Generalised linear models for aggregate claims; to Tweedie or not?

Oscar Quijano and José Garrido*
Concordia University, Montreal, Canada

(First version of May 20, 2013, this revision December 2, 2014)

Abstract

The compound Poisson distribution with gamma claim sizes is a very common model for premium estimation in Property and Casualty insurance. Under this distributional assumption, generalised linear models (GLMs) are used to estimate the mean claim frequency and severity, then these estimators are simply multiplied to estimate the mean aggregate loss.

The Tweedie distribution allows to parametrise the compound Poisson-gamma (CPG) distribution as a member of the exponential dispersion family and then fit a GLM with a CPG distribution for the response. Thus, with the Tweedie distribution it is possible to estimate the mean aggregate loss using GLMs directly, without the need to previously estimate the mean frequency and severity separately.

The purpose of this educational note is to explore the differences between these two estimation methods, contrasting the advantages and disadvantages of each.

1 Introduction

In recent years there have been remarkable developments in insurance modelling with dependence between frequency and severity. There are situations in which considering them independent is not unreasonable. When this is

*Corresponding author: jose.garrido@concordia.ca. The authors gratefully acknowledge the partial financial support of NSERC grant 36860-2012.

the case, the compound Poisson distribution with gamma claim sizes is a common model used for the aggregated loss. A common way of fitting the parameters of this distribution given a sample is using GLMs to estimate the mean claim frequency and severity separately, then these estimators are simply multiplied to estimate the mean aggregate loss.

The Tweedie distribution is a re-parametrisation of the compound Poisson-gamma (CPG) distribution as a member of the exponential dispersion family, which allows to fit a GLM with a CPG distribution for the response. With the Tweedie distribution one can hence estimate the mean aggregate loss using GLMs directly, without the need to previously estimate the mean frequency and mean severity. The main purpose of this article is to explore the differences between these two estimation methods.

Section 2 gives a review of exponential dispersion families, the basic theory of generalised linear models and Tweedie distributions. Section 3 we discusses the use of GLMs for pure premium estimation. In this section we outline the main differences between the separated approach and the Tweedie GLM. Section 4 discusses the limitations of the Tweedie GLM. We exemplify this limitations with a simulated dataset. In Section 5 we fit both models to a publicly available dataset and we compare their results. Section 6 comments on modifications to the Tweedie GLM that help overcome its limitations. More advanced models that consider dependence are commented and cited in this section.

The R code used to generate the simulations, graphs and tables in this article can be found in `http://totweedieornot.sourceforge.net`, together with the programs documentation.

2 Generalised Linear Models

2.1 Exponential Dispersion Family

The exponential dispersion family (EDF) is a set of probability distributions used as the response distribution for GLMs. This section gives a brief introduction to the EDF and some of its main properties; for a more extensive presentation see Jørgensen [7].

An EDF is a set of distributions with densities of the form

$$f(x;\theta,\lambda) = c(x,\lambda)\exp(\lambda\left\{\theta x - \kappa(\theta)\right\}), \qquad \theta\in\Theta, \lambda\in\Lambda, \tag{1}$$

where Θ, called the canonical space, is an open interval that contains 0, $\Lambda \subset (0, \infty)$ is called the index set, θ and λ are called the canonical and index parameters, respectively, and κ is a smooth function.

Many well-known discrete and continuous distributions can be parametrised in this way. Table 1 shows some popular members of the EDF and gives its parameters in terms of the usual density/probability function parameters.

Distribution	θ	λ	Θ	Λ	$\kappa(\theta)$
Binomial(n, p)	$\ln\left(\frac{p}{1-p}\right)$	n	$\mathbb{R}$	$\mathbb{N}$	$\ln\left(\frac{1+\exp(\theta)}{2}\right)$
Poisson(μ_N)	$\ln(\mu_N)$	$--$	$\mathbb{R}$	$--$	$\exp(\theta) - 1$
Gamma(α, τ)	$1 - \tau$	α	$(-\infty, 1)$	$\mathbb{R}$	$\ln\left(\frac{1}{1-\theta}\right)$
N(μ, σ^2)	μ	σ^2	$\mathbb{R}$	$(0, \infty)$	$\frac{1}{2}\theta^2$

Table 1: EDF parametrisation of some well-known distributions

If X is a random variable with density/probability function as in (1), then its moment generating function (mgf) is given by

$$m(t) = \exp\left(\lambda\left[\kappa(\theta + t/\lambda) - \kappa(\theta)\right]\right), \qquad t \in \Theta_{\lambda,\theta} \tag{2}$$

where $\Theta_{\lambda,\theta} := \lambda(\Theta - \theta) = \{\theta^* : \theta^* = \lambda(\theta_0 - \theta) \text{ for some } \theta_o \in \Theta\}$. By taking derivatives of m with respect to t, it is possible to see that

$$\mathbb{E}[X] = \kappa'(\theta) \qquad \text{and} \qquad \mathbb{V}[X] = \frac{\kappa''(\theta)}{\lambda}\,. \tag{3}$$

From the relations above we see that the mean and the variance depend on the derivatives of κ, called the cumulant generator of the family, which is twice continuously differentiable with $\kappa'' > 0$. Hence the variance of X can be expressed in terms of its mean through a variance function.

If we denote by $\mu = \mathbb{E}[X]$ and the function $\tau = \kappa'$, then the variance function is defined as

$$\mathbf{V}(\mu) = (\kappa'' \circ \tau^{-1})(\mu)\,,$$

and from (3) it follows that $\mathbb{V}[X] = \frac{\mathbf{V}(\mu)}{\lambda}$. An important property of the variance function is that it characterises the EDF; this property is used for the construction of the Tweedie distribution.

2.2 Generalised Linear Models (GLMs)

GLMs provide a practical methodology for the segmentation of a portfolio of policy-holders. Here we review the main concepts and assumptions of GLMs.

A standard assumption in insurance practice is that there is a set of observable and quantifiable risk characteristics (for example: policy-holder age, neighbourhood, smoker) that allow to segment the population into groups of homogeneous risks. In a GLM context, these characteristics are represented by a vector $X = (X_1, X_2, \ldots, X_k)$. These $X_1, \ldots, X_k$ are called covariates or explanatory variables. As with classical regression models, the X_i's can be categorical or continuous.

Another element present in GLMs is a weight that is given to each observation. We explain its role with an example; suppose that we are interested in modelling the annual aggregate loss for a portfolio of insurance policies that is already divided into homogeneous groups. Now consider an individual who cancels his/her policy 6 months after the beginning of the contract. It is reasonable to believe that his/her aggregate loss does not follow the same distribution as that of a policy-holder exposed to risk the whole year. In this case we assign this observation a weight of 0.5 years. We denote such weights W.

Finally, the variable to be forecasted, e.g. frequency, severity, or pure premium, is denoted Y and called the response variable.

