Scandinavian Actuarial Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/sact20

Fitting Tweedie's compound poisson model to insurance claims data

Bent Jørgensen & Marta C. Paes De Souza [a]
[a] Centra de Estudos e Pesquisas em Seguros , COPPEAD , Ilha do Fundão, Caixa Postal 68514, Rio de Janeiro RJ , Brazil
Published online: 22 Dec 2011.

To cite this article: Bent Jørgensen & Marta C. Paes De Souza (1994) Fitting Tweedie's compound poisson model to insurance claims data, Scandinavian Actuarial Journal, 1994:1, 69-93, DOI: 10.1080/03461238.1994.10413930

To link to this article: http://dx.doi.org/10.1080/03461238.1994.10413930

Scand. Actuarial J. 1994; 1: 69–93

ORIGINAL ARTICLE

Fitting Tweedie's Compound Poisson Model to Insurance Claims Data

BENT JØRGENSEN and MARTA C. PAES DE SOUZA

Jørgensen B., Paes de Souza MC. Fitting Tweedie's compound Poisson model to insurance claims data. Scand. Actuarial J. 1994; 1: 69–93.

We discuss the estimation and inference problems for the Tweedie compound Poisson process and its application to tarification. For data in the form of the total claim and number of claims for a given time interval and a given exposure, the Tweedie process corresponds to a Poisson process of claims and gamma distributed claim sizes. The model has three parameters, namely the mean claim rate, a dispersion parameter and a shape parameter, and the exposure enters as a weight via the dispersion parameter. The Tweedie process is an exponential dispersion model for fixed value of the shape parameter, and hence regression models for the claim rate may be fitted as in generalized linear models. The shape parameter is estimated by maximum likelihood, and inference is based on the likelihood ratio test, rather than the usual analysis of deviance. A GLIM 4 program for estimation in the model is presented. *Key words: Car insurance, claims data, compound Poisson model, exponential dispersion models, exposure, generalized linear models, GLIM, orthogonal parameters, regression, risk theory, tarification, Tweedie model.*

1. INTRODUCTION

In risk theory, consider a model for the total claim $Z(w)$ corresponding to an exposure w, defined by

$$Z(w) = \sum_{i=1}^{N(w)} Z_i, \tag{1}$$

the sum being zero if $N(w)$, the number of claims, is zero. Here we assume that $N(w), Z_1, Z_2, \ldots$ are independent and the claims Z_i are identically distributed. We define exposure as $w = vt$, where v denotes the value insured and t denotes the time interval during which the value v is exposed to risk. Other definitions of exposure may also be relevant, depending on the context. We shall investigate the *Tweedie model*, corresponding to the special case of (1) in which $N(w)$ is Poisson distributed and the Z_i are gamma distributed. The Tweedie model generalizes the so-called non-central chi-squared distribution with zero degrees of freedom, see Jones (1987), Siegel (1985) and references therein.

Let $Y(w) = Z(w)/w$ be the observed claim rate per unit of exposure, and let $\mu = EY(w)$ denote the mean claim rate, assumed to be the same for all w. We use

the Tweedie model to solve the problem of tarification, that is, to explain μ as a function of independent variables. In particular, we analyze a set of Brazilian data on car insurance, in which the main independent variables are vehicle type and age, and geographic region.

The main problem in connection with tarification is to deal with the fact that the value $Y(w) = 0$ occurs with positive, in fact non-negligible, probability. No completely satisfactory method for dealing with this problem seems to be currently available, in spite of the fact that the compound Poisson model has occupied a central position in risk theory for many years. A second problem is to incorporate the exposure into the analysis in a suitable way, taking into account that the precision of $Y(w)$ as an estimator of μ increases with w.

In particular, there seems to be some confusion concerning the choice of dependent variable in tarification analysis, and often the incidence and severity components are analyzed separately. The simplest suggestion is to analyze observations of $Z(w)$, the total claim, by least-squares regression, ignoring the fact that the distribution of $Z(w)$ is certainly far from being Gaussian, and is unlikely to have constant variance. This is the approach taken by for example Hallin & Ingenbleek (1983). The variable $N(w)$ may be analyzed by contingency table methods, and the results then combined with an analysis of individual claim sizes, as suggested by for example Andrade e Silva (1989). McCullagh & Nelder (1989, pp. 296–300) suggest analyzing the conditional distribution of $Y(w)$ given $N(w)$ by a gamma distribution, with the value of $N(w)$ entering as weight in the analysis, such that in particular zero values of $Y(w)$ are given zero weight. This approach is appropriate if the distribution of individual claim sizes is of interest.

A further possibility, proposed by Coe & Stern (1982) in connection with the analysis of rainfall data, is to analyze the proportion of zeros by logistic regression, and positive values of $Y(w)$ by methods for regression for positive observations. The advantage of this approach is that well-known techniques may be used, but otherwise its justification is mainly empirical.

As in any statistical analysis, the question of choosing the dependent variable is secondary to the more fundamental one of creating a suitable statistical model, which in turn suggests the proper form of the analysis. In the present case we propose to analyze the likelihood based on the joint distribution of $Y(w)$ and $N(w)$. The choice of parameter to be modelled by regression then determines the final form of the analysis. The question is hence to find a suitable form of the compound Poisson model that allows the fitting of appropriate regression models.

The Tweedie model may be parameterized by three parameters, namely the mean claim rate μ, a dispersion parameter σ^2, and a shape parameter α. The marginal distribution of $Y(w)$ is a member of the Tweedie family of exponential dispersion models (Jørgensen, 1987), which allows us to use methods from generalized linear models for fitting regression models for μ, whereas the distribution of $N(w)$ contains information mainly on the shape and dispersion parameters.

If only $Y(w)$, but not $N(w)$, is observed, our approach may be used in connection with the EM-algorithm (Dempster, Laird & Rubin, 1977) to fit the same kind of

regression models for the claim rate, cf. Yáñez Canal (1992). We shall not investigate this approach here, but we mention that one of the effects of not observing $N(w)$ is that less information on the dispersion and shape parameters is available. Otherwise the analysis follows much the same pattern.

The compound Poisson model with gamma distributed claim sizes used here is classical in risk theory, but much of the motivation for the particular form of our analysis comes from exponential dispersion models and the resultant connection with generalized linear models. Our main motivation is to model the claim rate μ by means of a regression model, as appropriate in tarification, taking exposure into account in an appropriate way. The model is based on parametric assumptions, but as illustrated by our analysis of the Brazilian data (Section 5), an analysis of residuals may be used for confirming that the chosen parametric form of the model is appropriate for a given set of data. The fact that the model has three parameters gives it a considerable flexibility, such that it may sometimes fit given data, even if the detailed assumptions related to (1) are not satisfied exactly.

Because we model the claim rate directly, the same estimate of the claim rate results from a fixed claim amount, independently of whether it arose from a single large claim or several small claims. This is correct as long as the claim rate is the quantity of interest, but as mentioned above, the number of claims contains information mainly about the dispersion and shape parameters in our model.

Compared with separate analyses of incidence and severity, along the lines mentioned earlier, our approach provides a considerable simplification, because a single joint analysis of incidence and severity is used. Because the model is based on specific assumptions regarding incidence and severity, separate information about each of these may be extracted from the model. This usually obviates the need for separate analyses of incidence and severity, unless some very special kind of information is needed.

We begin the paper by a detailed derivation of the Tweedie model, in Section 2, after giving some prerequisite results for exponential dispersion models in Section 1.1. We then consider parameter estimation for the model in Section 3 and inference in Section 4. Section 5 presents an analysis of a set of Brazilian data on car insurance.

1.1. Some prerequisites

We now review some basic facts about univariate exponential dispersion models from Jørgensen (1987) needed in the following. Let Y be a random variable with distribution $ED(\mu, \sigma^2)$, defined by the probability density function

$$p(y; \theta, \lambda) = a(\lambda; y) \exp[\lambda\{y\theta - \kappa(\theta)\}], \tag{2}$$

with respect to a suitable σ-finite measure on $\mathscr{R}$, where μ is the mean of (2), $\sigma^2 = 1/\lambda$ is the dispersion parameter, and a and κ are suitable functions. The model (2) is called an *exponential dispersion model*. If we define $\tau(\theta) = \kappa'(\theta)$, then $\mu = \tau(\theta)$ and Var $Y = \sigma^2 V(\mu)$, where $V(\mu) = \tau'(\tau^{-1}(\mu))$ is the variance function.

