REGRESSION ANALYSIS OF QUANTITY DATA WITH EXACT ZEROES*

Gordon K. Smyth
Department of Mathematics, University of Queensland,
Brisbane, Australia

Abstract

Measurements of the magnitude or duration of physical phenomenon have the property that they are positive and continuous, except for the possibility of exact zeroes when the phenomenon does not occur. Such data cannot be transformed to normality by power transformations or any other means, and special treatment of the zero observations is usually required. The approach of this paper is to model quantity data using a family of exponential family distributions intermediate between the Poisson and the gamma families. These families have the feature of power mean-variance relationships with exponent between one and two. Regression modelling is possible using the established framework of generalized linear models. With suitable assumptions this approach allows the information in both the zero and positive observations to contribute to the estimation of all parts of the model.

1 Introduction

McCullagh and Nelder (1989) discuss four distributions, namely the normal, Poisson, gamma and inverse-Gaussian, which fit into the generalized linear model framework and which have variances proportion to some power of the mean, i.e., for which $\text{var}(Y) = \phi E(Y)^\theta$ for some ϕ and θ. For these four distributions θ is 0, 1, 2 and 3 respectively. Jørgensen (1987) gave a more general definition of a generalized linear model distribution and showed that in fact any value outside the interval $(0, 1)$ is possible for θ. Those distributions with $1 < \theta < 2$ turn out to be Poisson mixtures of gamma distributions (Tweedie, 1984, p. 592; Jørgensen, 1987, p. 140). They are continuous and positive, except for an atom at zero.

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

The distribution with $\theta = 1.5$ is the noncentral chi-squared distribution on zero degrees of freedom discussed by Siegel (1979, 1985).

Data with exact zeros and power mean-variance relationships with $1 < \theta < 2$ are common, for example threshold models, weather variables such as wind speed, rainfall, snowfall, and population size (Perry 1981, 1985). The author's interest in the problem was motivated by an experiment, conducted by the Department of Child Health, University of Queensland, to compare the effects of six sleeping positions on gastro-esophageal reflux (GER) in infants (Ting et al, 1993). Esophageal pH level, blood oxygen saturation, heart rate and apnoea were monitored for each of twenty infants with a history of GER during eight hour nights. The infants were placed in one sleeping position for the first half of the night and another position for the second half. Summary pH measurements were calculated for each two hour period. The four quarters of each night were not found to differ significantly with respect to pH levels, and sleeping positions were randomly assigned, so the experiment has the structure of an unbalanced randomized block experiment with block size two. Children are considered to be suffering reflux when the pH-level in their esophagus falls below four. The reflux variables considered most meaningful by medical researchers are the percentage time in reflux, called the reflux index (RI), and the area between the pH trace and the pH=4 line (AU4). The latter variable is the value of (4−pH level) integrated over the time spent in episodes. Ting et al (1993) found the sleeping positions to be significantly different, with prone generally the best and supine or inclined supine the worst, with respect to presence/absence of reflux, mean pH, number of reflux episodes, and oxygen saturation. However a full information analysis of RI and AU4 was not possible because these variables have mixed distributions being continuous except for the positive probability of exact zeros. In this paper the use of the Poisson-gamma generalized linear models is investigated for RI and AU4.

Although Jørgensen (1987, 1992) has used a Poisson-gamma generalized linear model to analyse the amount spent by Amazonian peasants on hiring outside labour power, there are a number of outstanding problems. One is that there has been no satisfactory way to estimate the index parameter θ. See discussion by Gilchrist (1987) and Burridge (1987) of Jørgensen's 1987 paper. There is no apriori reason to prefer one value over another for θ is the interval $(0, 1)$ in most applications. Another problem is that there has been no way to check the Poisson-gamma distributional form in a regression context. In this paper it is shown that maximum likelihood can be satisfactorily applied to estimate θ, and the provision of the explicit likelihood function opens to way to check the distributional form using quantile residuals as defined by Dunn and Smyth (1996).