As usual we use $\boldsymbol{x}$, w and y to denote observations from X, W and Y, respectively.

We have the following distributional assumptions for GLMs:

1. Given $X = \boldsymbol{x}$ and $W = w$, the distribution of Y belongs to some fixed EDF, i.e.

$$f_{Y|X,W}(y|\boldsymbol{x}, w) = c(y, \lambda) \exp(\lambda \{\theta y - \kappa(\theta)\}) . \tag{4}$$

2. There exists ϕ such that for every $\boldsymbol{x}$ and w, the index parameter is given by

$$\lambda = \frac{w}{\phi} .$$

   This implies that λ varies only with the value of W. Here ϕ is called the dispersion parameter.

3. Let $\mu_{\boldsymbol{x}} =: \mathbb{E}[Y|X = \boldsymbol{x}, W = 1]$ be the expected response for explanatory variables $X = \boldsymbol{x}$ and weight $W = 1$. There exists a fixed vector

$\boldsymbol{\beta} = (\beta_1, ..., \beta_k) \in \mathbb{R}^k$ and a one-to-one differentiable function h such that

$$\mu_{\boldsymbol{x}} = h(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}).$$

If g denotes the inverse of h, then writing the last equation in terms of g gives

$$g(\mu_{\boldsymbol{x}}) = \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}\,,$$

where g is called the link function, it relates the linear explanatory term $\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}$ to the expected response $\mu_{\boldsymbol{x}}$.

In practice, both the response distribution and the link function need to be chosen. There is substantial flexibility for the choice of g since it can be any one-to-one function. Nevertheless, in insurance practice a log-link function $g = \ln$ is usually used, since it results in a multiplicative rating structure (see Ohlsson and Johansson [10, Section 1.3]).

The parameters $\boldsymbol{\beta}$ and ϕ of the GLM are estimated from sample values. Their maximum likelihood estimators (MLEs) are found using numerical methods. Popular software, like R or SAS, have pre-programmed packages for this.

2.3 The Tweedie Families of Distributions

The Tweedie families are EDFs that are defined through the variance function. An EDF is called a *Tweedie Family* if the domain of its variance function $\mathbf{V}$ is $(0, \infty)$ with

$$\mathbf{V}(\mu) = \mu^p,$$

for some $p \in \mathbb{R}$.

The Tweedie families contain many distributions, characterised by the value of p. Table 2 presents the well known distributions that can be seen as a Tweedie family for different values of p.

In addition, it is known that for $p < 0$ the Tweedie families characterise distributions that are supported on $\mathbb{R}$, while for $p > 1$ it characterises distributions that are supported on $(0, \infty)$. For $p \in (0, 1)$ it is know that there is no EDF with such variance function power. Here we focus on the case $p \in (1, 2)$.

An EDF can be parametrised also in terms of its mean instead of its canonical parameter. Using this parametrisation, we denote by $Tw(p, \mu, \lambda)$

Value of p	**Distribution**
$p = 0$	normal
$p = 1$	Poisson
$p \in (1, 2)$	compound Poisson - gamma
$p = 2$	gamma
$p = 3$	inverse Gaussian

Table 2: Tweedie distributions for different values of p

a Tweedie distribution with variance function exponent p, mean μ and index parameter λ. We will denote by $CPG(\mu_N, \alpha, \tau)$ a compound Poisson distribution with Poisson rate μ_N and jump size distribution gamma(α, τ), i.e $CPG(\mu_N, \alpha, \tau)$ represents the distribution of a random variable S with

$$S = \sum_{i=0}^{N} Y_i \,,$$

where $Y_0 = 0$, N follows a Poisson(μ_N) distribution, $Y_1, Y_2, \ldots$ is a iid gamma(α,τ) sequence independent from N. The form of the gamma density used here is

$$f(x) = \frac{\tau^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} \exp(-\tau x)\,, \qquad x > 0\,,$$

where $\alpha > 0$ is called the shape parameter and $\tau > 0$ the rate parameter.

To see that for $p \in (1, 2)$ the Tweedie family corresponds to a compound Poisson-gamma simply compare the moment generating functions of both distributions (for details see Jørgensen [7]). This way it is also possible to express one parametrisation in terms of the other. In fact, if $p \in (1, 2)$, $\mu > 0$ and $\lambda > 0$, then

$$Tw(p, \mu, \lambda) = CPG\left(\frac{\lambda\mu^{2-p}}{2-p}, -\frac{p-2}{p-1}, \frac{\lambda\mu^{1-p}}{p-1}\right)\,. \tag{5}$$

Similarly, for $m, \alpha, \tau > 0$,

$$CPG(\mu_N, \alpha, \tau) = Tw\left(\frac{\alpha+2}{\alpha+1}, \frac{\mu_N\alpha}{\tau}, \frac{(\mu_N\alpha)^{\frac{\alpha+2}{\alpha+1}-1}\tau^{2-\frac{\alpha+2}{\alpha+1}}}{\alpha+1}\right)\,. \tag{6}$$

3 Modelling P/C Premiums Using GLMs

This section starts with a review of some basic risk theory definitions.

The *duration* of a policy is the amount of time a policy is in force. It is usually measured in years.

Following the terminology in Ohlsson and Johansson [10], we define the following three key ratios:

- The *claim frequency* is the number of claims divided by the duration, i.e. the average number of claims per unit time.
- The *claim severity* is the total claims divided by the number of claims, i.e. the average size per claim.
- The *pure premium* is the total claims divided by the duration, i.e. the average claim amount paid per unit time.

GLMs can be used to estimate the three quantities above. Notice that all three are divided by a volume measure. This volume measure, called the exposure, is used as a weight in the GLM. Insurance is justified by the strong law of large numbers. For this reason it is desirable to have many observations in each class (a class is defined by one of the possible combinations of values of the explanatory variables). Thus it is customary to use only categorical variables as covariates and discretise continuous covariates. We follow this practice in this article.

Here we focus on the case where the pure premium is modelled by a compound Poisson-gamma distribution. This is equivalent to assuming that the claim frequency follows a Poisson distribution, that the claim severity follows a gamma distribution and that frequency is independent from severity.

Denote by N the number of claims within a year, by Y_i the amount of the i-th claim in the year and by S the aggregate claims:

$$S = \sum_{i=1}^{N} Y_i \,.$$

Here $\mathbb{E}[N]$, $\mathbb{E}[Y_1]$ and $\mathbb{E}[S]$ correspond to the claim frequency, claim severity and pure premium respectively. Under the independence assumption it follows that $\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[Y_1]$.

Now, if GLMs are to be used to estimate $\mathbb{E}[S]$ under these assumptions, we propose to consider also the use of a Tweedie distribution with $p \in (1, 2)$

for the response variable. In order to do this, p has to be estimated. A numerical method for the maximum likelihood estimator of p is given in Gilchrist and Drinkwater [5]. This method has now been implemented in R within the Tweedie package (see Dunn [4]).