For a given exponential dispersion model $ED(\mu, \sigma^2)$, we define the corresponding *exponential convolution model* $ED^*(\theta, \lambda)$ by

$$Y \sim ED(\tau(\theta), \lambda^{-1}) \Leftrightarrow \lambda Y \sim ED^*(\theta, \lambda). \tag{3}$$

The corresponding probability density function for the variable $Z = \lambda Y$ has the form

$$p(z; \theta, \lambda) = a^*(\lambda; z) \exp\{z\theta - \lambda\kappa(\theta)\}, \tag{4}$$

for a suitable function a^*. The model $ED^*(\theta, \lambda)$ has mean $m = \lambda\mu$, variance $\lambda V(\mu) = \lambda V(m/\lambda)$, and satisfies the convolution formula

$$ED^*(\theta, \lambda_1) \star \ldots \star ED^*(\theta, \lambda_k) = ED^*(\theta, \lambda_1 + \ldots + \lambda_k), \tag{5}$$

where $\star$ denotes convolution.

The convolution formula (5) implies that there exists a stochastic process $Z(t)$ satisfying $Z(0) = 0$ with stationary and independent increments, such that the increment $Z(t+s) - Z(t)$ has distribution $ED^*(\theta, s\rho)$, say. If the model is infinitely divisible, so s varies in $\mathscr{R}_+$, we have a continuous time process, in fact a Lévy process. See Jørgensen (1992) for further details about this type of process.

The *Tweedie class* of exponential dispersion models is defined as having variance function $V(\mu) = \mu^p$, for some p in the set $(-\infty, 0] \cup [1, \infty)$ (Tweedie, 1984; Jørgensen, 1987). The corresponding exponential dispersion model, denoted $ED^{(p)}(\mu, \sigma^2)$, satisfies the scale transformation property

$$cED^{(p)}(\mu, \sigma^2) = ED^{(p)}(c\mu, c^{2-p}\sigma^2) \tag{6}$$

for $c > 0$, in fact this property characterizes the Tweedie class. Especially interesting from a risk theory point of view is the case $p \in (1, 2)$ (which is what we mean when we refer to the Tweedie class in the following), because in this case we have compound Poisson distributions, as will be explained in Section 2. Other important special cases of the Tweedie class are $p = 0$, 1, 2 and 3, which correspond to respectively the normal, Poisson, gamma and inverse Gaussian distributions.

2. THE TWEEDIE MODEL

We now consider in detail the Tweedie model. For the moment, we assume that $v = 1$, so that the exposure w is equal to time t. Hence, let us assume in (1) that $N(t)$ follows a Poisson process with rate m, such that the distribution of $N(t)$ is $Po(tm)$, and let us assume that the distribution of the individual claims is

$$Z_i \sim Ga(-\theta, -\alpha), \tag{7}$$

where $Ga(-\theta, -\alpha)$ denotes the gamma distribution with density

$$\frac{(-\theta)^{-\alpha}}{\Gamma(-\alpha)} z^{-\alpha-1} e^{\theta z},$$

with $z > 0$ and $\alpha, \theta < 0$. The parameters α and θ are taken as negative in order to make the notation in the following agree with the standard notation for the Tweedie class. Then

$$Z(t) \mid N(t) \sim Ga(-\theta, -\alpha N(t)).$$

We now introduce the parameter $\lambda > 0$ defined by

$$m = \lambda^{1-\alpha}\kappa_\alpha(\theta),$$

where

$$\kappa_\alpha(\theta) = \frac{\alpha - 1}{\alpha}\left(\frac{\theta}{\alpha - 1}\right)^\alpha.$$

This gives the following joint density function for $(Z(t), N(t))$, for $n \geqslant 1$ and $z > 0$,

$$f_{Z(t),N(t)}(z, n; \theta, \lambda, \alpha) = \frac{\{\lambda^{1-\alpha}t\kappa_\alpha(-1/z)\}^n}{\Gamma(-n\alpha)n!\, z}\exp\{\theta z - \lambda^{1-\alpha}t\kappa_\alpha(\theta)\}, \tag{8}$$

and

$$P(N(t) = 0) = P(Z(t) = 0) = \exp\{-\lambda^{1-\alpha}t\kappa_\alpha(\theta)\}.$$

2.1. The distribution of the total claim $Z(t)$

We may now derive the marginal distribution of $Z(t)$, given by

$$f_{Z(t)}(z; \theta, \lambda, \alpha) = \sum_{n=1}^{\infty} \frac{\{\lambda^{1-\alpha}t\kappa_\alpha(-1/z)\}^n}{\Gamma(-n\alpha)n!\, z}\exp\{\theta z - \lambda^{1-\alpha}t\kappa_\alpha(\theta)\},$$

for $z > 0$, and

$$P(Z(t) = 0) = \exp\{-\lambda^{1-\alpha}t\kappa_\alpha(\theta)\}.$$

By comparing with (4), we find that this is an exponential convolution model of the form

$$ED^*(\theta, \lambda^{1-\alpha}t),$$

the density being taken with respect to the measure given by the Lebesgue measure plus a unit mass in zero.

In this particular case, the convolution formula (5) takes the form

$$ED^*(\theta, \lambda^{1-\alpha}t_1) \star \ldots \star ED^*(\theta, \lambda^{1-\alpha}t_k) = ED^*(\theta, \lambda^{1-\alpha}(t_1 + \ldots + t_k)).$$

This equation shows that if we accumulate claims over non-overlapping time intervals (giving independent contributions), then the total claim follows a member of the Tweedie class with t given by the total time interval. This implies that the development until now may be repeated if instead of time t we use exposure w, provided that the exposure behaves in an additive way, and that non-overlapping exposures generate independent claims. In particular, the exposure defined as $w = vt$, where v is the value insured and t is the time during which the value v is at

risk, satisfies this requirement. From now on, we hence derive the distribution of $Z(w)$ etc., where $Z(w)$ is the total claim corresponding to the exposure w.

2.2. The distribution of the claim rate Y(w)

By (3) the distribution of $\bar{Z}(w) = Z(w)/(\lambda^{1-\alpha}w)$ is an exponential dispersion model. In fact, an easy calculation (see e.g. Jørgensen, 1987) shows that the corresponding variance function is $V(\mu) = \mu^p$ for some $1 < p < 2$, where

$$p = \frac{\alpha - 2}{\alpha - 1}.$$

This then implies that

$$\bar{Z}(w) \sim ED^{(p)}\left(\lambda^{\alpha-1}\mu, \frac{\lambda^{\alpha-1}}{w}\right),$$

where $\mu = \kappa_\alpha'(\theta/\lambda)$. Furthermore, by (6) we obtain

$$Y(w) \sim ED^{(p)}\left(\mu, \frac{\sigma^2}{w}\right), \tag{9}$$

where $\sigma^2 = 1/\lambda$.

Formula (9) is central in our development, because it is of the form required for constructing generalized linear models, and puts at our disposal the standard tools for inference in generalized linear models. In particular, (9) shows that the exposure w enters as weight in the analysis. The density function of $Y(w)$ is given in Section 2.3.

Due to the fact that Tweedie (1984) seems to have been the first to study the compound Poisson model with gamma claims from the point of view of what is now known as exponential dispersion models, we have chosen to let his name be associated with the process (1) in the case studied here.

The main parameter in (9) is the mean claim rate μ, and σ^2 is a dispersion parameter. As may be seen from (9), the exposure w enters the model via σ^2. In particular, we note that Var $Y(w)$ is proportional to σ^2/w, making the variance small if the exposure is large, showing that the model takes proper account of the amount of exposure.

The parameter α is a shape parameter, in fact $(-\alpha)^{-1/2}$ is the coefficient of variation of the claim-size distribution (7). This makes α an important index in connection with risk theory. Thus, for example a small coefficient of variation in (7) corresponds to a policy of life insurance type, where the claim size is constant or almost so. Similarly, a large coefficient of variation corresponds to for example car insurance, where the variability of the claim-size is large. The parameter α hence characterizes an important aspect of the risk process, and would be an important tool for classification in risk theory.

