# Fitting Tweedie's compound Poisson model to pure premium with the EM algorithm

Guangyuan Gao*

June 28, 2022

**Abstract**

We consider the situation when the number of claims is unavailable, and a Tweedie's compound Poisson model is fitted to the observed pure premium. Currently, there are two different models based on the Tweedie distribution: a single generalized linear model (GLM) for mean and a double generalized linear model (DGLM) for both mean and dispersion. Although the DGLM approach facilitates the heterogeneous dispersion, its soundness relies on the accuracy of the saddlepoint approximation, which is poor when the proportion of zero claims is large. For both models, the power variance parameter is estimated by considering the profile likelihood, which is computationally expensive. We propose a new approach to fit the Tweedie model with the EM algorithm, which is equivalent to an iteratively re-weighted Poisson-gamma model on an augmented data set. The proposed approach addresses the heterogeneous dispersion without needing the saddlepoint approximation, and the power variance parameter is estimated during the model fitting. Numerical examples show that our proposed approach is superior to the two competing models.

**Keywords:** Tweedie's compound Poisson model; Tweedie distribution; Exponential dispersion family; The EM algorithm; Generalized linear model.

*Center for Applied Statistics and School of Statistics, Renmin University of China, 100872 Beijing, China.



Electronic copy available at: https://ssrn.com/abstract=4148316

# 1 Introduction

In insurance pricing, historical loss and risk factors are analyzed. Statistical models are implemented to regress random variables of loss on risk factors. Claim frequency $\lambda$, claim severity $\tau$ or pure premium $\mu$ are estimated from the statistical models. Policies are priced in parallel to their risk factors $\boldsymbol{x}$. Many statistical models have been used for insurance pricing, such as generalized linear models (GLMs), generalized additive models, generalized mixed effect models, mixture models, etc. Two crucial components in every model are the distribution and the regression function. A more suitable distribution for the loss variable would lead to a better model fitting. A more flexible regression function would explore more predictive potentials in the risk factor space. This paper assumes a Poisson-gamma GLM framework, which is the most commonly used model setting in insurance pricing. More specifically, we assume that the number of claims $N$ follows a Poisson distribution, the individual claim amount $Z$ follows a gamma distribution, and preprocessed risk factors $\boldsymbol{x}$ have a linear systematic effect on model parameters. The total claim amount is denoted by $S$, the exposure by $e$, and the observed pure premium by $Y = S/e$.

We discuss the Poisson-gamma GLM under two scenarios: (1) both the number of claims $N$ and total claim amount $S$ (or equivalently $Y$) are observed; (2) only total claim amount $S$ (or equivalently $Y$) is observed. Smyth and Jørgensen (2002) have studied Tweedie's compound Poisson model fitting under the two scenarios. In the first scenario of both $N$ and $S$ are observed, there are two modelling approaches: (1) two separate models consisting of a Poisson GLM for claim numbers $N$ and a gamma GLM for average claim sizes $S/N$; (2) a joint Tweedie's compound Poisson regression model for $(Y, N)$. Delong et al. (2021) have investigated the two approaches and provided theoretical evidence supporting the industry preference for the first approach of separate modelling. In the second scenario of only $S$ available, a marginal Tweedie's compound Poisson regression model is fitted to the observed pure premium $Y = S/e$; see Smyth and Jørgensen (2002). Since a marginal Tweedie's compound Poisson distribution given the power variance parameter $p$ belongs to the exponential dispersion family (Tweedie et al., 1984), there exist two types of GLMs: a single Tweedie GLM for mean $\mu$ with homogeneous dispersion $\phi$ and a double generalized linear model (DGLM) for both mean $\mu$ and dispersion $\phi$. For more detail on DGLM, please refer to Nelder and Pregibon (1987) and Smyth (1989).