Nevertheless, the common practice currently in the insurance industry is to fit a Poisson response GLM for the claim frequency, then a separate gamma response GLM for the claim severity and to multiply the means of the two GLMs to obtain an estimation for the pure premium. We will refer to this method of estimation as the separate Poisson-gamma approach (SPGA). Let us analyse the differences between this method and a Tweedie GLM.

3.1 Separate Poisson-gamma Approach (SPGA)

First consider the SPGA, starting with the fit of the frequency model. Assume that after the analysis of residuals and deviance we end up with k_N covariates. Let $\boldsymbol{\beta}^N$ denote the vector of corresponding regression coefficients. Similarly, let us assume that k_Y covariates are retained for the severity model and let $\boldsymbol{\beta}^Y$ be its vector of coefficients.

Since we might have used different covariates for each model, we need to unify the portfolio classification of these two models so we can multiply the frequency and severity estimators for each class. This can be done in the following way: Assume that the frequency and severity GLMs have k covariates in common. Define a common design matrix $\boldsymbol{X}_*$ in which the first k columns correspond to the common covariates, the next $k_N - k$ columns correspond to the covariates that are used only in the frequency model and the last $k_Y - k$ columns correspond to the covariates that are only used in the severity model. Then define an adjusted frequency vector of parameters $\boldsymbol{\beta}^N_* := (\boldsymbol{\beta}^N_1, \boldsymbol{\beta}^N_2, \mathbf{0})$, where $\boldsymbol{\beta}^N_1$ is a vector of dimension k with the values of the coefficients from $\boldsymbol{\beta}^N$ that correspond to the common covariates. $\boldsymbol{\beta}^N_2$ is a vector of dimension $k_N - k$ with the values of $\boldsymbol{\beta}^N$ that correspond to the covariates that appear only in the frequency model and $\mathbf{0}$ is a vector of zeros with dimension $k_Y - k$.

Similarly, define an adjusted severity vector of coefficients $\boldsymbol{\beta}^Y_* := (\boldsymbol{\beta}^Y_1, \mathbf{0}, \boldsymbol{\beta}^Y_2)$ where $\boldsymbol{\beta}^Y_1$ has dimension k, $\mathbf{0}$ has dimension $k_N - k$ and $\boldsymbol{\beta}^Y_2$ has dimension $k_Y - k$. Then, denoting by S^*_i the response variable of the i-th class and with $\boldsymbol{X}^*_i$ the i-th row of $\boldsymbol{X}_*$, we have for a log-link function that

$$\mathbb{E}(S^*_i) := \mu^*_i = \exp\left\{\boldsymbol{X}^*_i(\boldsymbol{\beta}^N_* + \boldsymbol{\beta}^Y_*)\right\} .$$

Here, $S_i^* \sim CPG(N_i^*, \alpha, \tau_i)$, where

$$\mathbb{E}(N) := \mu_{N_i^*} = \exp\left(\boldsymbol{X}_i^* \boldsymbol{\beta}_*^N\right)$$

is the mean of the frequency for the i-th class, and

$$\mathbb{E}(Y_i) := \mu_{Y_i^*} = \frac{\alpha}{\tau_i} = \exp(\boldsymbol{X}_i^* \boldsymbol{\beta}_*^Y)\,,$$

is the mean of the severity for the i-th class, with α and τ_i being the (constant) shape parameter and rate parameter, respectively, of the gamma distribution. By (6) this implies that the distribution of the i-th class is a $Tw(p^*, \mu_i^*, \lambda_i^*)$, where

$$p^* = \frac{\alpha+2}{\alpha+1}\,, \qquad \mu_i^* = \frac{\mu_{N_i^*}\alpha}{\tau_i} \quad \text{and} \quad \lambda_i^* = \frac{\left(\frac{\mu_{N_i^*}\alpha}{\tau_i}\right)^{\frac{\alpha+2}{\alpha+1}-1}\tau_i}{\alpha+1}\,. \tag{7}$$

3.2 Tweedie GLM

In turn, now consider the Tweedie GLM. Assume that for this model we have k_S covariates, let $\boldsymbol{\beta}^S$ be the vector of coefficients with $\boldsymbol{X}$ and S_i being the corresponding design matrix and response variable for the i-th class. In this model S_i follows a $Tw(p, \mu_i, \lambda)$ distribution, where $p \in (1,2)$ and $\lambda > 0$ are fixed for all the classes and

$$\mathbb{E}(S_i) := \mu_i = \exp\left(\boldsymbol{X}_i \boldsymbol{\beta}^S\right)\,.$$

3.3 Comparing the Models

Considering these above definitions, we can now enumerate the following important differences between these two models.

3.3.1 Number of Parameters

The SPGA has $k_N + k_Y + 1$ parameters: k_N betas for the Poisson GLM, k_Y betas for the gamma GLM plus one for its dispersion parameter. On the other hand the Tweedie GLM has $k_S + 1$: the k_S betas and the dispersion parameter. Usually the SPGA will use the most parameters: for example if for the Tweedie GLM we use the same explanatory variables as in the SPGA (i.e. all the explanatory variables from the frequency and severity models),

then $k_S = k_N + k_Y - k + 1$, so in this case we have k parameters more than the Tweedie GLM. Thus, in general, from the point of view of number of parameters the Tweedie GLM is a simpler model.

3.3.2 Variability of the Dispersion Parameter

From the discussion above we see that the index parameter λ_i of the SPGA varies along the different classes while the one in the Tweedie GLM remains constant. The main issue with having a constant index parameter is that according to the model, for any class, the larger the mean the larger the variance. This is because the variance of the i-th class is given by $\frac{\mathbb{E}[S_i^*]^p}{\lambda}$. Thus, the distributional assumptions of the model will not be satisfied in portfolios where this is not the case. On the other hand, denoting by μ_i^*, $\mu_{Y_i^*}$ and $\mu_{N_i^*}$ the means for the i-th class given by the SPGA for the aggregated loss, severity and frequency, respectively, we have that the variance of the i-th class is given by any of the following expressions:

$$\frac{\alpha+1}{\alpha}\,\mu_{Y_i^*}\,\mu_i^* = \frac{\alpha+1}{\alpha}\,\mu_{N_i^*}\,\mu_{Y_i^*}^2 = \frac{\alpha+1}{\alpha}\,\frac{(\mu_i^*)^2}{\mu_{N_i^*}}\,,$$

which is not strictly monotone with respect to μ_i^*. There is a way to cope with this limitation of the Tweedie GLM which consists in assigning covariates to the dispersion parameter. Such models are called double generalised linear models and are defined in Smyth and Verbyla [12].