The convolution formula (5) takes the following form when applied to (9). If $Y_i \sim ED^{(p)}(\mu, \sigma^2/w_i)$, $i = 1, \ldots, k$ are independent, then

$$\frac{1}{w_+}\sum_{i=1}^{k} w_i Y_i \sim ED^{(p)}\left(\mu, \frac{\sigma^2}{w_+}\right), \tag{10}$$

where $w_+ = w_1 + \ldots + w_k$. Formula (10) says that the method is closed with respect to pooling of claims with the same claim rate, the pooling corresponding to simply adding up the individual exposures.

2.3. The joint distribution of Y(w) and N(w)

From (8) we obtain the following joint density for $(Y(w), N(w))$,

$$f_{Y(w),\,N(w)}(y, n; \theta, \lambda, \alpha) = \frac{\{(\lambda w)^{1-\alpha}\kappa_\alpha(-1/y)\}^n}{\Gamma(-n\alpha)n!\,y} \exp[\lambda w\{\theta_0 y - \kappa_\alpha(\theta_0)\}] \tag{11}$$

and

$$P(Y(w) = 0) = P(N(w) = 0) = \exp\{-\lambda w \kappa_\alpha(\theta_0)\},$$

where $\theta_0 = \theta\lambda^{1/(1-\alpha)}$.

We may now calculate the marginal density of $Y(w)$ as follows

$$f_{Y(w)}(y; \theta, \lambda, \alpha) = \sum_{n=1}^{\infty} \frac{\{(\lambda w)^{1-\alpha}\kappa_\alpha(-1/y)\}^n}{\Gamma(-n\alpha)n!\,y} \exp[\lambda w\{\theta_0 y - \kappa_\alpha(\theta_0)\}] \tag{12}$$

and

$$P(Y = 0) = \exp\{-\lambda w \kappa_\alpha(\theta_0)\}.$$

Also, we may calculate the conditional distribution of $N(w)$ given $Y(w)$. Define

$$p(y, n; \lambda, \alpha) = \frac{\{(\lambda w)^{1-\alpha}\kappa_\alpha(-1/y)\}^n}{\Gamma(-n\alpha)n!\,y}.$$

For $y > 0$ we obtain the conditional probability function

$$f_{N(w)\,|\,Y(w)}(n \mid y; \lambda, \alpha) = \frac{p(y, n; \lambda, \alpha)}{\sum_{i=1}^{\infty} p(y, i; \lambda, \alpha)}, \tag{13}$$

for $n = 1, \ldots$. In particular, the conditional mean of $N(w)$ given $Y(w)$ is

$$E(N(w) \mid Y(w)) = \frac{\sum_{n=1}^{\infty} np(y, n; \lambda, \alpha)}{\sum_{n=1}^{\infty} p(y, n; \lambda, \alpha)}. \tag{14}$$

Note that (13) does not involve θ_0, because $Y(w)$ is sufficient for θ_0 when (λ, α) is known. The conditional expectation (14) may be useful in connection with the EM-algorithm, as mentioned in the Introduction.

3. PARAMETER ESTIMATION

Consider independent pairs of data of the form $(Y_1, N_1), \ldots, (Y_m, N_m)$, such that the ith pair $(Y_i, N_i) = (Y(w_i), N(w_i))$, based on the exposure w_i, follows the Tweedie model (11) with parameters $(\mu_i, \sigma^2/w_i, \alpha)$. This implies in particular that

$$Y_i \sim ED^{(p)}\left(\mu_i, \frac{\sigma^2}{w_i}\right). \tag{15}$$

We consider estimation of the parameter vector $(\beta, \sigma^2, \alpha)$ under this model, based on maximum likelihood, assuming the regression structure

$$\eta_i = \log(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_j, \tag{16}$$

where $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$. Let $X = \{x_{ij}\}$ denote the design matrix of the model, and let us denote observed quantities of Y_i etc. by the corresponding lower-case letters.

The model given by the marginal distribution of Y_i in (15) and (16) is a standard generalized linear model, a fact that will be useful for various purposes in the following. We refer the reader to Jørgensen (1987) and McCullagh & Nelder (1989) for the terminology of generalized linear models used below.

In (16) we assume a logarithmic link function, although other links may be used. In principle, the form of the link function should be confirmed by the data, but the logarithmic link function has the advantage of giving a multiplicative form for the net premium, corresponding to common actuarial practice, where the premium is calculated as a product of a base rate and various correction factors. This shows an important advantage of generalized linear models, namely that the error distribution and the link function may be freely combined, in contrast to classical linear models, where the identity link is the only possibility. See Chang & Fairly (1978) for some related discussion.

In the following we let $\mu_i = \mu_i(\beta)$ denote μ_i considered as a function of β according to the relation given by (16), and similarly for η_i. The log likelihood for the parameters $(\beta, \sigma^2, \alpha)$ is given by

$$L(\beta, \sigma^2, \alpha) = \sum_{i=1}^{m}\Big[n_i\{(1-\alpha)\log(w_i/\sigma^2) + \log \kappa_\alpha(-1/y_i)\} - \log\Gamma(-n_i\alpha) + \frac{w_i}{\sigma^2}\{y_i\mu_i^{1/(\alpha-1)}(\alpha-1) - \kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\}\Big].$$

As noted in Section 2.3, the conditional distribution of N_i given Y_i does not depend on μ_i, corresponding to a factorization of the form

$$f_{Y,N}(y, n; \beta, \sigma^2, \alpha) = f_Y(y, n; \beta, \sigma^2, \alpha) f_{N|Y}(n \mid y; \sigma^2, \alpha).$$

We note that for α fixed, the estimate of β may be found from the first factor alone, that is from the marginal distribution of Y_i in (15). Another important result is that the parameter (σ^2, α) is orthogonal to β. This follows from a general result for orthogonality in partly exponential families (Barndorff-Nielsen, 1978, p. 184), due to the fact that for fixed value of (σ^2, α), the model (11) is an exponential family with mean value parameter μ. These two results will be important in the following.

3.1. Estimation based on Y for α known

It is useful to consider first estimation based on the marginal distribution for Y, (15), for α fixed, because this provides the kernel of the general estimation procedure below. Since the model given by (15) and (16) is a generalized linear model, it may be fitted by the iterative least-squares algorithm of Nelder &

Wedderburn (1972), implemented in GLIM. We recall that the estimate of β in this method is the same, whatever the value of σ^2.

The iterative weighted least-squares algorithm for calculating the maximum likelihood estimate of β is of the form

$$\beta^* = (X^T W X)^{-1} X^T W z \tag{17}$$

where β^* denotes the updated value, X is the design matrix, W is a diagonal weight matrix with elements

$$W_{ii} = \frac{w_i}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2,$$

and where

$$z_i = \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}, \tag{18}$$

both evaluated at the previous value of β.

3.2. The profile likelihood for α

The general structure of the algorithm for estimating the full set of parameters $(\beta, \sigma^2, \alpha)$ relies on a number of properties of the profile likelihood for α. The algorithm as such is explained in Section 3.3, but we now consider the relevant aspects of the profile likelihood.

The profile likelihood for α (considering for the moment β as known) is $L(\beta, \hat{\sigma}_\alpha^2, \alpha)$, where $\hat{\sigma}_\alpha^2$ is the maximum likelihood estimate for σ^2 for β and α known. The estimate is

$$\hat{\sigma}_\alpha^2 = \frac{-\sum_{i=1}^{m} w_i \{ y_i \mu_i^{1/(\alpha-1)} (\alpha - 1) - \kappa_\alpha (\mu_i^{1/(\alpha-1)} (\alpha - 1)) \}}{n_+ (1 - \alpha)}, \tag{19}$$

where $n_+ = n_1 + \ldots + n_m$. Hence the profile likelihood for α is

$$\tilde{L}_\beta(\alpha) = L(\beta, \hat{\sigma}_\alpha^2, \alpha) = (1 - \alpha) \sum_{i=1}^{m} n_i \log(w_i / \hat{\sigma}_\alpha^2) + \sum_{i=1}^{m} n_i \log \kappa_\alpha(-1/y_i)$$

$$- \sum_{i=1}^{m} 1_{\{n_i \neq 0\}} \log \Gamma(-n_i \alpha) + n_+ (\alpha - 1).$$

Figure 1 shows a simulated example of a profile likelihood for α, for a sample of $m = 50$ with $\mu_i = 2$, $\sigma^2 = 0.5$ and $\alpha = -1$ ($p = 1.5$). Figure 2 shows the same profile likelihood, but as a function of p. The main characteristics of the profile likelihood are that there is a central region with a local maximum, whereas the likelihood tends to infinity at $p = 1$, respectively $\alpha = -\infty$, giving rise to a local minimum of the profile likelihood to the left of the local maximum. A small simulation study with varying values of the parameters indicated that the local maximum is always fairly near the true value of α, such that the corresponding estimator probably corresponds to a consistent root of the likelihood equation. This is the estimator adopted in the following.