This paper focuses on the second scenario, where only the total claim amount is available. It is worth noting that the current models (single GLM or DGLM) are not satisfactory. First, a single GLM with constant dispersion implicitly assumes that $\lambda^{1-p}\tau^{2-p}$ remains constant over all policies (Smyth and Jørgensen, 2002), which is not valid for many real data. Second, a DGLM relies on the accuracy of the saddlepoint approximation of the probability density function, which is poor if the exposure is small and the proportion of zero claims is large (Nelder and Pregibon, 1987; Jørgensen, 1997). For example, DGLM is not suitable for policy-level data. Third, the estimation of power variance parameter $p$ is usually done by considering the profile log-likelihood for $p$. That is, we have to fit different models with every $p \in (1, 2)$ to find the optimal one, which is computationally expensive.

Above weaknesses motivate our proposed model-fitting method for the marginal Tweedie's compound Poisson model. The idea is relatively straightforward. Since there are difficulties with fitting the marginal Tweedie's compound Poisson model for observed pure premium $Y$, we will stay with the joint model for $(Y, N)$. However, $N$ is not observed. Thus, we propose estimating parameters with the expectation-maximization (EM) algorithm (Dempster et al., 1977), which



Electronic copy available at: https://ssrn.com/abstract=4148316

is a natural choice when having unobserved latent data. The proposed model-fitting method overcomes the weaknesses associated with the single GLM and the DGLM. First, the proposed method addresses the heterogeneous dispersion without needing the saddlepoint approximation. Therefore, dispersion modelling can be done for policy-level data by the proposed method. Second, the power variance parameter and regression coefficients are estimated simultaneously during the EM iterations, which takes less computing time than the single GLM or the DGLM.

The rest of the paper is structured as follows. Section 2 reviews Tweedie's compound Poisson model. Section 3 presents our proposed model-fitting method. Section 4 studies a simulated data example and a real data example to illustrate the proposed method and compare it with competing models. Section 5 concludes the paper with several important findings. The R code for Section 4 is publicly available at `https://github.com/sxpyggy/fitting-tweedieCP-with-EM`.

## 2 Tweedie's compound Poisson model

We first set up the compound Poisson model with gamma-distributed individual claim amount (Poisson-gamma model). Then we revisit the connection between the Poisson-gamma model and Tweedie's compound Poisson model. We highlight several challenges with fitting a marginal Tweedie's compound Poisson model.

### 2.1 Poisson-gamma model

The number of claims $N_i$ incurred by policy $i$ follows a Poisson distribution with mean $e_i\lambda_i$:

$$N_i \overset{\text{ind.}}{\sim} \text{Poisson}(e_i\lambda_i), \text{ for } i = 1, \ldots, n,$$

where $e_i > 0$ is the exposure of policy $i$ in years and $\lambda_i > 0$ is the annual claim frequency of policy $i$. The individual claim amount $Z_j$ made by policy $i$ follows a gamma distribution with shape $\alpha$ (dispersion $1/\alpha$) and mean $\tau_i$ (scale $\tau_i/\alpha$):

$$Z_j \overset{\text{i.i.d.}}{\sim} \text{gamma}(\alpha, \tau_i), \text{ for } j = 1, \ldots, N_i, \text{ if } N_i > 0,$$

where we assume that the shape parameter $\alpha$ is constant over all policies $i = 1, \ldots, n$. The total claim amount made by policy $i$ is

$$S_i = \begin{cases} \sum_{j=1}^{N_i} Z_{i,j}, & \text{if } N_i > 0, \\ 0, & \text{if } N_i = 0. \end{cases}$$

The conditional distribution of $S_i$ given a positive number of claims $N_i > 0$ is a gamma distribution with shape $N_i\alpha$ and mean $N_i\tau_i$. Therefore, the conditional distribution of observed pure premium $Y_i = S_i/e_i$ given a positive number of claims $N_i > 0$ is a gamma distribution with shape $N_i\alpha$ and mean $N_i\tau_i/e_i$ (scale $\tau_i/(\alpha e_i)$).