2 Poisson-gamma generalized linear models

Let $f(y;\mu,\phi,\theta)$ be the probability density function of a univariate random variable Y, and suppose that $E(Y)=\mu$ and $\mathrm{var}(Y)=\phi\mu^{\theta}$. We seek a density which satisfies

$$\frac{\partial \log f}{\partial \mu} = \frac{y-\mu}{\phi\mu^{\theta}} \tag{1}$$

since this form for the derivative characterizes generalized linear model distributions (Smyth, 1991). Integrating, f must satisfy

$$\log f = \int^{\mu} \frac{y-\eta}{\phi\eta^{\theta}} d\eta = \frac{1}{\phi}\left(y\frac{\mu^{1-\theta}}{1-\theta} - \frac{\mu^{2-\theta}}{2-\theta}\right) + c(y,\phi,\theta)$$

where $c(y,\phi,\theta)$ is some function not depending on μ. It is a condition here that θ is not equal to 1 or 2. Writing $\nu=\mu^{1-\theta}/(1-\theta)$ and $\kappa(\nu)=[(1-\theta)\nu]^{(2-\theta)/(1-\theta)}/(2-\theta)$, this is

$$\log f = \frac{1}{\phi}(y\nu-\kappa(\nu)) + c(y,\phi,\theta)$$

which, for given θ, is of the form required for a generalized linear model distribution (McCullagh and Nelder, 1989). Note that $\dot\kappa(\nu)=\mu$ and $\ddot\kappa(\nu)=\mu^{\theta}$ in accordance with the usual generalized linear model theory.

The moment generating function of Y is

$$\begin{aligned} M_Y(t) &= \int \exp\{\frac{1}{\phi}[y(\nu+t\phi)-\kappa(\nu)] + c(y,\phi,\theta)\}dy \\ &= \exp\{\frac{1}{\phi}[\kappa(\nu+t\phi)-\kappa(\nu)]\} \end{aligned}$$

so the cumulant generating function is

$$\begin{aligned} \log M_Y(t) &= \frac{1}{\phi}[\kappa(\nu+t\phi)-\kappa(\nu)] \\ &= \frac{1}{\phi}\frac{\mu^{2-\theta}}{2-\theta}[(1+t\phi(1-\theta)\mu^{\theta-1})^{(2-\theta)/(1-\theta)}-1] \end{aligned}$$

This can be compared with the cumulant generating function of $Z=X_1+\ldots+X_N$, where N is Poisson(λ) and, conditional on N, the X_i are independent gamma(α,τ), which is

$$\log M_Z(t) = \lambda[(1-\tau t)^{-\alpha}-1]$$

Note that Z is a Poisson mixture of gamma distributions since Z given N is gamma($N\alpha$,τ). We see by identifying terms in the cumulant generating functions that Y has the same distribution as Z with

$$\lambda = \frac{1}{\phi}\frac{\mu^{2-\theta}}{2-\theta}, \quad \alpha = \frac{2-\theta}{\theta-1}, \quad \tau = \phi(\theta-1)\mu^{\theta-1}$$

The requirement that the gamma shape parameter α be positive means that the representation of Y as a Poisson mixture of gamma random variables is valid only for $1 < \theta < 2$. Note that $\lambda > 0$ and $\tau > 0$ imply that $\mu > 0$ and $\phi > 0$ also.

The density function f can now be written as

$$\begin{aligned} f(y;\mu,\phi,\theta) &= P(N=0)d_0(y) + \sum_{j=1}^{\infty} P(N=j) f_{Z|N=j}(y) \\ &= e^{-\lambda} d_0(y) + \sum_{j=1}^{\infty} \frac{\lambda^j e^{-\lambda}}{j!} \frac{y^{j\alpha-1} e^{-y/\tau}}{\tau^{j\alpha}\Gamma(j\alpha)} \end{aligned}$$

where d_0 is the Dirac delta function at zero and $f_{Z|N}$ is the conditional density of Z given N. Therefore

$$\log f = \begin{cases} -\lambda + \log d_0(y) & y = 0 \\ -y/\tau - \lambda - \log y + \log W(y,\lambda,\alpha,\tau) & y > 0 \end{cases}$$

where

$$W(y,\lambda,\alpha,\tau) = \sum_{j=1}^{\infty} \frac{\lambda^j (y/\tau)^{j\alpha}}{j!\Gamma(j\alpha)}$$

Tweedie (1984, p. 586) has identified W as an instance of Wright's (1933) generalized Bessel function. The function is not expressible however in terms of the more common Bessel functions.