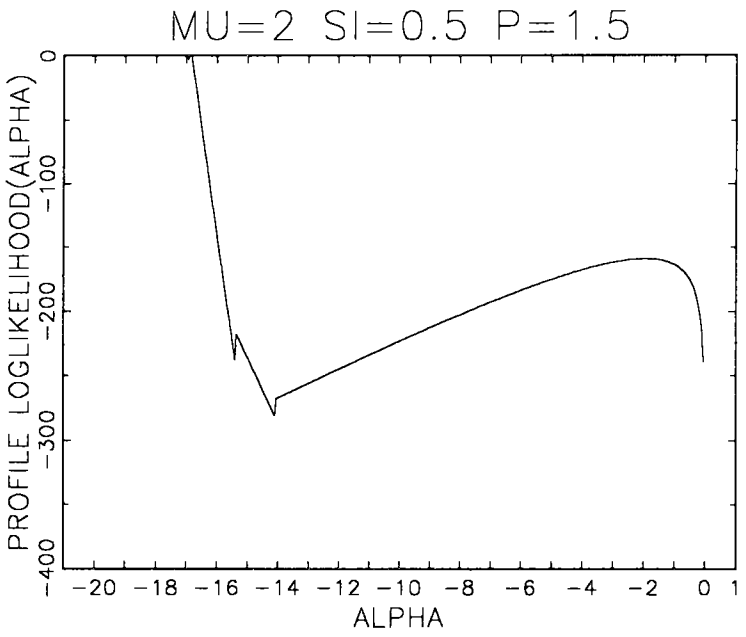

Fig. 1. The profile log likelihood for α, simulated data.

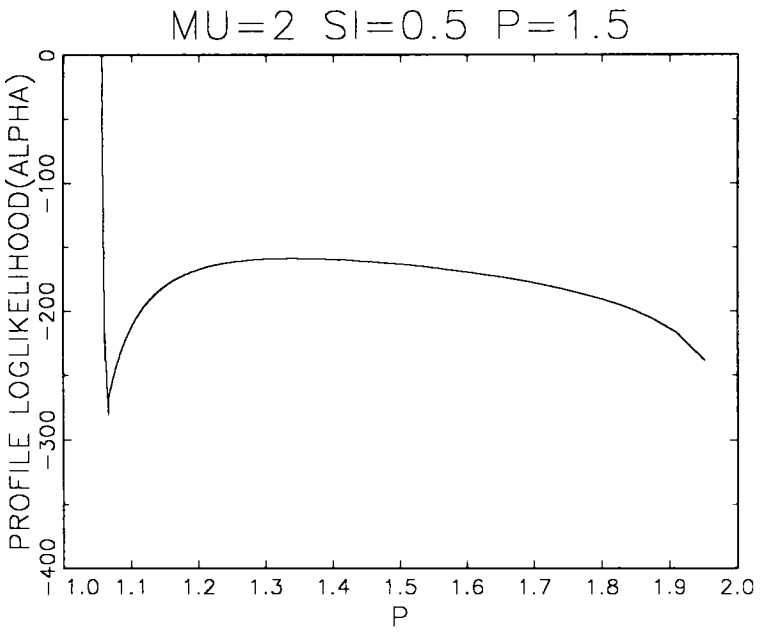

Fig. 2. The profile log likelihood for p, simulated data.

The phenomenon that the profile likelihood tends to infinity at $p = 1$ is due to the fact that the corresponding limiting form of the Tweedie class is proportional to the Poisson distribution, which is discrete, leading for p near 1 to a Dirac delta form of the density on a set of lattice points. A further unusual property of the profile likelihood is the small oscillations near the local minimum, which may be due to the superposition of the Dirac delta functions just mentioned.

An algorithm for obtaining the local maximum of the profile likelihood must keep the iterations away from the area around and to the left of the local minimum, an area that we refer to as the *dangerous* region. A second problem is to always maintain the value of α negative. Note that the likelihood based on (Y, N) cannot be extended to values of p bigger than 2, because the Tweedie class can no longer be interpreted as a compound Poisson model in this case.

The need to maintain the iterations in the central area of the parameter space suggests that we look for a parameter transformation to symmetrize the profile likelihood. Neither of the parameters α and p meet this objective, because the profile likelihood is skewed towards the left for p, and towards the right for α as is evident from Figs. 1 and 2. We have found that the parameter

$$\xi = -\log(-\alpha)$$

symmetrizes the profile likelihood to a reasonable extent, as evidenced by Fig. 3. This parameter has the further advantage that its domain is $\mathscr{R}$, avoiding any problems caused by a finite upper terminal for the domain.

The Newton-Raphson algorithm applied to the profile likelihood for ξ, but expressed in terms of α, is given by the iteration

$$\alpha^* = \alpha \exp\left\{\frac{H_\beta(\alpha)}{-\alpha H'_\beta(\alpha)}\right\}, \tag{20}$$

where α^* denotes the updated value, and where

$$H_\beta(\alpha) = -\alpha \frac{\partial \tilde{L}_\beta}{\partial \alpha}(\alpha)$$

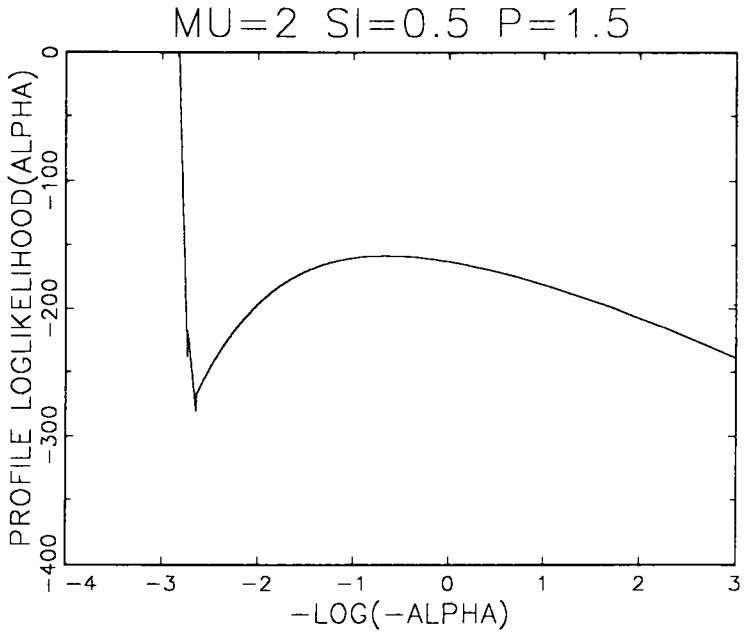

Fig. 3. The profile log likelihood for ξ, simulated data.

is the profile score function for ξ, expressed in terms of α. The derivatives of the profile likelihood $\tilde{L}_\beta$ are given in Appendix A.

3.3. The algorithm

We now consider the algorithm for estimating the full set of parameters $(\beta, \sigma^2, \alpha)$. The basic form of the algorithm is that of a stabilized Newton-Raphson algorithm, but the detailed design of the algorithm relies on the properties of the likelihood derived above. The algorithm could be programmed in any language with matrix-handling facilities, but we describe an implementation in GLIM 4. This implementation takes advantage of GLIM's language for specifying linear models, a fact that eliminates a lot of programming effort, and makes it easy to explore models in an interactive fashion, thus preserving some of the main advantages of GLIM. The program for the algorithm is given in Appendix C.

The main property of the likelihood is the orthogonality of the parameters β and (σ^2, α), which makes the expected information matrix block diagonal, with an $m \times m$ block corresponding to β, and a 2×2 block corresponding to (σ^2, α). The inverse information matrix is hence also block diagonal, and corresponds to inverting the blocks individually. This in turn leads to a separation of the Newton-Raphson equations in two separate equations, one for each of the two parameter blocks. The updating of for example β may thus be performed as if the parameters (σ^2, α) were known, and similarly for updating (σ^2, α). The algorithm hence swops between updating β via (17) and updating (σ^2, α) via (19) and (20), continuing until convergence. Note that starting values for the iteration (17) may be obtained from the data, by taking $\eta_i = g(y_i) = \log y_i$ in (18), with a suitable modification for $y_i = 0$.