The joint distribution of $Y_i$ and $N_i$ is given by

$$\begin{aligned} f_{Y,N}(y, m) &= f_{Y|N=m}(y) f_N(m) \\ &= \begin{cases} f_{poi}(0; e_i\lambda_i) & \text{for } m = 0; \\ f_{poi}(m; e_i\lambda_i) f_{gam}(y; m\alpha, m\tau_i/e_i) & \text{for } m > 0, \end{cases} \end{aligned} \tag{2.1}$$



Electronic copy available at: https://ssrn.com/abstract=4148316

where $f_{poi}$ is the Poisson probability mass function and $f_{gam}$ is the gamma probability density function. By summing (2.1) over $N$, we obtain the marginal distribution of $Y_i$, which has a probability mass $f_{poi}(0; e_i\lambda_i)$ at $y = 0$ and a probability density function $f_Y(y)$ at $y \in (0, \infty)$

$$\begin{aligned} f_Y(y) &= \sum_{m=1}^{\infty} f_{Y,N}(y, m) \\ &= \sum_{m=1}^{\infty} \frac{(e_i\lambda_i)^m \exp(-e_i\lambda_i)}{m!} \frac{y^{m\alpha-1} \exp\left(-\frac{y}{\tau/(e_i\alpha)}\right)}{\Gamma(m\alpha)(\tau_i/(e_i\alpha))^{m\alpha}} \\ &= \exp\left(-\frac{e_i\alpha}{\tau_i}y - e_i\lambda_i\right) \sum_{m=1}^{\infty} \left(\left(\frac{e_i^{\alpha+1}y^{\alpha}\lambda_i\alpha^{\alpha}}{\tau_i^{\alpha}}\right)^m \frac{1}{m!\Gamma(m\alpha)y}\right). \end{aligned} \tag{2.2}$$

The probability density function $f_Y(y)$ for positive $y$ involves a summation of infinitely many complicated terms.

## 2.2 Tweedie's compound Poisson model

It can be shown that the marginal distribution of $Y_i$ belongs to the exponential distribution family (Jørgensen and De Souza, 1994):

$$\begin{aligned} f_Y(y) =& c(y; \phi_i/e_i, p) \exp\left(\frac{e_i}{\phi_i}\left(y\frac{\mu_i^{1-p}}{1-p} - \frac{\mu_i^{2-p}}{2-p}\right)\right) \\ =& c(y; \phi_i/e_i, p) \exp\left(\frac{e_i}{\phi_i}y\theta_i - \kappa_p(\theta_i)\right), \end{aligned} \tag{2.3}$$

where the normalized constant is

$$c(y; \phi_i/e_i, p) = \begin{cases} 1 & \text{for } y = 0; \\ \sum_{m=1}^{\infty} \left(\frac{(e_i/\phi_i)^{\alpha+1}y^{\alpha}}{(p-1)^{\alpha}(2-p)}\right)^m \frac{1}{m!\Gamma(m\alpha)y} & \text{for } y > 0. \end{cases} \tag{2.4}$$

The relationship between the Tweedie parameters $\mu_i, \phi_i, p$ and the Poisson-gamma parameters $\lambda_i, \alpha, \tau_i$ are as follows:

$$\begin{aligned} \text{Mean: } \mu_i &= \lambda_i\tau_i, \\ \text{Dispersion: } \phi_i &= \frac{\lambda_i^{1-p}\tau_i^{2-p}}{2-p}, \\ \text{Power: } p &= \frac{\alpha+2}{\alpha+1}. \end{aligned} \tag{2.5}$$

The variance function of $Y_i$ is given by

$$\text{Var}(Y_i) = \frac{\phi_i\mu_i^p}{e_i}.$$

Thus, the observed pure premium $Y$ follows a Tweedie distribution with power variance parameter $p \in (1, 2)$, which belongs to the exponential dispersion family. Currently, parameters are estimated either in a single Tweedie GLM for mean $\mu_i$ with constant dispersion $\phi$ or in a DGLM for both mean $\mu_i$ and dispersion $\phi_i$. There are several challenges with the two models. First, a single Tweedie GLM assumes the dispersion $\phi_i = (\lambda_i^{1-p}\tau_i^{2-p})/(2-p)$ as constant, which is



Electronic copy available at: https://ssrn.com/abstract=4148316

rarely being the case. Second, dispersion modelling is intricate due to the sum of infinitely many terms in (2.4). Thus, the saddlepoint approximation is applied to simplify (2.3) such that the dispersion can be modelled by using unit deviance, called double GLM (Nelder and Pregibon, 1987; Smyth, 1989). However, the saddlepoint approximation is not accurate when $e_i/\phi_i$ is small and the proportion of zero claims is large, which is quite common in individual policy data. Third, the regression coefficients in $\mu, \phi$ and the power variance parameter $p$ cannot be estimated simultaneously. Those challenges motivate our proposed method in the next section.