Feller (1968) and Jørgensen (1987) call the distribution of Y compound Poisson, while Johnson and Kotz (1971) call distributions of this type compound gamma. Here it will be called simply Poisson-gamma in recognition of its various characterizations as a Poisson mixture of gammas, as a Poisson sum of gammas, or as an exponential family intermediate between the Poisson and gamma families. The Poisson-gamma family intersects the noncentral χ^2 family at $\theta = 1.5$. The $\chi^2_\nu(2\lambda)$ distribution can be expressed as the mixture of gamma$(N + \nu/2, 2)$ distributions where N is Poisson(λ), so the Poisson-gamma distribution with $\alpha = 1$ and $\beta = 2$ is $\chi^2_0(2\lambda)$. As $\theta \uparrow 2$ the distribution approaches a gamma$(\alpha'$,$\tau')$ distribution with $\alpha' = 1/\phi$ and $\tau' = \phi\mu$. As $\theta \downarrow 1$ the distribution of Y/ϕ approaches Poisson(λ).

3 Parameter Estimation

In generalized linear model applications observations $Y_1, \ldots, Y_n$ will be assumed independent with common ϕ and θ and with means μ_i which satisfy a link-linear model

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

where the $\mathbf{x}_i$ are vectors of covariates and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients.

For any given value of θ, maximum likelihood estimates of $\boldsymbol{\beta}$ and an unbiased estimate of ϕ can be calculated as for a generalized linear model, for example using the `$OWN` directive of GLIM (Payne, 1985) or using the `make.family` function of S-Plus. As always in a generalized linear model, the maximum likelihood estimate of $\boldsymbol{\beta}$ does not depend on the dispersion parameter ϕ, but it does depend on θ. The scoring iteration for $\boldsymbol{\beta}$ can be written as

$$\boldsymbol{\beta}^{k+1} = (X^T W X)^{-1} X^T W \mathbf{z}$$

where $W = \text{diag}(\dot{g}(\mu_i)^2 \mu_i^\theta)^{-1}$, $\mathbf{z}$ is a working vector with components $\dot{g}(\mu_i)(y_i - \mu_i) + g(\mu_i)$, and all terms on the right-hand side are evaluated at the current estimate $\boldsymbol{\beta}^k$. The iteration may be started at $\mu_i = y_i$, and converges reliably to the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ for most link functions.

It is apparent from (1) that $\partial^2 \log f / \partial\phi\partial\mu$ and $\partial^2 \log f / \partial\theta\partial\mu$ have expectation zero so that μ is orthogonal to both ϕ and θ. Since the likelihood depends on $\boldsymbol{\beta}$ only through the μ_i, it is also true that $\boldsymbol{\beta}$ is orthogonal to ϕ and θ. One consequence of this is that the standard errors for $\boldsymbol{\beta}$ obtained from the above generalized linear model will be correct Fisher information standard errors even if ϕ and θ have been estimated from the same data.

Maximum likelihood estimators of ϕ and θ can be obtained by directly maximizing the profile likelihood, by the pseudo-likelihood approach of Davidian and Carroll (1987), or by the extended quasi-likelihood approach of Nelder and Pregibon (1987). Given estimated values for $\boldsymbol{\beta}$ and θ, an unbiased estimate of ϕ can obtained from

$$\tilde{\phi} = \sum_{i=1}^{n} \frac{[y_i - \mu_i(\hat{\boldsymbol{\beta}})]^2}{\mu_i(\hat{\boldsymbol{\beta}})^\theta}$$

Given $\hat{\theta}$, this is essentially equivalent to the pseudo-likelihood estimate.

4 Examples

Data for which the variance increases with the mean more rapidly than direct proportionality but less rapidly than the mean squared is common. Consider the wind speed data analysed by Haslett and Raftery (1989) consisting of daily mean wind speeds at 12 meteorological stations in Ireland during the period 1961–1978.