Schematically, the algorithm takes the form:

- **Initialization**
  - —Let $\alpha = -0.10$.
  - —Let $\eta_i = g(y_i)$.
- **Repeat**
  - —Beta Step: Update β.
  - —Sigma Step: Update σ^2.
  - —Alpha Step: Update α.
- **until convergence**.

The initial value of α is chosen fairly near zero to avoid the iterations entering the dangerous region. However, the algorithm is fairly insensitive to the initial value for α, and when a sequence of similar models is fitted (for example adding or deleting variables from the regression model) the previous estimate for α is a very good starting value.

In practice, the Sigma and Alpha steps are not performed explicitly by the user in GLIM. The trick is to include the Sigma and Alpha steps in the macros for fitting (15) and (16). Since these macros are called in each iteration of the form (17) performed by GLIM, the result becomes the algorithm described above.

The stopping criterion used by GLIM is based on the relative decrease of the value of the deviance, whereas for our purpose it ought to be defined in terms of the increase of the likelihood. We hence take the precaution of using a fairly strict stopping criterion, to make sure that the iterations do not stop prematurely. For the first fit in a session, based on the pre-assigned starting value $\alpha = -0.1$, this tends to give about 12 to 15 iterations, whereas subsequent fits, in which the value of α from the previous fit is used as starting value, take less than 10 iterations. This is satisfactory compared with the average of about 4–6 iterations for a standard generalized linear model fitted in GLIM.

4. INFERENCE

In order to calculate standard errors for the parameter estimates in the model defined in Section 3, we again take advantage of the orthogonality of the parameters β and (σ^2, α). The orthogonality implies that the standard errors for the components of the maximum likelihood estimate $\hat{\beta}$ may be calculated based on the expected information for β as if (σ^2, α) were known. Furthermore, the expected information matrix for β has σ^2 entering as a proportionality factor. Since GLIM estimates σ^2 by $\tilde{\sigma}^2 = D/\mathrm{df}$, where D is the deviance and df the degrees of freedom for the model (15), the standard errors produced by GLIM need to be multiplied by the correction factor $\hat{\sigma}/\tilde{\sigma}$. This produces the standard errors for $\hat{\beta}$ based on the expected information matrix.

In a similar fashion, the standard errors of the maximum likelihood estimate $(\hat{\sigma}^2, \hat{\alpha})$ based on the expected information matrix may be calculated as if β were known. However, it is somewhat easier to use an approximate procedure based on the observed information matrix. Note that the expected orthogonality of the parameters β and (σ^2, α) implies approximate observed orthogonality of the same parameters. Hence, we may invert the (σ^2, α) block of the observed information matrix to obtain the (approximate) asymptotic variance matrix for $(\hat{\sigma}^2, \hat{\alpha})$. The resulting standard error for $\hat{\alpha}$ may also be obtained from the inverse observed profile information for α, calculated from the second derivative of $\tilde{L}_\beta(\alpha)$, cf. Barndorff-Nielsen (1988, p. 31). In practice, the first method turned out to be rather unstable, and we have hence used the observed profile information for the calculation of the standard error for $\hat{\alpha}$. For reference, the formulae for the calculation of the observed information matrix for (σ^2, α) are given in Appendix B.

4.1. The likelihood ratio test

Turning now to hypothesis testing, we note that, due to the observation of N jointly with Y, the usual analysis of deviance from generalized linear models is inefficient, because it does not take N into account. Instead, the inference for the Tweedie model is done via the likelihood ratio test, defined by

$$\mathrm{LR} = L(\hat{\beta}_1, \hat{\sigma}_1^2, \hat{\alpha}_1) - L(\hat{\beta}_2, \hat{\sigma}_2^2, \hat{\alpha}_2),$$

where the subscripts denote estimates under the hypotheses H_1 and H_2, respectively, H_2 being a sub-hypothesis of H_1. Assuming that the hypotheses are smooth, the

asymptotic distribution of $2 \times \text{LR}$ is a chi-squared distribution with suitable degrees of freedom.

It is nevertheless interesting to compare the F-test from analysis of deviance with the above likelihood ratio test. To illustrate, we consider the first test performed in Section 5, in connection with Table 1. The result for the likelihood ratio test is

$$2 \times \text{LR} = 9.288, \quad \text{df} = 17, \quad p = 0.93,$$

where p denotes the p-value, while the F-test based on the deviance is

$$\text{F} = 1.359, \quad \text{df} = (17,453), \quad p = 0.15.$$

The F-test based on Pearson's X^2 gives

$$\text{F} = 0.7335, \quad \text{df} = (17,453), \quad p = 0.77.$$

While the three approximate p-values are quite different, it is interesting to note that the best agreement is between the likelihood ratio test and the F-test based on the Pearson statistic, while the F-test based on the deviance produces a somewhat discrepant p-value. This confirms the fact that the estimator $\tilde{\sigma}^2$, which enters in the latter test, is in general biased, and hence produces a biased test. These results should be interpreted with some care, because the deviances on which the F-tests are based, depend on the estimates of α under the two hypotheses, which are stochastic, and differ between the hypotheses.

5. ANALYSIS OF THE BRAZILIAN CAR INSURANCE DATA

In 1987, data were collected from the automobile portfolios of the four major insurance companies of Brazil for the period March 1985 to February 1986. The data analyzed here concern policies for private cars with coverage for collision. The dependent variables and the weight were defined as follows:

Number of Claims (N)—The number of claims during the interval of exposure.

Total Claim (Z)—The total claim (corrected for inflation) during the interval of exposure.

Exposure (w)—The insured value (corrected for inflation) times the length of the exposure interval (in years).

The following five independent variables were included:

Model—Factor with 5 levels. Represents a classification of the model of the insured vehicle.

Age—Continuous variable. Represents age of insured vehicle (years) at the onset of the policy.

Deductible—Factor with 3 levels. Represents the value deducted from a claim before payment. The first level represents the case where there is no deduction.

Bonus—Factor with 7 levels. Indicates the discount in premium due to one or more previous policies with no claims during one year. The first level of the factor represents the case of no bonus.

Region—Factor with 19 levels. Represents the geographical region where the vehicle circulates.

Table 1. *Reduction of levels*

Factor	New level	Old levels	Description
Model	1	1, 2, 5	simple cars
	2	3, 4	expensive cars
Bonus	1	1, 2	
	2	3	
	3	4	
	4	5	
	5	6, 7	
Region	1	1, 2, 4	Manaus, Belém and Fortaleza
	2	3	Other cities in the North Region
	3	5, 6, 7	Recife, Salvador and other cities in the North East Region
	4	8, 9	Belo Horizonte, Vitória, other cities in the South East Region,
		13, 14, 15	Curitiba and Porto Alegre
	5	10, 11, 12	Greater Rio de Janeiro,
			Greater São Paulo and the State of São Paulo
	6	16	Other cities in the South Region
	7	17, 18, 19	Center East Region

The present analysis concerns a random sample of 4440 policies from the original database. The data were grouped according to the values of the dependent variables, and the accumulated values of w, $N(w)$ and $Z(w)$ were calculated together with the value of $Y(w)$, giving a total of 485 groups, of which 323 had no claims. Note that the justification for this calculation comes from the convolution property (10).

As our basic model, we adopt the Tweedie model as explained in Section 2, and the regression model (16). Our initial model includes the main effects for the five independent variables. Based on this model, we first investigate the question of reducing the number of levels for each of the four factors of the model. This is especially important for the variable Region, which has 19 levels. The results are given in Table 1, where we show the grouping of the variables Model, Bonus and Region. No grouping was done for the factor Deductible.

The grouping of the variable Region is partly based on previous knowledge, and partly on empirical findings from the data itself. As may be seen from Table 1, an important feature of the grouping consists in a distinction between the major city centres, like Rio de Janeiro and São Paulo, major state capitals, like Belo Horizonte and Vitória, and a number of further regional divisions that are natural given the geographical features of Brazil. To verify the validity of the grouping, we make a test based on the likelihood ratio test ($2 \times \mathrm{LR} = 9.288$; $\mathrm{df} = 17$; $p = 0.93$), showing a clear confirmation of the grouping. The resulting model, with 470 degrees of freedom, includes the main effects for the five variables, but with the reduced number of levels for the factors described in Table 1.