# 3 Fitting Tweedie's compound Poisson model with the EM algorithm

Rather than fitting marginal Tweedie's compound Poisson model (2.3) directly, we stay with separate Poisson-gamma modelling (2.1). Since the number of claims is not observed, the EM algorithm is a natural choice for parameter estimation. In this section, we first propose an EM algorithm for fitting a marginal Tweedie's compound Poisson model. Then this proposed EM algorithm is converted into iteratively re-weighted Poisson-gamma regressions on an augmented data set, which can be easily implemented in R.

## 3.1 The EM algorithm

The EM algorithm relies on the joint distribution of observed and latent data, which is usually more tractable than the marginal distribution of observed data. This is exactly true for Tweedie's compound Poisson model, where the joint distribution (2.1) is more tractable than the marginal distribution (2.3). In a nutshell, the EM algorithm contains two iterative steps: in the expectation step, the expectation of joint likelihood of $(Y, N)$ is calculated w.r.t. the conditional distribution of $N$ given $Y$ and current parameter estimates; in the maximization step, the expected joint likelihood is maximized to obtain the updated parameter estimates. We propose an EM algorithm for fitting a marginal Tweedie's compound Poisson model to the observed pure premium $Y_i$, as shown in Algorithm 1.

**Algorithm 1** The EM algorithm for marginal Tweedie's compound Poisson model.

1: Initialization. Set suitable initial parameter estimates $\hat{\theta}^{[0]}$
2: **for** $t = 1$ to $T$ **do**
3: Conditional distribution of latent data. Calculate the conditional distribution of $N$ given $Y$ under current estimated parameters $\hat{\theta}^{[t-1]}$, denoted by $f_{N|Y}(m; \hat{\theta}^{[t-1]})$.
4: Expectation. Calculate the expectation of the joint log-likelihood (2.1) w.r.t. the conditional distribution $f_{N|Y}(m; \hat{\theta}^{[t-1]})$, denoted by $Q(\theta|\hat{\theta}^{[t-1]})$.
5: Maximization. Find the parameters that maximize $Q(\theta|\hat{\theta}^{[t-1]})$:

$$\hat{\theta}^{[t]} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{[t-1]}).$$

6: **end for**
7: **return** $\hat{\theta} = \hat{\theta}^{[T]}$.

In the initialization, a Poisson GLM is fitted to the occurrence of claims $O_i = \mathbb{1}_{(0,\infty)}Y_i$ for



Electronic copy available at: https://ssrn.com/abstract=4148316

$i = 1, \ldots, n$. The MLEs of regression coefficients in $\lambda$ are denoted by $\hat{\beta}^{[0]}$, i.e., $\hat{\lambda}_i^{[0]} = \hat{\lambda}^{[0]}(\boldsymbol{x}_i) = \lambda(\boldsymbol{x}_i; \hat{\beta}^{[0]})$. A gamma GLM is fitted to the observed positive pure premium $Y_i = S_i/e_i$ for $\{i : Y_i > 0\}$. The MLEs of regression coefficients in $\tau$ are denoted by $\hat{\gamma}^{[0]}$, i.e., $\hat{\tau}_i^{[0]} = \hat{\tau}^{[0]}(\boldsymbol{x}_i) = \tau(\boldsymbol{x}_i; \hat{\gamma}^{[0]})$, and the MLE of shape parameter is denoted by $\hat{\alpha}^{[0]}$. We denote the parameter vector by $\theta = (\beta, \gamma, \alpha)$. By the Bayes theorem, the conditional distribution of $N$ given $Y$ and current parameter estimates $\hat{\theta}^{[t-1]}$ is derived as

$$f_{N|Y=y}(m; \hat{\theta}^{[t-1]}) = \begin{cases} \mathbb{1}_0(m) & \text{for } y = 0; \\ \frac{f_{Y,N}(y,m;\hat{\theta}^{[t-1]})}{f_Y(y;\hat{\theta}^{[t-1]})} & \text{for } y > 0. \end{cases} \tag{3.1}$$