3.3.3 Non-optimallity of the SPGA Parameters

As mentioned above, the i-th class follows a $Tw(p^*, \mu_i^*, \lambda_i^*)$ distribution. Now putting μ_i^* and λ_i^* in terms of the regression coefficients we have that

$$\begin{aligned}\mu_i^* &= \exp\left\{\boldsymbol{X}_i^*(\boldsymbol{\beta}_*^N + \boldsymbol{\beta}_*^Y)\right\}\,,\\ \lambda_i^* &= \frac{\alpha}{\alpha+1}\,.\end{aligned}$$

When we fit the model we estimate $\boldsymbol{\beta}^N$ and $\boldsymbol{\beta}^Y$ by maximising the likelihood equation for the frequency and severity GLMs respectively. These estimations do not correspond to the MLE of $\boldsymbol{\beta}^N$ and $\boldsymbol{\beta}^Y$ that we would find if we wrote the likelihood function of the joint analysis.

3.3.4 Loss of Information with the Tweedie GLM Estimation

In order to fit the Tweedie GLM, only the accumulated loss is used without using the number of claims. Thus, some information provided by the sample is lost. This issue can be solved by maximising the joint likelihood of (S, N) (aggregated loss and number of claims), instead of the likelihood of S. This is detailed in Jørgensen and de Souza [8].

3.4 Graphical and Empirical Comparisons; the Lift Chart

Given a sample from a specific portfolio, how can one chose the "best" model?

Let us first fit both models and analyse them individually. For the Tweedie GLM the usual methods for GLMs can be used (e.g. analysis of deviance and residual plots). For the SPGA the analyses are carried out separately for the frequency and severity GLMs.

If the individual analyses are not conclusive, it is important to note that it is not possible to write the Tweedie model as a special case of the SPGA. This is because the dispersion parameter λ_i of the SPGA varies with the different linear predictors in different classes, both for the frequency and severity GLMs. Thus a likelihood ratio test is not possible. One measure that can help in opting for a specific model is Akaike's information criterion (AIC) or the corrected AIC. The problem with this approach is that it requires the MLEs for both models.

Empirical graphical comparisons are also used to assess and compare the goodness of fit of these two models, like PP-plots and QQ-plots of the predicted values against the observed values.

Another graph that is useful in model testing is the lift chart. It comes in different versions, all essentially based on the same underlying idea. We outline here the lift chart version used in this article for goodness of fit comparisons. It is similar, to that given in Briere-Giroux et al. [2].

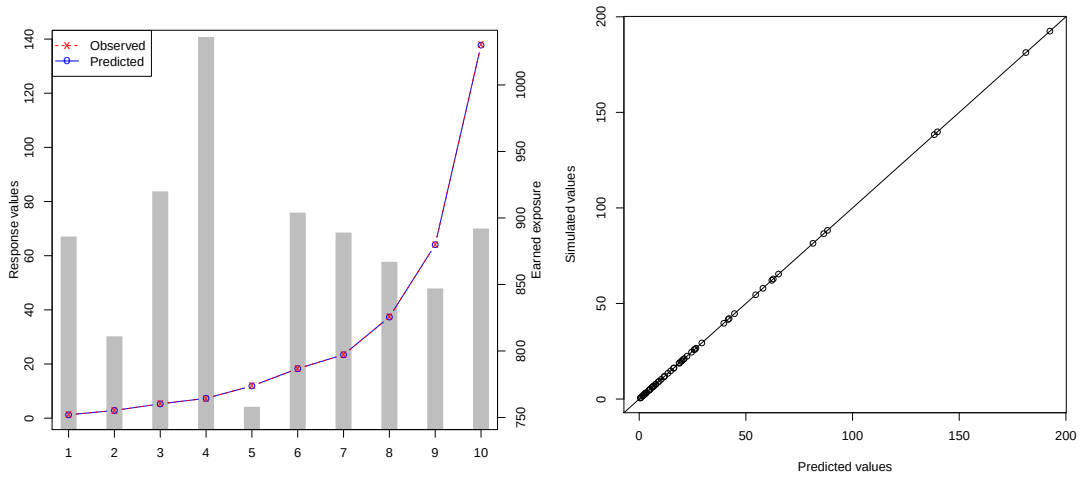
Suppose that a model is used to describe or predict a certain phenomenon and that observations of this phenomenon are available. The following steps are used in order to create a lift chart:

1. Using the model generate predictions for the observations.
2. Order the observations increasingly with respect to the predictions.
3. Divide the ordered data in groups with equal number of observations.

4. Plot the mean observation and mean prediction for each group.

When this chart is produced for a GLM, it is common to add bars that correspond to the exposure in each group. In these graphs, the scale of the vertical axis on the left corresponds to the mean computed for each group and the scale of the vertical axis on the right corresponds to the exposure.

Lift charts give information about two aspects of the model. On the one hand, by examining the trend on the curve for the observed means it is possible to see if the model identifies reasonably the groups that are more costly. On the other hand, the vertical distances between the predicted mean and the observed mean give an idea of how close the model predictions are to the observed data.



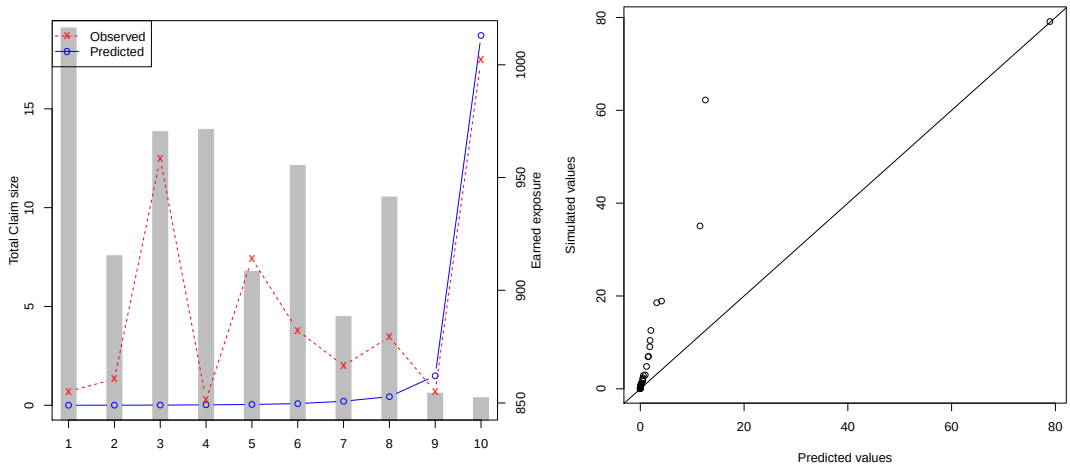
(a) Lift chart (b) QQ-plot

Figure 1: Example of lift chart for a good fit

Figure 1(a) is an example of a lift chart for a model that provides a very good fit. This graph was generated by simulating a gamma GLM with many observations. Then a GLM with gamma responses and a log-link function was fitted to this simulated data. The predictions from this fitted GLM and the simulated responses were then used to produce the lift chart. The illustrated

fit is very good since not only the observed values plot is increasing but also because the predictions are very close to the observed values. Figure 1(b) is a QQ-plot of the predictions against the simulated values that confirms the good fit.