Before turning to the question of interactions, we test the significance of each of the variables in turn against the full model. The results are given in Table 2. All five variables are significant.

Table 2. *Tests for main effects*

Variable	2 × LR	df	*p*
Region	25.97	6	0.0003
Model	34.25	1	0.0000
Bonus	48.78	4	0.0000
Age	26.01	1	0.0000
Deductible	22.72	2	0.0000

Table 3. *Tests for interactions*

Model	2 × LR	df	*p*
All two-factor interactions	–	–	–
Model. (Deductible + Age) + Bonus + Region	74.60	63	0.1503
Model + Deductible + Age + Bonus + Region	20.49	3	0.0002

5.1. Interactions

The question of interactions is very important for actuarial practice. According to standard practice, the premium is calculated as the product of a standard premium and various correction factors, corresponding to for example vehicle model, region, bonus etc. This corresponds to our model with log link and main effects, but no interactions, which corresponds to a multiplicative model for the claim rate.

The introduction of interactions complicates this pattern considerably. When an interaction exists between two given factors, the factors are replaced by a compound factor, given by a cross-table for the two factors. The corresponding correction factor should be calculated according to this table. The introduction of interactions should hence be justified and interpreted carefully, due to the extra complication involved.

In our analysis we have studied all interactions between pairs of variables, testing their significance individually. Table 3 shows a summary of this analysis, in the form of three models, of which one is the model

$$\text{Model.Age} + \text{Model.Deductible} + \text{Region} + \text{Bonus}, \tag{21}$$

in the standard GLIM notation, whereas the two other models represent the cases of no interactions (the model used above), and the model with all the possible interactions between two variables present. The tests based on the likelihood ratio statistic in Table 3 show that the model (21) is acceptable when compared with the model with all interactions, whereas the interactions between Model and Deductible, and between Model and Age, are in fact significant. Our final model is hence (21). The parameter estimates for this model is given in Table 4, in GLIM notation.

We have not considered interactions among three or more variables, because such interactions are difficult to interpret, and require many observations for their analysis. In the present case, even analysis of three-variable interactions would be

Table 4. *Parameter estimates for Model (21)*

Variable	Coefficient	s.e.
1	−4.078	0.633
Deductible (2)	−0.272	0.132
Deductible (3)	0.221	0.303
Model (2)	−0.574	0.133
Age	0.0864	0.0305
Bonus (2)	−1.036	0.223
Bonus (3)	−0.575	0.231
Bonus (4)	0.0270	0.201
Bonus (5)	−0.839	0.144
Region (2)	1.206	0.703
Region (3)	0.816	0.638
Region (4)	0.317	0.643
Region (5)	0.918	0.633
Region (6)	1.779	0.676
Region (7)	0.600	0.709
Deductible (2).Model (2)	−0.441	0.175
Deductible (3).Model (2)	−1.639	0.490
Model (2).Age	0.144	0.043

difficult, due to the many levels of each of the factors. However, while difficult to interpret, a statistically significant high order interaction may be important for tarification.

The estimate of ξ is -0.5179 with asymptotic standard error 0.0853, and the estimate of σ^2 is $\hat{\sigma}^2 = 0.4061$. The estimates of α and p are $\hat{\alpha} = -1.678$ and $\hat{p} = 1.373$. An asymptotic 95% confidence interval for ξ, based on the above standard error, is $[-0.6851, -0.3506]$, and the corresponding (asymmetric) confidence intervals of α and p, are, respectively, $[-1.984, -1.420]$ and $[1.335, 1.413]$. The latter interval does not include the value $p = 1.5$, corresponding to the non-central chi-squared distribution with zero degrees of freedom. This is not surprising, because the value $p = 1.5$ corresponds to an exponential claim-size distribution, and p in the interval (1.5, 2) gives a claim-size distribution with mode in zero, a feature that is unlikely for data on car insurance. The case of a value for p in the interval (1, 1.5) corresponds to a unimodal claim-size distribution with a positive mode, which is much more realistic for insurance data.

Figure 4 shows a residual plot for the final model, the deviance residuals for Y being plotted against the linear predictor η. The main function of this plot is to confirm the form of the variance function, and to highlight outliers if any are present. The graph shows that a number of negative residuals fall on a well-defined smooth curve, the remaining residuals lying above this curve. In fact, the residuals on the curve correspond to the zero observations in the data, giving the smallest possible residual for each given value of the linear predictor. As shown on the graph, this lower limit becomes lower and lower as the value of η increases, while for small values of η this lower limit is very close to zero.

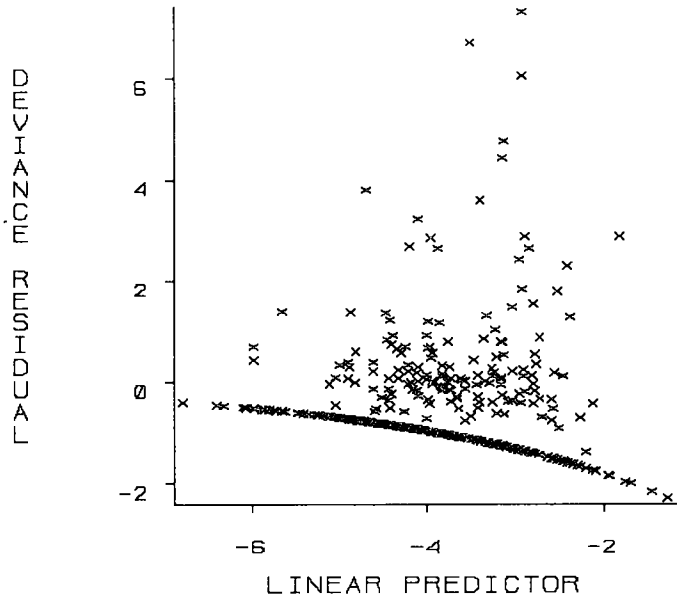

Fig. 4. Plot of deviance residuals versus linear predictor, car data.

This feature of the residual plot is hence in agreement with the model, but makes it more difficult to interpret the plot. The remaining residuals, corresponding to positive observations, seem to behave as one expects for residuals from a continuous model, except for the presence of some very large residuals. However, the number of large residuals is roughly in agreement with what one might expect with as many as 485 observations, and the plot hence does not give rise to questioning the model.

5.2. Interpretation of the model

Tables 5–8 show the correction factors entering in the model (21), obtained essentially by exponentiating the values of Table 4. The estimated mean claim rate

Table 5. *Correction factors for Model and Age*

Model	1	2
Factor	1.09^{Age}	1.15^{Age}

Table 6. *Correction factors for Deductible and Model*

Deductible		1	2	3
Model	1	1	0.7617	1.247
	2	0.5631	0.2758	0.1363

Table 7. *Correction factors for Region*

Region	1	2	3	4	5	6	7
Factor	1	3.340	2.262	1.373	2.504	5.924	1.823

Table 8. *Correction factors for Bonus*

Bonus	1	2	3	4	5
Factor	1	0.3549	0.5629	1.027	0.4322

for any given combination of factor levels may thus be obtained as a product of a base rate and the corresponding factors. The base level corresponds to level one for each of the for factors and Age zero, giving a claim rate of 0.0169. This is hence the cost per unit of exposure of a policy for a new car of Model 1 in Region 1 (Manaus, Belém and Fortaleza), with no Deductible and no Bonus.

A number of interesting conclusions follow from the values in Tables 5–8. From Table 5 we find that the claim rate increases faster with the age of the vehicle for the expensive car group than for the simple car group. However, if we take group 1 of the factor Deductible as a reference, we will find from the first column of Table 6 that the claim rate for a new car is lower for the expensive car group than for the inexpensive car group. A simple calculation shows that the two car groups present identical claim rates for Age of about 4 years. From 4 years onwards, the expensive car group presents the highest claim rate.

The same basic pattern is repeated for Deductible groups 2 and 3, the intersection being at respectively Age = 7 years and Age = 15.4 years. In the case of Deductible group 3, the claim rate for the inexpensive car group thus lies above the rate for the expensive car group during essentially the whole lifetime of the car. A possible explanation for this behaviour could lie in the fact that most cars change owner several times during their lifetime, in general passing on to owners having a lower socio-economic status, and perhaps a different driving pattern. The low claim rate for new and expensive cars and old and inexpensive cars may perhaps be explained by the fact that the former tend to be owned by middle-aged drivers with a high socio-economic status, while the latter tend to be owned by middle-aged drivers with a low socio-economic status. This indicates that age of the driver could be an important explanatory variable, as is known from other studies, see e.g. Andrade e Silva (1989).