Note that $f_Y(y)$ is Tweedie distribution (2.3), and all the distributions are evaluated at the current estimated parameters $\hat{\theta}^{[t-1]}$. In the expectation step, the expected log-likelihood is calculated as

$$\begin{aligned} & Q(\theta|\hat{\theta}^{[t-1]}) \\ =& \mathbb{E}_{N|Y}(\log L(\theta; \boldsymbol{y}, N)) \\ =& \mathbb{E}_{N|Y}\left(\sum_{i=1}^{n} \log f_{Y,N}(y_i, N; \theta)\right) \\ =& \sum_{i=1}^{n} \sum_{m=0}^{\infty} f_{N|Y=y_i}(m; \hat{\theta}^{[t-1]}) \log f_{Y,N}(y_i, m; \theta). \end{aligned} \tag{3.2}$$

In the maximization step, we find the parameters that maximize $Q(\theta|\hat{\theta}^{[t-1]})$:

$$\hat{\theta}^{[t]} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{[t-1]}).$$

This step is equivalent to weighted regressions on an augmented data set, which will be discussed in the next subsection.

### 3.2 The EM algorithm as iteratively re-weighted Poisson-gamma regressions on augmented data

The expected log-likelihood (3.2) can be further decomposed as follows:

$$\begin{aligned} & Q(\theta|\hat{\theta}^{[t-1]}) \\ =& \sum_{i=1}^{n} \sum_{m=0}^{\infty} f_{N|Y=y_i}(m; \hat{\theta}^{[t-1]}) \log f_{Y,N}(y_i, m; \theta) \\ =& \sum_{i:y_i=0} \log f_{Y,N}(y_i, 0; \theta) + \sum_{i:y_i>0} \sum_{m=1}^{\infty} f_{N|Y=y_i}(m; \hat{\theta}^{[t-1]}) \log f_{Y,N}(y_i, m; \theta) \\ \approx& \sum_{i:y_i=0} \log f_N(0; \theta) + \sum_{i:y_i>0} \sum_{m=1}^{M} f_{N|Y=y_i}(m; \hat{\theta}^{[t-1]}) \log f_{Y,N}(y_i, m; \theta) \\ =& \sum_{i:y_i=0} \log f_{poi}(0; e_i\lambda_i) + \sum_{i:y_i>0} \sum_{m=1}^{M} f_{N|Y=y_i}(m; \hat{\theta}^{[t-1]}) \log f_{poi}(m; e_i\lambda_i) + \\ & \sum_{i:y_i>0} \sum_{m=1}^{M} f_{N|Y=y_i}(m; \hat{\theta}^{[t-1]}) \log f_{gam}(y_i; m\alpha, m\tau_i/e_i), \end{aligned} \tag{3.3}$$



Electronic copy available at: https://ssrn.com/abstract=4148316

where $M$ is selected such that $\sum_{m=M+1}^{\infty} f_{N|Y=y_i}(m;\hat{\theta}^{[t-1]})$ is negligible, or equivalently, the sum of all conditional probabilities is close to sample size $n$, i.e.,

$$\sum_{i=1}^{n}\sum_{m=0}^{M} f_{N|Y=y_i}(m;\hat{\theta}^{[t-1]}) \approx n.$$

From (3.3), the expected log-likelihood $Q(\theta|\hat{\theta}^{[t-1]})$ can be decomposed into two parts: the Poisson part

$$\sum_{i:y_i=0} \log f_{poi}(0; e_i\lambda_i) + \sum_{i:y_i>0}\sum_{m=1}^{M} f_{N|Y=y_i}(m;\hat{\theta}^{[t-1]}) \log f_{poi}(m; e_i\lambda_i), \tag{3.4}$$