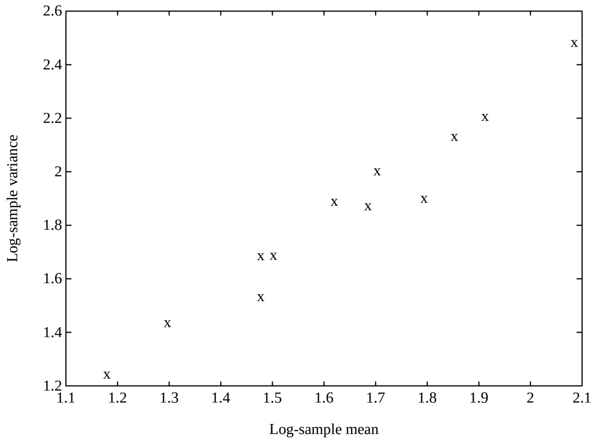

Figure 1: Plot of log-sample variance versus log-sample mean for daily mean wind speeds at 12 sites in Ireland.

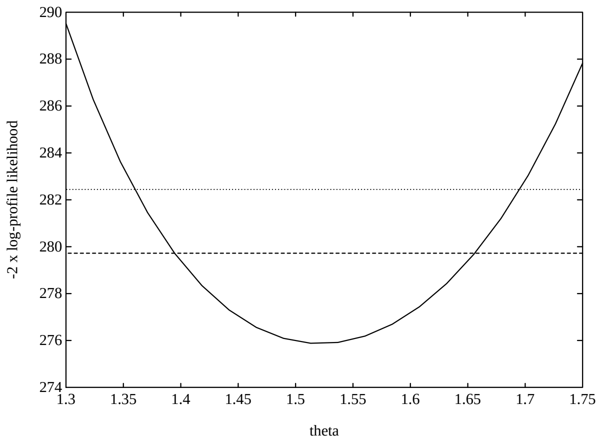

Figure 2: Profile plot of θ for January snowfall in Seattle, 1906 to 1960. The dashed and dotted lines indicate 95% and 99% confidence intervals.

Figure 1 shows that relationship between log-sample variance and log-sample mean over the 12 sites is strikingly linear. The slope of the least squares line is 1.30, suggesting that the Poisson-gamma distribution may be appropriate here. Sixteen out of 78888 observations are exactly zero for this data set.

The maximum likelihood approach allows estimation of θ for single samples and other data sets for which there is not a wide range of values for μ. Seigel (1985) considered January snowfall in Seattle, in inches, for the years 1906 to 1960. Figure 2 gives a profile likelihood plot for θ in this single sample problem, with nominal 95% and 99% confidence bands. Seigel effectively assumed $\theta = 1.5$; this value is close to the centre of the confidence intervals for θ, justifying his analysis.

An experiment conducted by Joseph Ting in the Department of Child Health,

Table 1: Minus twice the log-profile likelihood for the reflux experiment with and without sleeping position as a factor.

θ	Child	Child+Position	Difference
1.1	869.0505	836.2950	32.7554
1.2	699.5124	677.8852	21.6272
1.3	632.6674	615.7370	16.9304
1.4	599.2437	584.9011	14.3426
1.5	581.9255	569.3087	12.6168
1.6	574.5833	563.3471	11.2362
1.7	575.2719	565.4217	9.8502
1.8	585.4402	577.3253	8.1149
1.9	615.3785	609.8962	5.4822