The first row of Table 6 shows a remarkable pattern for the factor Deductible, in that the inexpensive car group presents a claim rate that is not decreasing with the level of Deductible, as one might have expected, but shows the highest claim rate for the highest level of Deductible, and the lowest claim rate for the middle group. We do not have any explanation for this behaviour, but it should be kept in mind that the data come from four different insurance companies, such that this

phenomenon might be the result of differences between the four portfolios unaccounted for by the explanatory variables. As pointed out by the referee, another possible explanation is that Deductible may be correlated with other known covariates, for example model of car, since there may be an income effect on Deductible for which model of car is a proxy.

Another important aspect related to the interpretation of these results is that the variables Model and Age are not independent of each other over time, because the available car models change from year to year, and furthermore, the characteristics of a given model of car may change slightly from year to year. In fact, the division of Model into just two groups is unfortunate, because it represents a discontinuity between the groups that may not be the ideal way to describe differences between different cars. A much better approach would be to represent car model by one or two continuous explanatory variables, such as for example volume of engine. This variable was used by for example Hallin & Ingenbleek (1983), although they actually grouped it into rather few groups.

Table 7 shows a remarkable variation from region to region. The region with the highest claim rate consists of the cities in the South region other than the two main capitals Curitiba and Porto Alegre, a region that presents a claim rate almost six times the lowest rate, in the Northern state capitals of Manaus, Belém and Fortaleza. The two main cities of Rio de Janeiro and São Paulo together with the interior of São Paulo State occupy a middle position.

The variation with bonus class also calls for a comment. According to Table 8, the lowest rate occurs for Bonus level 2, and the highest level for level 4. Again, this may reflect differences between the portfolios of the four insurance companies.

6. DISCUSSION

Within the framework of the compound Poisson model of the Tweedie form, we have obtained a fairly complete solution for analyzing regression models for the claim rate. Our solution puts results from generalized linear models at our disposal, in particular the implementation of the algorithm in GLIM, which provides a powerful and flexible tool for the analysis of claims data.

Our approach is based on a specific probability model, which facilitates the interpretation of the results. The Tweedie model provides a natural solution to the problem of exact zeros in the data, and takes the exposure into account in a satisfactory way. The interpretation of the parameter α as a kind of dispersion index for the claim distribution suggests that this parameter may be an important tool for classifying portfolios or insurance types. In this sense, the Tweedie model may be useful as a reference model.

The shortcoming of the Tweedie model, as for any compound Poisson model, is that it may be too simple, so that for example the Poisson process should be replaced by some more general mixed or compound process, and similarly for the claim distribution. Here, the approach based on exponential dispersion models may be useful, because as shown by Jørgensen (1987), many of the commonly used

mixed and compound distributions may be interpreted as mixtures of exponential dispersion models, in the sense defined by Jørgensen (1987). This may lead to useful generalizations of the Tweedie model.

APPENDIX

A. Derivatives of the profile likelihood

This appendix contains formulae for derivatives of the profile likelihood. Let ψ denote the digamma function. The derivatives of $\tilde{L}_\beta$ are

$$\frac{\partial \tilde{L}_\beta}{\partial \alpha}(\alpha) = \sum_{i=1}^{m} n_i\{\psi(-\alpha n_i) - \log y_i w_i\} + n_+\left\{\log\frac{\hat{\sigma}_\alpha^2}{1-\alpha} - \alpha^{-1}\right\}$$
$$+ \sum_{i=1}^{m} \frac{w_i}{\hat{\sigma}_\alpha^2}\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\left\{\alpha^{-1} - \frac{\alpha}{(\alpha-1)^2}\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right)\log\mu_i\right\}$$

and

$$\frac{\partial^2 \tilde{L}_\beta}{\partial \alpha^2}(\alpha) = \sum_{i=1}^{m} w_i\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))$$
$$\times\left[\alpha^{-1} - \frac{\alpha}{(\alpha-1)^2}\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right)\log\mu_i\right](n_+(\alpha-1)\hat{\sigma}_\alpha^2)^{-1}$$
$$\times\left[n_+ - \sum_{i=1}^{m}\frac{w_i}{\hat{\sigma}_\alpha^2}\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\left(\alpha^{-1} - \frac{\alpha}{(\alpha-1)^2}\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right)\log\mu_i\right)\right]$$
$$+ n_+[(1-\alpha)^{-1} + \alpha^{-2}] - \sum_{i=1}^{m} n_i^2\psi'(-\alpha n_i) + \sum_{i=1}^{m}\frac{w_i}{\hat{\sigma}_\alpha^2}\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))$$
$$\times\left[-\frac{\alpha}{(\alpha-1)^3}\log\mu_i\left[\left(-\frac{\log\mu_i}{\alpha-1} - 1\right)\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right) + \frac{2(\alpha-1)}{\alpha^2}\right] + \frac{2-\alpha}{\alpha^2(\alpha-1)}\right].$$

B. The information matrix for σ^2 and α

The observed information matrix for σ^2 and α may be found from the following derivatives of the log likelihood,

$$\frac{\partial^2 L}{\partial \alpha^2} = n_+[(1-\alpha)^{-1} + \alpha^{-2}] - \sum_{i=1}^{m} n_i^2\psi'(-\alpha n_i) + \sum_{i=1}^{m}\frac{w_i}{\sigma^2}\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\frac{-\log\mu_i}{(\alpha-1)^2}$$
$$\times\left[-\frac{\alpha}{(\alpha-1)^2}\log\mu_i\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right) + \alpha^{-1}\right]$$
$$+ \sum_{i=1}^{m}\frac{w_i}{\sigma^2}\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\left\{\left(1 - \frac{\log\mu_i}{\alpha-1}\right)\frac{1}{\alpha(\alpha-1)} - \frac{2-\alpha}{\alpha^2(1-\alpha)}\right\},$$

$$\frac{\partial^2 L}{\partial \alpha\,\partial\sigma^2} = \frac{n_+}{\sigma^2} - \sum_{i=1}^{m}\frac{w_i}{\sigma^4}\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\left\{\frac{\alpha}{\alpha-1}\left(1 - \frac{\log\mu_i}{\alpha-1}\right)\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right) + \alpha^{-1}\right\},$$

$$\frac{\partial^2 L}{\partial(\sigma^2)^2} = \frac{n_+(1-\alpha)}{\sigma^4} + \frac{2\alpha}{\sigma^6}\sum_{i=1}^{m} w_i\kappa_\alpha(\mu_i^{1/(\alpha-1)}(\alpha-1))\left(\frac{y_i}{\mu_i} - \alpha^{-1}\right).$$

C. GLIM program

The algorithm was implemented using a beta-test version of GLIM 4. Our experience is based on a PC compatible micro computer with 80386 chip and 80387 mathematical co-processor. A number of specific features of GLIM 4 are used, so the algorithm does not work in the past version, GLIM 3.77. For further information regarding GLIM 4, contact NAG Ltd, Wilkinson House, Jordan Hill Road, Oxford, OX2 8DR, UK.