and the gamma part

$$\begin{aligned} &\sum_{i:y_i>0}\sum_{m=1}^{M} f_{N|Y=y_i}(m;\hat{\theta}^{[t-1]}) \log f_{gam}(y_i; m\alpha, m\tau_i/e_i) \\ =&\sum_{i:y_i>0}\sum_{m=1}^{M} m f_{N|Y=y_i}(m;\hat{\theta}^{[t-1]}) \log f_{gam}(y_i; \alpha, m\tau_i/e_i). \end{aligned} \tag{3.5}$$

Therefore, in the maximization step, the MLEs of Poisson parameters $\beta$ and gamma parameters $\gamma, \alpha$ can be obtained by fitting a weighted Poisson GLM and a weighted gamma GLM, respectively, to an augmented data set. This augmented data set contains $n_0$ policies without claims and $M(n-n_0)$ synthetic policies with claims. More specifically, for each policy with claims $(Y_i, \boldsymbol{x}_i, e_i)$ we generate $M$ synthetic policies $(Y_i, N_i = 1, \boldsymbol{x}_i, e_i), \ldots, (Y_i, N_i = M, \boldsymbol{x}_i, e_i)$. We denote the index of the augmented data set by $i = 1, \ldots, n^+ = n_0 + M(n-n_0)$ with a slight abuse of notation $i$. The weighted Poisson regression is fitted to all augmented data points, while the weighted gamma regression is fitted to $M(n-n_0)$ synthetic policies with claims.

We find that the EM algorithm shown in Algorithm 1 is equivalent to iteratively re-weighted Poisson-gamma regressions on an augmented data set, as shown in Algorithm 2. The conditional distribution calculation and expectation steps in Algorithm 1 can be treated as "weights update", while the maximization step in Algorithm 1 is equivalent to "weighted regressions". To end this section, we list three critical points when implementing Algorithm 2 in practice. First, $M$ should be sufficiently large such that $\sum_{i=1}^{n}\sum_{m=0}^{M} f_{N|Y=y_i}(m;\hat{\theta}^{[t-1]}) \approx n$. The choice of $M$ depends on the claim frequency; see the simulated data example in Section (4.1). Second, in the weighted Poisson-gamma regressions, the offset terms are specified as $\log e_i$ and $\log N_i - \log e_i$, respectively. Note that we assume a logarithm link function for both regressions. Third, we must obtain the MLE of shape parameter $\alpha$ at each iteration. If the shape parameter is estimated using the residuals typically done by R, then the algorithm will not converge.

# 4 Numerical examples

In this section, we study simulated and real data examples. In both examples, we compare the proposed method with two competing models: a single Tweedie GLM with homogeneous dispersion and a Tweedie DGLM for both mean and dispersion modelling.



Electronic copy available at: https://ssrn.com/abstract=4148316

**Algorithm 2** The iteratively re-weighted Poisson-gamma regressions on the augmented data.

1: Initialization. Set suitable initial parameter estimates $\hat{\theta}^{[0]} = (\hat{\beta}^{[0]}, \hat{\gamma}^{[0]}, \hat{\alpha}^{[0]})$

2: Data augmentation. Keep the policies without claims. For each policy with claims, generate $M$ synthetic policies with the same $Y, \boldsymbol{x}, e$ but different $N \in \{1, \ldots, M\}$.