University of Queensland, provides a data set with several factors. Children are considered to be suffering reflux if the ph-level in their esophagus falls below 4, and this experiment compared the effect on reflux of six sleeping positions. Twenty babies of a few months age with a history of reflux had their ph-levels recorded during an 8 hour night. Two responses were tried. The first was $-\log(1-\mathrm{RI})$, where RI is the percentage time in reflux. The second was $-\log(1-\mathrm{AU4}/\mathrm{Duration\ times\ 4})$. The response variable was the amount by which ph-level fell below 4, integrated over the time spent in reflux, and this was recorded for each 2 hour period. Each baby was assigned to two sleeping positions, one in the first half and one in the second half of the night. A Poisson-gamma generalized linear model was fitted with child as a 20 level factor and sleeping position as a 6 level factor. The log-link was used, and the logarithm of the exact duration of each of the nominal two hour periods was set as offset. Factors comparing the four 2 hour periods during the night were found to be unimportant. Four children who experienced no reflux during the night in either position were excluded from the analysis. Table 1 gives minus twice the log-profile likelihood (excluding the Dirac delta term) with and without sleeping position for various values of θ. The likelihood is maximized by θ about 1.6 both with and without sleeping position. Looking at a finer grid of θ values in Table 2, the likelihood is maximized without position at 1.64, with position at 1.62, and the likelihood ratio test for sleeping position allowing estimation of θ is about 10.8. As a χ^2 variable on 5 degrees of freedom this corresponds to a p-value of 0.055. Such a result from such a small study suggests further investigation.

Table 2: Minus twice the log-profile likelihood for the reflux experiment with and without sleeping position as a factor.

θ	Child	Child+Position	Difference
1.60	574.5833	563.3471	11.2362
1.61	574.2969	563.1937	11.1032
1.62	574.0880	563.1183	10.9697
1.63	573.9567	563.1211	10.8355
1.64	573.9031	563.2028	10.7004
1.65	573.9279	563.3640	10.5639
1.66	574.0320	563.6062	10.4258
1.67	574.2165	563.9309	10.2857
1.68	574.4833	564.3400	10.1433
1.69	574.8342	564.8360	9.9982
1.70	575.2719	565.4217	9.8502

Acknowledgements

The author also wishes to thank Joseph Ting and Dr John Vance of The University of Queensland for permission to use the reflux data.

References

Burridge, J. (1987). Discussion of Dr Jørgensen's paper. *J. R. Statist. Soc. B*, **49**, 150–151.

Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *J. Amer. Statist. Ass.*, **82**, 1079–91.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *J. Computat. Graph. Statist.*, to appear.

Gilchrist, R. (1987). Discussion of Dr Jørgensen's paper. *J. R. Statist. Soc. B*, **49**, 145–147.

Haslett, J. and Raftery, A. E. (1989). Space-time modelling with long-memory dependence: assessing Ireland's wind power resource. *Appl. Statist.*, **38**, 1–50.

Jørgensen, B. (1987). Exponential dispersion models. *J. R. Statist. Soc. B*, **49**, 127–162.

Jørgensen, B. (1992). *The theory of exponential dispersion models and analysis of deviance*. Monografias de Matemátika No. 51, Instituto de Mathemátika pura e Aplicada, Rio de Janeiro.

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

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–231.

Payne, C. D. (1985). *The GLIM System Release 3.77 Manual.* Numerical Algorithms Group: Oxford.

Perry, J. N. (1981). Taylor's power law for dependence of variance on mean in animal populations. *Appl. Statist.*, **30**, 254–263.

Perry, J. N. (1985). Adès: new ecological families of species-specific frequency distributions that describe repeated spatial samples with an intrinsic power-law variance-mean property. *J. Animal Ecology*, **54**, 931–953.

Seigel, A. F. (1979). The noncentral chi-squared distribution with zero degrees of freedom and testing for uniformity. *Biometrika*, **36**, 707–19.

Seigel, A. F. (1985). Modelling data containing exact zeroes using zero degrees of freedom. *J. Roy. Statist. Soc. B*, **47**, 267–71.

Smyth, G. K. (1991). Exponential dispersion models and the Gauss-Newton algorithm. *Austral. J. Statist.* 33: 57–64.

Ting, J., Smyth, G. K., Cleghorn, G., Masters, B., and Vance, J. C. (1993). The relationship between gastro-esophageal reflux and sleeping position in infants (implication for suddent infant death syndrome). Technical Report, University of Queensland.

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 Institute Golden Jubilee International Conference.* (Eds. J. K. Ghosh and J. Roy), pp. 579–604. Calcutta: Indian Statistical Institute.

Wright, E. M. (1933). On the coefficients of power series having essential singularities. *J. London Math. Soc.*, **8**, 71–9.