for these effects, there is no significant effect for Spacing or Plant by Zone interaction and only a marginally significant main effect for Plant ($p = 0.02$). A table of mean RLD by Stock and Zone allows us to interpret the effects for Stock and Zone (Table 5). We see that stocks M26 and MM106 are very similar and have greater RLD that Mark stock. The Inner zone shows greater RLD than the outer for all stocks, however the relative difference is greatest for Mark stock. A more complete treatment of this data might include fitting a generalized linear mixed model with plant as a random effect. Such an analysis is beyond the scope of this paper and would be unlikely to change our qualitative conclusions since the differences between plants are not too large.

We finish this example by comparing maximum likelihood with another strategy which has been suggested for estimating parameters in a generalized linear model variance function, namely the extended quasi-likelihood (EQL) proposed by Nelder and Pregibon (1987). For exponential dispersion models, EQL is equivalent to using the saddlepoint approximation to the density

$$f(y;\mu,\phi) \approx [2\pi\phi V(y)]^{-1/2} \exp\left\{-\frac{1}{2\phi} d(y,\mu)\right\} \{1 + O(\phi)\}$$

(Jørgensen, 1997; Smyth and Verbyla, 1999). The saddlepoint approximation as given above is not defined at zero for power variance functions but Nelder and Pregibon (1987) suggest using $V(y + 1/6)$ in place of $V(y)$ to allow evaluation at $y = 0$. While this strategy has proved effective for count data, it seems inappropriate here with a partially continuous response and no quantum gap between exact zeros and positive obvervations. In the root length density data, some of the positive responses are as small as 0.003008. We have experimented with estimating ϕ and p from the EQL using $V(y + c)$ in place of $V(y)$ and with various values of c. The estimators turn out to be extremely sensitive to the choice of c — the estimated value for p can be made to be

anything from near 0 to nearly 2 by varying c between 0 and 1/6. The most sensible estimates for ϕ or p are obtained when c is somewhat smaller than the smallest positive value of y, say $c = 0.003/3$ or $c = 0.003/6$. It seems inappropriate though to make c depend on the observed data, so we conclude that EQL is not well suited to this type of mixed data with exact zeros. The Tweedie model approach is invariant to a change in the unit of measurement, apart from an obvious rescaling of the μ and ϕ, while transformations which involve $y + c$ for a pre-specified constant c are not.

10 Conclusion

This paper has considered the numerical computation of probability densities for Tweedie models. The motivation has been to extend the range of response distributions for which generalized linear model-type analyses can be done. Apart from the appealing power mean-variance relationship, Tweedie EDMs are natural candidates for modeling quantitative data on arbitrary measurement scales because they are the only EDMs which are closed under scale transformations, i.e., changes in the unit of measurement. The ability to compute densities facilitates maximum likelihood estimation of the variance function and dispersion parameter as well as diagnostic checking of the response distribution.

Numerically efficient strategies have been developed for evaluating series expansions for the density functions and their derivatives when $p > 1$. The series are evaluated only for those terms that make a contribution to the final result. The method shows excellent relative accuracy for a wide range of parameter values. However we have also shown that the number of terms necessary can increase without bound for certain parameter values and, even more serious, that full accuracy is not obtainable in floating point arithmetic for certain parameter values when $p > 2$ regardless of the number of terms evaluated.

The Tweedie models have closed form characteristic functions. The series expansions for the density functions arise from an analytic approach to the inversion integral for the density function in terms of the characteristic function. Dunn (2001) has also investigated a more direct numerical approach to inverting the characteristic function based on numerical integration methods for oscillating functions. The numeric inversion method provides a means to evaluate the Tweedie densities when the series approach fails, and this work will be published elsewhere.

The saddlepoint approximation has been shown to perform poorly for the root length density data of Section 9.2. This agrees with the rule of thumb given in Smyth and Verbyla (1999) for judging whether the saddlepoint approximation will be adequate, which rules out saddlepoint approximations for data sets with exact zeros. On the other hand, the saddlepoint approximation is judged adequate for the poison data of Section 9.1.

The software package "tweedie" has been developed for the R (R Development Core Team 2004) programming environment to implement the methods developed in this paper. The package includes functions for the Tweedie density, distribution function, quantile function and random number generation using both the series expansion and numeric inversion methods. It is available from `http://www.sci.usq.edu.au/staff/-`

`dunn/twhtml/home.html`. The generalized linear model family function `tweedie()`, which does not in itself require likelihood calculations, is in the "statmod" package, also written by the authors of this paper, available from `http://www.statsci.org/r/`.

11 Acknowledgements

The authors wish to thank Nihal de Silva of HortResearch, New Zealand, for the data used in Section 9.2.

References

Aalen, O. O. 1992. Modelling heterogeneity in survival analysis by the compound Poisson distribution. Ann. Appl. Probab. 2: 951–972.

Abramowitz, M. and Stegun, I. A. (eds) 1965. A Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. Dover Publications, New York.

Bar-Lev, S. K. and Enis, P. 1986. Reproducibility and natural exponential families with power variance functions. Ann. Statist. 14: 1507–1522.

Bar-Lev, S. K. and Stramer, O. 1987. Characterizations of natural exponential families with power variance functions by zero regression properties. Probab. Theory Related Fields 76: 509–522.