Figure 2(a) illustrates a lift chart for a poorly fitting model. This graph was generated by simulating a GLM with inverse Gaussian responses and log-link, but then fitting a GLM with normal responses and a log-link to the simulated values. The graph shows that the fit is inadequate because the observed values plot is not increasing and also because many of the predictions are far from the observed values. A QQ-plot of this same data is given in Figure 2(b), which confirms the bad fit seen in the lift chart.



(a) Lift chart

(b) QQ-plot

Figure 2: Example of lift chart for a poor fit

4 Limitations of the Tweedie GLM

As mentioned above, one limitation of the Tweedie is that its mean increases with its variance. Besides, some recommend the SPGA over the Tweedie for

other reasons. For instance we quote the arguments presented in Ohlsson and Johansson [10]:

1. Claim frequency is usually much more stable than claim severity and often much of the power of rating factors is related to claim frequency: these factors can then be estimated with greater accuracy.
2. A separate analysis gives more insight into how a rating factor affects the pure premium.

(Note that above rating factor is a synonym of explanatory variable).

These arguments are questionable: using (5), given a Tweedie GLM for the aggregate claims, we can not only obtain predictions for the claim frequency and severity, but it also induces a GLM structure for each of them (i.e., a linear predictor with a link function for the mean). Thus potentially, the same insight gained with the separated approach can also be obtained with the Tweedie GLM. More specifically, assume that a Tweedie GLM is fitted to model the pure premium of a portfolio. Now consider one policyholder with a model prediction for the pure premium of μ. Also, let μ^P and μ^G denote, respectively, the means of the Poisson frequency and gamma severity induced by the Tweedie GLM. Assuming a log-link function, we have that

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{k} x_i \beta_i \,,$$

where $\beta_0, \ldots, \beta_k$ are the fitted coefficients, and $x_1, \ldots, x_k$ the values of the covariates for this policy-holder. By (5), we then have that

$$\mu^P = \frac{\lambda \mu^{2-p}}{2-p} = \frac{\lambda}{2-p} \left[\exp(\beta_0) \prod_{i+1}^{k} \exp(\beta_i x_i) \right]^{2-p} \,,$$

which implies that

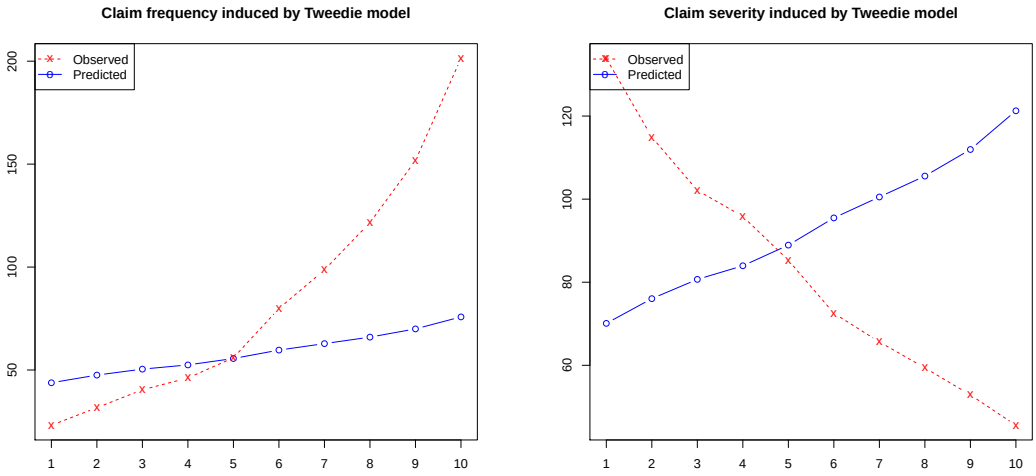
$$\ln(\mu^P) = \beta_0(2-p) + \ln\left(\frac{\lambda}{2-p}\right) + \sum_{i=1}^{k} x_i \beta_i (2-p) \,.$$

Denoting by $\beta_0^P = \beta_0(2-p) + \ln\left(\frac{\lambda}{2-p}\right)$ and $\beta_i^P = \beta_i(2-p)$, for $1 \leq i \leq k$, we get a GLM structure for the expected frequency, μ^P, with coefficients

$\beta_0^P, \ldots, \beta_k^P$. In a similar way, for the severity we have that

$$\ln(\mu^G) = \beta_0^G + \sum_{i=1}^{k} x_i \beta_i^G ,$$

where $\beta_0^G = \frac{2-p}{\lambda} + \beta_0(p-1)$ and $\beta_i^G = \beta_i(p-1)$. Thus, the Tweedie GLM for the pure premium induces GLMs for the frequency and the severity that use the same covariates.



(a) Lift chart

(b) QQ-plot

Figure 3: Example of induced frequency and severity models from the Tweedie GLM

For this argument to be of practical use, the adequacy of these induced models should be comparable to that of the separate GLMs for frequency and severity. Now, these induced models have an important limitation, that makes this unlikely: under the Tweedie GLM, a larger pure premium implies both, a larger claim frequency and claim severity. To see why this is the case, let μ_i be the mean of the i-th class of a Tweedie GLM, that means

$$\mu_i = \exp\left(\boldsymbol{X}_i \boldsymbol{\beta}^S\right) .$$

By (6), this corresponds to a $CPG(\frac{\lambda\mu_i^{2-p}}{2-p}, -\frac{p-2}{p-1}, \frac{\lambda\mu_i^{1-p}}{p-1})$. Hence the means for the claim frequency and claim severity are given by $\frac{\lambda\mu_i^{2-p}}{2-p}$ and $\frac{(2-p)\mu_i^{p-1}}{\lambda}$, respectively. Which shows that both increase or decrease with μ_i.

A simulated illustration exemplifies dramatically this limitation of the Tweedie. We simulated separate Poisson and gamma distributed datasets in such a way that for the classes with higher pure premium, there is also a higher claim frequency but a smaller claim severity. A Tweedie GLM is then fitted for the pure premium and, using (6), the induced predicted means for the claim frequency and severity are obtained. The lift charts that correspond to these predictions against the simulated values are given in Figure 3. We see that for the frequency the predictions are far from the simulated values, but at least the model is capable of explaining what classes are more expensive than others (this is because the observed values graph is increasing). On the other hand, for the severity graph, we see that it detects cheaper classes as more expensive. This shows that the Tweedie is not adequate for this simulated data.