3: **for** $t = 1$ to $T$ **do**

4: Weights update. Calculate the weights for number of claims $(N_i)_{i=1:n^+}$ in the Poisson regression as

$$\omega_i = \begin{cases} 1 & \text{if } m_i = 0, \\ f_{N|Y=y_i}(m_i; \hat{\theta}^{[t-1]}) & \text{if } m_i > 0. \end{cases}$$

Calculate the weights for observed positive pure premium $Y_i > 0$ in the gamma regression as

$$\omega_i^* = m_i f_{N|Y=y_i}(m_i; \hat{\theta}^{[t-1]}).$$

5: Weighted regressions. Fit a weighted Poisson regression to the $n^+$ augmented policies $(N_i, \boldsymbol{x}_i, e_i)_{i=1:n^+}$ with mean

$$e_i \lambda_i = e_i \lambda(\boldsymbol{x}_i; \beta)$$

and weights $\omega_i$ to obtain the updated MLEs of $\hat{\beta}^{[t]}$. Fit a weighted gamma regression to the $M(n - n_0)$ synthetic policies $(Y_i, N_i, \boldsymbol{x}_i, e_i)_{i:Y_i>0}$ with mean

$$N_i \tau_i / e_i = N_i \tau(\boldsymbol{x}_i; \gamma)/e_i$$

and weights $\omega_i^*$ to obtain the updated MLEs of $\hat{\gamma}^{[t]}, \hat{\alpha}^{[t]}$. Update $\hat{\theta}^{[t]} = (\hat{\beta}^{[t]}, \hat{\gamma}^{[t]}, \hat{\alpha}^{[t]})$.

6: **end for**

7: **return** $\hat{\theta} = \hat{\theta}^{[T]}$.

### 4.1 Simulated data example

We conduct simulation studies to demonstrate the advantages of our proposed method over the competing models. We consider two different exposure levels. In case 1, the average exposure is around five years, so the proportion of zero claims is low. In case 2, the average exposure is around one year, so the proportion of zero claims is high. We expect the DGLM to perform better in case 1 than in case 2. We generate $n = 10,000$ data points for training models and another $n^* = 10,000$ data points for comparing the out-of-sample performance of different modelling approaches.

The underlying model is specified as follows:

$$\begin{aligned} N_i &\overset{\text{ind.}}{\sim} \text{Poisson}(e_i \lambda_i), \text{ for } i = 1, \ldots, n, \\ Z_{i,j} &\overset{\text{i.i.d.}}{\sim} \text{gamma}(\alpha, \tau_i), \text{ for } j = 1, \ldots, N_i, \text{ if } N_i > 0, \\ S_i &= \begin{cases} \sum_{j=1}^{N_i} Z_{i,j}, & \text{for } N_i > 0, \\ 0, & \text{for } N_i = 0, \end{cases} \\ Y_i &= S_i / e_i, \end{aligned}$$



Electronic copy available at: https://ssrn.com/abstract=4148316

where the parameters are assumed as

$$\begin{aligned}
\eta_i &= \eta(\boldsymbol{x}_i) = 0.5 \tanh\left(\tanh(x_{i,1}) - \tanh(x_{i,2}) - \tanh(x_{i,3}) + \tanh(x_{i,4}) - 0.8x_{i,5}\right) - 2, \\
\lambda_i &= \lambda(\boldsymbol{x}_i) = \exp(\eta(\boldsymbol{x}_i)), \\
\zeta_i &= \zeta(\boldsymbol{x}_i) = 0.8 \tanh(-(x_{i,2}-1)^2 - (x_{i,3}-1)^2) + \log 2000, \\
\tau_i &= \tau(\boldsymbol{x}_i) = \exp(\zeta(\boldsymbol{x}_i)), \\
\alpha &= 1.
\end{aligned} \tag{4.1}$$

Covariates $x_1, x_2, x_3, x_4$ are generated from a normal distribution with mean of 0 and standard deviation of 0.5, independently, and covariate $x_5$ is generated from a Bernoulli distribution with probability of 0.5. We use the hyperbolic tangent function to make the effects of different covariates comparable and to cap the effect of outlying covariate values at $(-1, 1)$.

The histograms of true $\lambda_i, \tau_i, \mu_i, \phi_i$ are shown in Figure 1. The average annual claim frequency is 12.49%, the average claim size is 986.56, the average pure premium is 120.78, and the average dispersion of pure premium is 183.45. The histogram of dispersion implies that a single Tweedie GLM with homogeneous dispersion is not appropriate. We draw the pair plots for the linear predictors $\eta_i, \zeta_i$ and logged $\mu_i, \phi_i$ against the covariates in Figures 6, 7, 8 and 9, respectively, in the appendix. No significant non-linearity is observed. Thus, the GLMs with logarithm link function would be sufficient to capture the major systematic effects of covariates, and we will not conduct covariate transformation or selection in the following.