Box, G. E. P. and Cox, D. R. 1964. An analysis of transformations (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 26: 211–252.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. SIAM J. Scientific Computing. 16: 1190–1208.

Cox, D. R. 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 105–123. University of California Press, Berkeley, CA.

Cox, D. R. 1962. Further results on tests of separate families of hypotheses. Journal of the Royal Statistical Society B 24: 406-424.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. J. R. Statist. Soc. B, 49: 1–39.

Davidian, M. 1990. Estimation of variance functions in assays with possible unequal replication and nonnormal data. Biometrika 77: 43–54.

Davidian, M., Carroll, R. J. and Smith, W. 1988. Variance functions and the minimum detectable concentration in assays. Biometrika 75: 549–556.

de Silva H. N., Hall, A. J., Tustin, D. S. and Gandar, P. W. 1999. Analysis of distribution of root length density of apple trees on different dwarfing rootstocks. Ann. Botany, 83: 335–345.

Dunn, P. K. 2001. Likelihood-Based Inference for Tweedie Exponential Dispersion Models. Unpublished PhD Thesis, University of Queensland.

Dunn, P. K. and Smyth, G. K. 1996. Randomized quantile residuals. J. Comput. Graph. Statist. 5: 236–244.

Feller, W. 1968. An Introduction to Probability Theory and its Applications, Volume I, third edition. John Wiley, New York.

Gilchrist, R. and Drinkwater, D. 1999. Fitting Tweedie models to data with probability of zero responses. Proceedings of the 14th International Workshop on Statistical Modelling, Graz, pp. 207–214.

Haberman, S. and Renshaw, A. E. 1996. Generalized linear models and actuarial science. The Statistician, 45: 407–436.

Haberman, S., and Renshaw, A. E. 1998. Actuarial applications of generalized linear models. In Statistics in Finance, D. J. Hand and S. D. Jacka (eds), Arnold, London.

Hougaard, P. 1986. Survival models for heterogeneous populations derived from stable distributions. Biometrika, 73: 387–396.

Hougaard, P. , Harvald, B. and Holm, N. V. 1992. Measuring the similarities between the lifetimes of adult Danish twins born between 1881–1930. J. Amer. Statist. Assoc. 87: 17–24.

Johnson, N. L. and Kotz, S. 1970. Continuous Univariate Distributions–I, Boston: Houghton Mifflin Company.

Jørgensen, B. 1987. Exponential dispersion models (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 49: 127–162.

Jørgensen, B. 1997. The Theory of Dispersion Models. Chapman and Hall, London.

Jørgensen, B. and Paes de Souza, M. C. 1994. Fitting Tweedie's compound Poisson model to insurance claims data. Scand. Actuar. J. 1: 69–93.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models, second edition. Chapman and Hall, London.

Millenhall, S. J. 1999. A systematic relationship between minimum bias and generalized linear models. 1999 Proceedings of the Casualty Actuarial Society 86: 393–487.

Murphy, K. P., Brockman, M. J., and Lee, P. K. W. (2000). Using generalized linear models to build dynamic pricing systems. Casualty Actuarial Forum, Winter 2000.

Nelder, J. A. 1994. An alternative view of the splicing data. J. Roy. Statist. Soc. Ser. C 43: 469–476.

Nelder, J. A. and Lee, Y. 1992. Likelihood, quasi-likelihood and pseudolikelihood: some comparisons, J. R. Stat. Soc. Ser. B Stat. Methodol. 54: 273–284.

Nelder, J. A. and Pregibon, D. 1987. An extended quasi-likelihood function. Biometrika 74: 221–232.

Nelder, J. A. and Wedderburn, R. W. M. 1972. Generalized linear models. J. Roy. Statist. Soc. Ser. A 135: 370–384.

Pereira, B. de B. 1977. Discriminating among separate models: a bibliography. Internat. Statist. Rev. 45: 163–172.

Perry, J. N. 1981. Taylor's power law for dependence of variance on mean in animal populations. J. Roy. Statist. Soc. Ser. C 30: 254–263.

R Development Core Team 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from `http://www.R-project.org`.

Renshaw, A. E. 1994. Modelling the claims process in the presence of covariates. ASTIN Bulletin 24: 265–286.

Seigel, A. F. 1979. The noncentral chi-squared distribution with zero degrees of freedom and testing for uniformity. Biometrika 66: 381–386.

Seigel, A. F. 1985. Modelling data containing exact zeros using zero degrees of freedom. J. R. Stat. Soc. Ser. B 47: 267–271.

Smyth, G. K. 1989. Generalized linear models with varying dispersion, J. R. Stat. Soc. Ser. B 51: 47–60.

Smyth, G. K. 1996. Regression analysis of quantity data with exact zeros. Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management. Technology Management Centre, University of Queensland, 572–580.

Smyth, G. K. and Verbyla, A. P. 1999. Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics 10: 695–709.

Smyth, G. K., and Jørgensen, B. 2002. Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. ASTIN Bulletin 32: 143–157.

Tweedie, M. C. K. 1984. An index which distinguishes between some important exponential families. Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Indian Statistical Institute, Calcutta, pp. 579–604.

Wright, E. M. 1933. On the coefficients of power series having essential singularities. J. London Math. Soc. 8: 71–79.