5 Example: Canadian Automobile Insurance Claims for 1957-1958

Here we fit both models to a publicly available dataset and compare their results. The data was originally used in Bailey and Simon [1], but it is now available in complete form at `http://www.statsci.org/data/general/carinsca.html`, where the following description of the data is also available:

```
The data give the Canadian automobile insurance experience for policy
years 1956 and 1957 as of June 30, 1959. It includes virtually
every insurance company operating in Canada and was collated by the
Statistical Agency (Canadian Underwriters' Association - Statistical
Department) acting under instructions from the Superintendent of
Insurance. The data given here is for private passenger automobile
liability for non-farmers for all of Canada, excluding Saskatchewan.

The variable Merit measures the number of years since the last claim
on the policy. The variable Class is a collation of age, sex, use and
```

Variable	Description
Merit	Merit Rating:
	3 - licensed and accident free 3 or more years
	2 - licensed and accident free 2 years
	1 - licensed and accident free 1 year
	0 - all others
Class	1 - pleasure, no male operator under 25
	2 - pleasure, non-principal male operator under 25
	3 - business use
	4 - unmarried owner or principal operator under 25
	5 - married owner or principal operator under 25
Insured	Earned car years
Premium	Earned premium in 1000's
	(adjusted to what the premium would have been
	had all cars been written at 01 rates)
Claims	Number of claims
Cost	Total cost of the claim in 1000's of dollars

Table 3: Variables for Canadian Automobile Insurance Dataset

`marital status. The variables Insured and Premium are two measures of the risk exposure of the insurance companies`

Table 4 reports the data, to which we added the column group, corresponding to what we called class in previous sections, i.e. a possible combination of values of the covariates. It cannot be called class here to avoid any confusion with the explanatory variable *Class* in the dataset. All the models in this section were fitted with R (which is available at the webpage)

5.1 Frequency Model

First compare the frequency models. A Poisson model was fitted. After some analysis of the variables, 4 interactions terms were added to the regression model:

- Class 1 with Merit 3
- Class 3 with Merit 3
- Class 4 with Merit 3

Group	Merit	Class	Insured	Premium	Claims	Cost
1	3	1	2757520	159108	217151	63191
2	3	2	130535	7175	14506	4598
3	3	3	247424	15663	31964	9589
4	3	4	156871	7694	22884	7964
5	3	5	64130	3241	6560	1752
6	2	1	130706	7910	13792	4055
7	2	2	7233	431	1001	380
8	2	3	15868	1080	2695	701
9	2	4	17707	888	3054	983
10	2	5	4039	209	487	114
11	1	1	163544	9862	19346	5552
12	1	2	9726	572	1430	439
13	1	3	20369	1382	3546	1011
14	1	4	21089	1052	3618	1281
15	1	5	4869	250	613	178
16	0	1	273944	17226	37730	11809
17	0	2	21504	1207	3421	1088
18	0	3	37666	2502	7565	2383
19	0	4	56730	2756	11345	3971
20	0	5	8601	461	1291	382

Table 4: Data for Canadian Automobile Insurance Claims for 1957-1958

- Class 1 with Merit 2

In what follows we label these interactions as C1M3, C3M3, C4M3 and C1M2 respectively. The summary of the R output for this regression is given in Table 5.

The last column in the table shows that all the variables are significant. In the Poisson GLM, the dispersion parameter is equal to 1, it is not estimated. In this dataset the weight (the Insured variable) is high for all classes. This implies that we can use a χ^2 distribution with 8 degrees of freedom as an approximation for the deviance (see section 3.6 in Jørgensen [6]).

This distribution is commonly used for goodness of fit tests. Here the null hypothesis is that the data follows the fitted GLM. Thus, rejecting this test implies that our model might not be adequate for the data. Otherwise the deviance does not give evidence of lack of fit. The p-value of this test is $\mathbb{P}(\chi^2_8 \geq 7.3344) = 0.501$, thus the deviance does not present evidence of lack of fit.

	Estimate	Std. Error	z value	$\mathbb{P}(>\|z\|)$
(Intercept)	-1.9839	0.0048	-409.39	0.0000***
Merit1	-0.1478	0.0072	-20.57	0.0000***
Merit2	-0.1610	0.0132	-12.24	0.0000***
Merit3	-0.3746	0.0134	-28.00	0.0000***
Class2	0.1627	0.0126	12.94	0.0000***
Class3	0.3786	0.0098	38.54	0.0000***
Class4	0.3755	0.0088	42.47	0.0000***
Class5	0.0758	0.0150	5.05	0.0000***
C1M3TRUE	-0.1830	0.0140	-13.05	0.0000***
C3M3TRUE	-0.0666	0.0165	-4.04	0.0001***
C4M3TRUE	0.0580	0.0164	3.54	0.0004***
C1M2TRUE	-0.1039	0.0161	-6.46	0.0000***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Null deviance:	33854.1582 on 19 degrees of freedom			
Residual deviance:	7.3344 on 8 degrees of freedom			

Table 5: Summary table for Poisson GLM

The deviance residuals plot of this model is shown in Figure 4. The points that are labeled with a number correspond to the residuals that are more than 1.5 standard deviations away from the residuals mean. The number corresponds to the corresponding group of the residual. The dashed line corresponds to $x = 0$, and the lower and upper dashed lines correspond to the mean residuals, plus or minus 1.5 standard deviations, respectively.

We see no apparent trend, but residuals that correspond to groups 10, 12 and 17 are significantly further from 0 than the rest. In order to improve the fit we tried adding interactions between other variables in the data, in all cases the residuals of these three observations were still the largest (in absolute value).

An analysis of deviance helps to determine if some variables should be taken out. Let D_1, k_1 be the deviance and number of parameters of our current model and D_2, k_2 the deviance and number of parameters of the same model without some variable. Then as the number of observations or the smallest weight increases, we approximately have $D_2 - D_1 \sim \chi^2_{(k_1-k_2)}$ to test for inclusion of this variable Using the function `drop1` in R it is possible to obtain the p-value of this test for all the variables in the model; Table 8

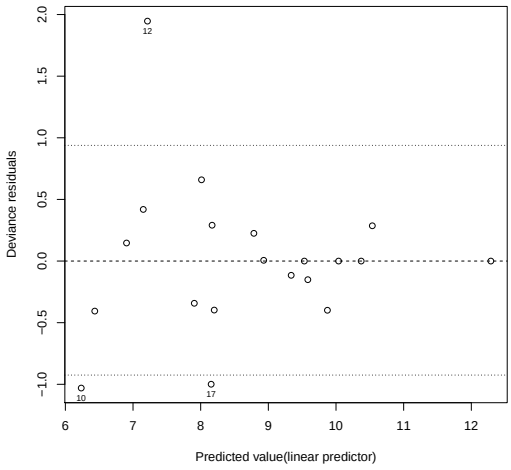


Figure 4: Residuals plot for the frequency model

shows the results.

