Fitting Tweedie's Compound Poisson Model to Insurance Claims Data: Dispersion Modelling

Gordon K. Smyth and Bent Jørgensen
Department of Statistics and Demography, University of Southern Denmark, Denmark
gks@statdem.ou.dk and bentj@statdem.ou.dk

Introduction

We reconsider the problem considered by Jørgensen and de Souza (1994), namely that of producing fair and accurate tariffs based on aggregated insurance data giving numbers of claims and total costs for the claims. Jørgensen and de Souza (1994) assumed Poisson arrival of claims and gamma distributed costs for individual claims. These assumptions imply that the total cost of claims in each category over a given time period follows a Tweedie compound Poisson distribution. Jørgensen and de Souza (1994) directly modelled the parameter of interest, namely the risk or expected cost of claims per insured unit, μ say. They observed that the dependence of the likelihood function on μ is as for a linear exponential family, so that modelling similar to that of generalized linear models is possible.

In this paper we observe that it is generally necessary to model the dispersion of the costs as well as their mean. In order to model the dispersion we use the framework of double generalized linear models developed by Smyth (1989) and Smyth and Verbyla (1999). Modelling the dispersion increases the precision of the estimated tariffs. This approach has the added benefit that the case where only the total cost of claims and not the number of claims has been recorded can be handled with no extra complication.

The method used by Jørgensen and de Souza (1994) implicitly assumes that explanatory variables affect the expected cost of claims μ by simultaneously increasing or decreasing both the frequency of claims and the average claim size. In practice however, some explanatory factors will have a greater impact on the frequency of claims than on their size, while other variables may impact more on the size of claims. It is also possible for certain factors, such as a no-claims bonus, to affect the frequency of claims and the claim size on opposite directions. This does not invalidate the method used by Jørgensen and de Souza (1994), but it does mean that insurance claims data is likely to display non-constant dispersion, so that it is necessary to model the dispersion in order to obtain efficient estimation of μ.

Double generalized linear models allow the simultaneously modelling of both the mean and the dispersion in generalized linear models. Estimation of the dispersion is effected by a second generalized linear model, the dispersion submodel, in which the responses are the unit deviances from the original model. The unit deviances are distributed approximately as $\phi_i \chi_1^2$, where ϕ_i is the dispersion for the ith observation, so the dispersion submodel has gamma responses. When modelling insurance data with counts of claims as well as total costs, we use the same double generalized linear model framework, but modify the definition of the response and the weights in the dispersion submodel. When only the total claim costs are observed and not the claim counts, the definitions of the response and weights in the dispersion submodel revert to their customary values.

Tweedie's Compound Poisson Model

Let N_i be the number of claims observed in the ith category and Z_i be the total claim size.

Suppose that the number of units at risk (perhaps measured in policy years) is t_i, and write $Y_i = Z_i/t_i$ for the claim per unit at risk. We suppose that N_i is Poisson distributed with mean $\lambda_i t_i$, and that the amount of each claim is gamma distributed with mean τ_i and shape parameter α. The probability that N_i and Y_i are zero is $e^{-\lambda_i}$. Jørgensen and de Souza (1994) observed that the parameter is interest is $\mu_i = E(Y_i) = \lambda_i \tau_i$. From Jørgensen (1987, 1997) it is known that Y_i follows a Tweedie linear exponential family as μ_i varies, with variance function $V(\mu_i) = \mu_i^p$ where $p = (\alpha+2)/(\alpha+1)$ is between 1 and 2. It follows that $\mathrm{var}(Y_i) = \phi_i \mu_i^p$ where ϕ_i is a dispersion parameter. It is easy to show that $\phi_i = \lambda_i^{1-p}\tau_i^{2-p}/(2-p)$. The joint density of N_i and Y_i can be parametrized in terms of μ_i, ϕ_i and p instead of λ_i, τ_i and α. This parametrization has the advantage that it focuses attention of the parameter of interest and two other parameters which are orthogonal to it.

It can be seen that any factor which increases the frequency of claims λ_i without affecting their size will increase the mean μ_i and decrease the dispersion ϕ_i. Any factor which increases the average claim size τ_i without increasing their frequency will increase both the mean and the dispersion. A factor which affects the mean but not the dispersion must affect λ_i and τ_i is such as way that $\lambda_i^{1-p}\tau_i^{2-p}$ remains constant.

We assume link-linear models $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ and $h(\phi_i) = \mathbf{z}_i^T\boldsymbol{\gamma}$ where $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of covariates and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients. If g and h are logarithmic, then this is equivalent to log-linear modelling of the frequency and claim size separately. We assume that α and hence p does not vary between cases. We estimate $\boldsymbol{\beta}$ by maximum likelihood from the marginal density of the observed costs y_i, this being sufficient for $\boldsymbol{\beta}$. We estimate $\boldsymbol{\gamma}$ using approximate REML from the joint likelihood of the y_i and the observed counts n_i. Both steps are implemented as generalized linear models. The method has been implemented as an S-Plus function `tariff` available from the URL `http://www.maths.uq.edu.au/~gks/s/`.

REFERENCES

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

Jørgensen, B. (1997). *Theory of Dispersion Models.* Chapman and Hall, London.

Jørgensen, B., and de Souza, M. C. P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 69–93.

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

Smyth, G. K., and Verbyla, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics.* To appear.

RÉSUMÉ

Nous reconsidérons le problème de produire des tarifs précis basés sur des données agrégées d'assurance donnant des nombres de réclamations et de coûts totaux pour chaque catégorie. Nous assumons l'arrivée de Poisson des réclamations et des coûts distribués par gamma pour différentes réclamations. Ces prétentions impliquent que tout le coût de réclamations dans chaque catégorie sur une période indiquée de temps suit un distribution Tweedie Poisson composé. Nous prouvons qu'il est généralement nécessaire de modeler la dispersion des coûts aussi bien que leur moyen. Ceci est fait en utilisant le cadre des modèles linéaires généralisés doubles.