The following program is an example of a GLIM command sequence to run the estimation program. We assume that the data are available in a file with columns Y, N, w, x, where, for illustration, x denotes a single independent variable.

```
$UNITS m           ! Number of observations
$DATA Y N w x      ! Name data columns
$INPUT 11          ! Read data matrix
$INPUT 12 TWEEDIE  ! Input Tweedie subfile
$YVAR Y            ! Dependent variable is claim rate
$WEIGHT w          ! Weight is exposure
$USE TWEEDIE N     ! Call set-up macro with N as argument
$FIT x             ! Fit a model
$DISPLAY E         ! Beta estimates and standard errors
$USE DEVIATION     ! Confidence intervals for alpha and p
$FIT               ! Fit further models and perform tests etc.
$STOP              ! End session
```

Here follows the main GLIM macros for the estimation algorithm.

```
!************************************************************
$SUBFILE TWEEDIE!
!
!
!-----------------------------------------------------------
!  Author: Bent Jorgensen, IMPA, Rio de Janeiro, Brazil
!  Version 1.0 for beta-test version of GLIM 4 December 1991
!  Main macros:
!    TWEEDIE Set up own macros for the Tweedie compound
!        Poisson model to estimate alpha and sigma**2
!        Formal arguments:
!           %1   Vector with number of claims
!        Macro arguments:
!           None
!    Output: Standard GLIM output, values of alpha, sigma**2
!                and likelihood for each iteration
!
!    DEVIATION Calculate confidence intervals
!              for alpha , xi and p
!        Formal arguments:
!           None
!        Macro arguments:
!           None
```

```
!    Output: standard error for xi and corresponding
!            95% confidence intervals  for alpha and p
!    Example of use:
!        $YVAR Y
!        $WEIGHT w
!        $USE TWEEDIE N
!        $FIT x
!        $USE DEVIATION
!-------------------------------------------------------------
!
!
!
$MAC TWEEDIE ! Set-up macro
!*************************************************************
  $CAL N_=%1 !
  $CAL %A = -0.10 : %P = (%A-2)/(%A-1)  ! Starting value
  $PRINT 'ALPHA START = ' %A '  P = ' %P !
  $ERROR 0 MODEL  ! Define compound Poisson model
  $LINK 0 MYLOG ! Define link function
  $CAL %LP = %LOG(%YV+%EQ(%YV,0)) ! Starting value for eta
  $RECYCLE 20 1 0.000001 ! Stopping criteria
  $CAL %VA = 1 : %DI = 1 : %FV = 0 : %DR = 0 !
  $CAL %N = %CU(N_) ! Total number of claims
  $CAL L_ = (%LOG(%YV+%EQ(%YV,0))+%LOG(%PW+%EQ(%PW,0)))*N_ !
$$ENDMAC !
!
$MAC MYLOG ! Link function
!*************************************************************
  $CAL %FV = %EXP(%LP) : %DR = 1 / %FV ! Logarithmic link
$$ENDMAC !
!
$MAC MODEL ! Main macro
!*************************************************************
  $CAL %P = (%A-2)/(%A-1) ! Value of p
  $CAL %VA=%FV**%P ! Variance function
  $CAL %DI=2*(%A-1)*((%A-1)*%NE(%YV,0)*(%YV+%EQ(%YV,0)) !
   **(%A/(%A-1))/%A !
       +%FV**(%A/(%A-1))*(%A**-1-%YV/%FV)) ! Deviance
  $CAL KALPHA_ = (%A-1)/%A*%FV**(%A/(%A-1)) !
  $CAL DALPHA_ = %PW*(%YV*(%A-1)*%FV**(1/(%A-1))-KALPHA_) !
  $CAL %S = - %CU(DALPHA_)/%N/(1-%A) ! Scale parameter
  $CAL K_ =-N_*%A+%EQ(N_,0) !
  $CAL LOGV_ = %NE(%YV,0)*(-%LOG(%YV+%EQ(%YV,0)) !
      +N_*((1-%A)*(%LOG(%PW/%S)+!
    %LOG(1-%A))-%A*%LOG(%YV+%EQ(%YV,0))-%LOG(-%A))-%LGA(K_)-!
    %LGA(N_+1))+%PW/%S*((%A-1)*%FV**(1/(%A-1))*%YV-KALPHA_) !
  $CAL %Z1 = %CU(LOGV_) ! Log likelihood
  $CAL %Z = %EQ(%ITN,1) ! If first iteration,
  $SKIP %Z ! skip the remainder of the macro
```

```
$CAL H_ = %DIG(K_) !
$CAL I_= %PW*KALPHA_*(-%LOG(%FV)*%A/(%A-1)**2 !
*(%YV/%FV-%A**-1)+%A**-1) !
$CAL %F = -%CU(L_)+%N*(%LOG(%S)-%LOG(1-%A)-%A**-1)+!
  %CU(H_*N_)+%CU(I_/%S) ! Derivative of the log likelihood
$CAL J_ = N_**2*%TRG(K_) !
$CAL M_ = %PW/%S*KALPHA_*(-%LOG(%FV)*%A/(%A-1)**3
*((-%LOG(%FV)/(%A-1)-1)*!
(%YV/%FV-%A**-1)+2/%A**2*(%A-1))+(2-%A)/%A**2/(%A-1)) !
$CAL %G = %CU(I_/%N/(%A-1))/%S*(%N-%CU(I_)/%S) !
    +%N*(1/(1-%A)+%A**-2)-!
          %CU(J_)+%CU(M_) ! Second derivative
  $CAL %A = %A*%EXP(-%F/(%F+%A*%G)) ! Update alpha
  $CAL %X=-%LOG(-%A) !
  $CAL %P=(%A-2)/(%A-1) !
  $PRINT: '   ITNO   SIGMA**2    ALPHA  P    XI    LOGLIK' !
  $PRINT %ITN %S %A %P %X %Z1 !
$$ENDMAC !
!
$MAC DEVIATION ! Standard deviation of estimators
!************************************************************
  $NUMBER  DEVX = 0  !
  $CAL %B=%A*%F+%A**2*%G !
       ! Second derivative with respect to xi
  $CAL DEVX=%SQRT(-1/%B) !
  $CAL %E=-%EXP(-%X-1.96*DEVX) !
  $CAL %T=-%EXP(-%X+1.96*DEVX) !
  $CAL %Q=(%E-2)/(%E-1) !
  $CAL %R=(%T-2)/(%T-1) !
  $PRINT 'XI    STANDARD ERROR FOR XI ' !
  $PRINT %X DEVX !
  $PRINT 'CONFIDENCE INTERVAL FOR ALPHA'!
  $PRINT %T %E !
  $PRINT 'CONFIDENCE INTERVAL FOR P'!
  $PRINT %R %Q !
  $ENDMAC !
!
$RETURN !
$FINISH !
```


ACKNOWLEDGEMENTS

We are grateful to Roberto Westenberger, Centro de Estudos e Pesquisas em Seguros, for putting the data at our disposal, to Gabriel Yáñez Canal for computational assistance, and to Vibeke Thinggård and the referee for some useful comments on the paper. The research was done at Instituto de Matemática Pura e Aplicada, Rio de Janeiro.

REFERENCES

Andrade e Silva, J. M. (1989). An application of generalized linear models to Portuguese motor insurance. *XXI ASTIN Colloquium* 633–649.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. Wiley, Chichester.

Barndorff-Nielsen, O. E. (1988). *Parametric statistical models and likelihood. Lecture Notes in Statistics 50*, Springer-Verlag, Berlin.

Chang, L. & Fairly, W. B. (1978). Pricing automobile insurance under multivariate classification of risks: additive versus multiplicative. *J. Inst. Actuaries Student's Soc.* **22,** 75–98.

Coe, R. & Stern, R. D. (1982). Fitting models to daily rainfall data. *J. Appl. Meterology* **21,** 1024–1031.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39,** 1–38.

Hallin, M. & Ingenbleek, J.-F. (1983). The Swedish automobile portfolio in 1977. A Statistical study. *Scand. Actuarial J.* 49–64.

Jones, M. C. (1987). On the relationship between the Poisson-exponential model and the non-central chi-squared distribution. *Scand. Actuarial J.* 104–109.

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc. Ser. B* **49,** 127–162.

Jørgensen, B. (1992). Exponential dispersion models and extensions: A review. *Internat. Statist. Rev.* **60,** 5–20.

McCullagh, P. & Nelder, J. A. N. (1989). *Generalized linear models*, 2nd edn. Chapman and Hall, London.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135,** 370–384.

Siegel, A. F. (1985). Modelling data containing exact zeroes using zero degrees of freedom. *J. Roy. Statist. Soc. Ser. B* **47,** 267–271.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and new directions. Proceedings of the Indian Statistical Golden Jubilee International Conference* (Eds. J. K. Ghosh & J. Roy), pp. 579–604. Indian Statistical Institute, Calcutta.

Yáñez Canal, G. (1992). *Estimação na Familia Tweedie e uma Aplicação a Seguros de Automóveis.* (Estimation in the Tweedie Family and an Application in Automobile Insurance). Masters dissertation (in Portuguese), Instituto de Matemática Pura e Aplicada, Rio de Janeiro.

Revised version received December 1993

Address for correspondence:
Department of Statistics
University of British Columbia
2021 West Mall
Vancouver B.C.
Canada V6T 172

Centro de Estudos e Pesquisas em Seguros
COPPEAD
Ilha do Fundão
Caixa Postal 68514
21945 Rio de Janeiro RJ
Brazil