	Df	Deviance	p-value
<none>		7.33	
Merit	3	1052.56	0.0000
Class	4	2544.10	0.0000
C1M3	1	180.81	0.0000
C3M3	1	23.66	0.0001
C4M3	1	19.84	0.0004
C1M2	1	48.71	0.0000

Table 6: p-values for testing the fit of reduced models, without one variable

We see that the p-values are very small for all variables. Hence, we reject the null hypothesis for the reduced models, and therefore keep the full model.

5.2 Severity Model

For the severity model we used a gamma GLM with a log-link function. After analysis, no interaction terms were added. Table 7 gives a summary of the R output for this regression.

	Estimate	Std. Error	t value	$\mathbb{P}(>\|t\|)$
(Intercept)	-1.1746	0.0155	-75.58	0.0000 ***
Merit1	-0.0687	0.0261	-2.63	0.0220 *
Merit2	-0.0702	0.0291	-2.41	0.0327 *
Merit3	-0.0567	0.0163	-3.48	0.0046 **
Class2	0.0827	0.0264	3.13	0.0087 **
Class3	0.0158	0.0183	0.86	0.4048
Class4	0.1598	0.0194	8.23	0.0000 ***
Class5	-0.0814	0.0391	-2.08	0.0593 .
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
(Dispersion parameter for gamma family taken to be 13.25825)				
Null deviance:	1556.0 on 19 degrees of freedom			
Residual deviance:	156.9 on 12 degrees of freedom			

Table 7: Summary for the gamma GLM fit

To test the goodness of fit we assume that the weights in the model (the Claims variable) are sufficiently large (the smallest one is 487) to assume that the residual deviance divided by the estimated dispersion parameter follows a χ^2 distribution with 12 degrees of freedom. The null hypothesis yields a p-value of 0.46, hence we cannot reject the model.

Figure 5 gives the deviance residuals plot for the gamma GLM model. The dashed lines have the same meaning as in the Poisson GLM residual plot. Finally, the analysis of deviance in Table 8 obtained using the `drop1` function in R again shows that the p-value is significant for both variables, indicating the use of the full model.

	Df	Deviance	$\mathbb{P}(>F)$
<none>		156.90	
Merit	3	342.54	0.0211
Class	4	1262.71	0.0000

Table 8: p-values for testing the fit of reduced severity models, without one variable

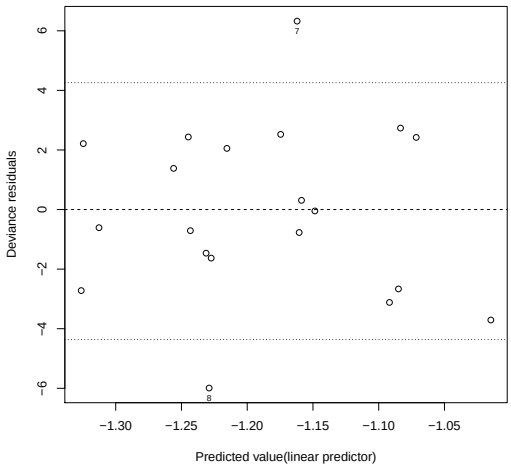


Figure 5: Residuals plot for the severity model

5.3 The Tweedie Model

After analysing the model two interaction terms were added to the Tweedie GLM: C1M3 and C4M3. These variables have the same meaning as in the frequency model. The summary R output for this model is given in Table 9.

The goodness of fit test for this model gives a p-value of 0.489, thus there is no evidence of lack of fit. The deviance residuals plot for this model is given in Figure 6, which does not show any clear pattern.

The test in Table 10 suggests to reject any reduced model and keep the full model.

5.4 SPGA vs Tweedie

The predictions of the SPGA model are obtained by multiplying the separate mean frequency and severity predictions for each class. For these separate predictions an exposure of 1 is used, and is then adjusted for the empirical

	Estimate	Std. Error	t value	$\mathbb{P}(>\|t\|)$
(Intercept)	-3.1549	0.0181	-174.54	0.0000 ***
Class2	0.2747	0.0377	7.28	0.0000 ***
Class3	0.3731	0.0335	11.15	0.0000 ***
Class4	0.5266	0.0353	14.91	0.0000 ***
Class5	0.0209	0.0464	0.45	0.6621
Merit1	-0.2201	0.0273	-8.05	0.0000 ***
Merit2	-0.3045	0.0296	-10.30	0.0000 ***
Merit3	-0.4675	0.0340	-13.76	0.0000 ***
C1M3TRUE	-0.1535	0.0356	-4.32	0.0015 **
C4M3TRUE	0.1153	0.0524	2.20	0.0525 .
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
(Dispersion parameter for Tweedie family taken to be 76.59105)				
Null deviance:	301049.59 on 19 degrees of freedom			
Residual deviance:	724.36 on 10 degrees of freedom			

Table 9: Summary table for Tweedie GLM

	Df	Deviance	F value	Pr(>F)
<none>		724.36		
Class	4	28613.75	96.26	0.0000 ***
Merit	3	20240.32	89.81	0.0000 ***
C1M3	1	2132.19	19.44	0.0013 **
C4M3	1	1094.56	5.11	0.0473 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Table 10: p-values for testing the fit of the tweedie model without each variable

values in the dataset; the exposure for the aggregated model is the variable `Insured`.

Table 11 shows the predicted means for both models as well as the observed mean. The latter is obtained by dividing `Cost` by `Insured` for each class. Note that the values are in thousands of dollars. We can see that the predictions from both models are close to each other, as well as to the observed means.

The lift curves for both models are given in Figure 7. We see that both distinguish reasonably well the classes that are more expensive, and that overall, the predicted and observed values are close to each other.

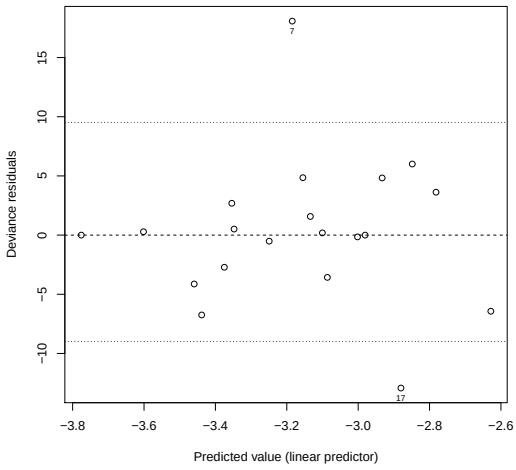


Figure 6: Residuals plot for the Tweedie model

In this example both models seem to explain the data equally well. The SPGA has 21 parameters while the Tweedie GLM only has 11. Thus the Tweedie GLM offers a more parsimonious option and is therefore preferred in this example.

It is important to mention that this preference for the Tweedie GLM is only as a pure premium model. We do not recommend its use here to draw conclusions about the frequency and severity since the Tweedie induced separate predicted means do not give a good fit.

6 Notes on Recent Developments

Modifications to the Tweedie GLM have been proposed to help overcome some of its limitations. We have already mentioned Jørgensen and de Souza [8] for the use of a joint likelihood and the double generalised linear models of Smyth and Verbyla [12].

More precisely, the joint likelihood approach includes the frequency so that the information given by the number of claims is not lost. Jørgensen and de Souza [8] gives an algorithm to find the MLE for the joint likelihood.

In double generalised linear models Smyth and Verbyla [12] also use explanatory variables to let the dispersion parameter vary with the class. This

Group	SPGA mean	Tweedie mean	Observed mean	Group	SPGA mean	Tweedie mean	Observed mean
1	0.022988	0.022916	0.022916	11	0.034220	0.034220	0.033948
2	0.035282	0.035165	0.035224	12	0.043738	0.045038	0.045137
3	0.038314	0.038802	0.038755	13	0.050769	0.049695	0.049634
4	0.049964	0.050768	0.050768	14	0.058447	0.057938	0.060743
5	0.027449	0.027283	0.027320	15	0.034028	0.034942	0.036558
6	0.030389	0.031449	0.031024	16	0.042489	0.042643	0.043107
7	0.043095	0.041391	0.052537	17	0.054308	0.056124	0.050595
8	0.050022	0.045672	0.044177	18	0.063038	0.061928	0.063267
9	0.057588	0.053247	0.055515	19	0.072572	0.072200	0.069998
10	0.033527	0.032113	0.028225	20	0.042251	0.043543	0.044413

Table 11: Comparison of the SPGA and Tweedie predicted means vs the observed means

solves the problem of monotonicity between the mean and variance of the Tweedie GLM. Smyth and Jørgensen [11] apply double generalised linear models with a Tweedie response to an auto insurance dataset.

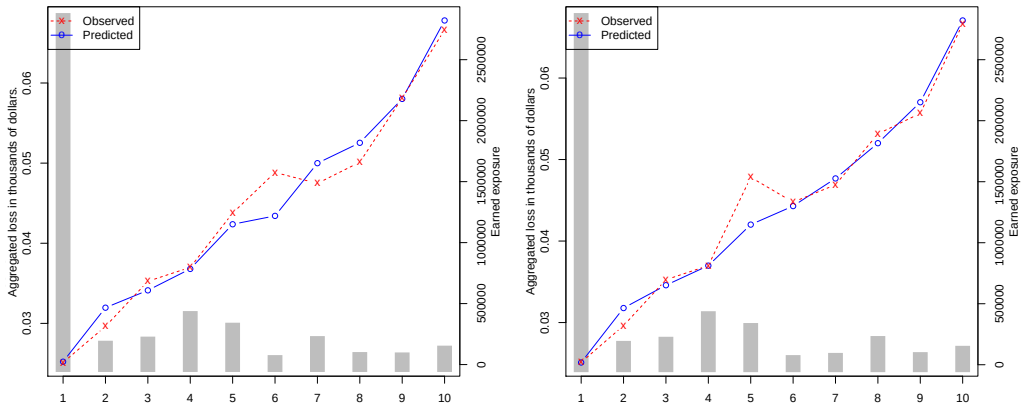
The SPGA and Tweedie GLMs reviewed here assume independence between the frequency and severity components. Relaxing this independence assumption has led recently to important results. Song [13] uses Gaussian copulas to construct multivariate GLMs with dependence in the joint distribution with different marginals. This construction also works for the case where some of the variables are discreet and other are continuous. This method is used in Czado et al. [3] to construct a bivariate Poisson-gamma GLM with dependence. Krämer et al. [9] extends this work by allowing general copulas. It also proposes a method to choose an optimal copula family.

7 Summary and Conclusions

The SPGA and Tweedie GLMs are two alternative ways of modelling the pure premium when a compound Poisson-gamma (CPG) distribution is assumed for the aggregate loss.

The Tweedie GLM is a simpler model. By the parsimony principle the Tweedie GLM should be preferred whenever both models explain the data similarly well.

The separate frequency and severity GLMs induced by the Tweedie GLM



(a) SPGA lift chart

(b) Tweedie GLM lift chart

Figure 7: Lift charts for both models

are only useful under very restrictive conditions. This makes them unreliable for gaining insight beyond the pure premium.

In conclusion, with regards to the pure premium, one should Tweedie rather than not, whenever the parsimony principle applies. By contrast, other models should, in general, be used in order to draw conclusions about the claim frequency and severity.

Acknowledgements

The authors are sincerely grateful to the Editor and an anonymous referee for their constructive comments that helped improve this educational note.

References

[1] Robert A. Bailey and Leroy Simon. Two studies in automobile insurance ratemaking. *ASTIN Bulletin*, 1(4):192–217, 1960.

[2] Guillaume Briere-Giroux, Jean-Felix Huet, Robert Spaul, Andy Staudt, and David Weinsier. Predictive modeling for life insurers, 2010. URL

https://www.soa.org/files/pdf/research-pred-mod-life-huet.pdf.

[3] Claudia Czado, Rainer Kastenmeier, Eike Christian Brechmann, and Aleksey Min. A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal*, (4):278–305, 2012.

[4] Peter K Dunn. *tweedie: Tweedie exponential family models*, 2014. R package version 2.2.1.

[5] R. Gilchrist and D. Drinkwater. Fitting tweedie models to data with probability of zero responses. *Proceedings of the 14th International Workshop on Statistical Modelling*, pages 207–214, 1999.

[6] Bent Jørgensen. *The Theory of Exponential Dispersion Models and Analysis of Deviance*. Instituto de Matemática Pura e Aplicada, (IMPA), Brazil, 1992.

[7] Bent Jørgensen. *The Theory of Dispersion Models*. Chapman & Hall, London, 1997.

[8] Bent Jørgensen and Marta C. Paes de Souza. Fitting tweedie's compound poisson model to insurance claims data. *Scandinavian Actuarial Journal*, pages 69–93, 1994.

[9] Nicole Krämer, Eike C. Brechmann, Daniel Silvestrini, and Claudia Czado. Total loss estimation using copula-based regression models. *Insurance: Mathematics and Economics*, 53(3):829 – 839, 2013.

[10] Esbjörn Ohlsson and Björn Johansson. *Non-Life Insurance Princing with Generalized Linear Models*. Springer-Verlag, Berlin, 2010.

[11] G.K. Smyth and B. Jørgensen. Fitting tweedie's compound poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, 32(1), 2002.

[12] Gordon K. Smyth and Arūnas P. Verbyla. Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics*, 10(6):695–709, 1999.

[13] Peter Song. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.