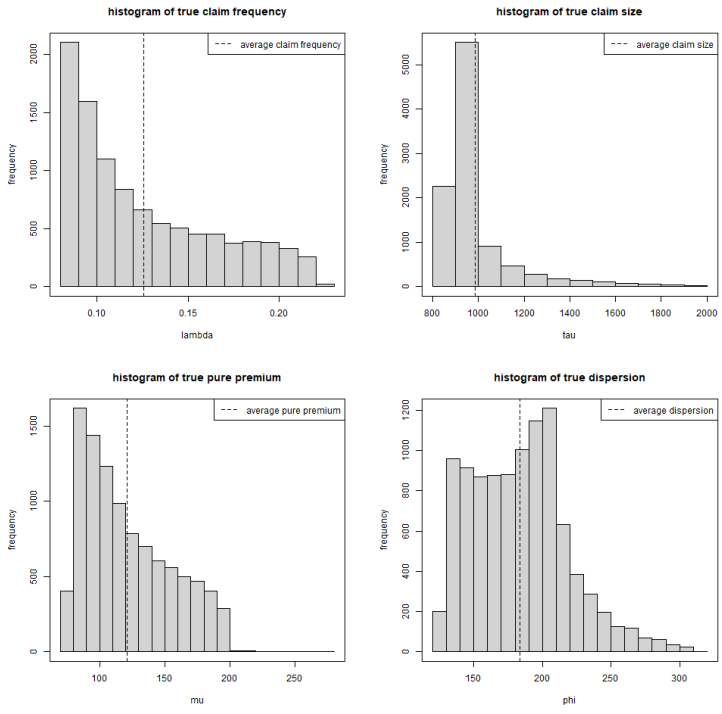



Figure 1: The histograms of true $\lambda_i, \tau_i, \mu_i, \phi_i$.



Electronic copy available at: https://ssrn.com/abstract=4148316

#### 4.1.1 Case 1: Large exposure

Suppose that exposure $e_i$ is generated from a uniform distribution in $(4, 6)$. The proportion of zero claims is 53.36%. In the augmented data, there are $n_0 = 5,336$ policies without claims and $M(n-n_0) = 5 \times 4,664 = 23,320$ synthetic policies, resulting in $n^+ = 28,656$ augmented policies in total. We implement the proposed method with the regression functions

$$\begin{aligned}
\log \mathbb{E}(N_i) &= \log e_i + \log \lambda(\boldsymbol{x}_i; \beta) \\
&= \log e_i + \beta^\top \boldsymbol{x}_i \\
&= \log e_i + \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}, \\
\log \mathbb{E}(Y_i) &= \log N_i - \log e_i + \log \tau(\boldsymbol{x}_i; \gamma) \\
&= \log N_i - \log e_i + \gamma^\top \boldsymbol{x}_i \\
&= \log N_i - \log e_i + \gamma_0 + \gamma_1 x_{i,1} + \gamma_2 x_{i,2} + \gamma_3 x_{i,3} + \gamma_4 x_{i,4} + \gamma_5 x_{i,5},
\end{aligned} \tag{4.2}$$

where the logarithm link function is assumed, and $\log e_i$ and $\log N_i - \log e_i$ are treated as the offsets in the linear predictors for $N$ and $Y$, respectively. The algorithm converges in six iterations, as shown in Figure 2. The sum of conditional probabilities over the augmented data is

$$\sum_{i=1}^{n^+=28,656} f_{N|Y=y_i}(m_i; \hat{\theta}) = 9,997.72 \approx n,$$

justifying the selection of $M = 5$. The estimated parameters are listed in Table 1. The MLEs of Tweedie parameters $\mu_i, \phi_i, p$ can be derived using (2.5). The power variance parameter is estimated as $\hat{p} = 1.4992$, which is quite close to the true $p = 1.5$.

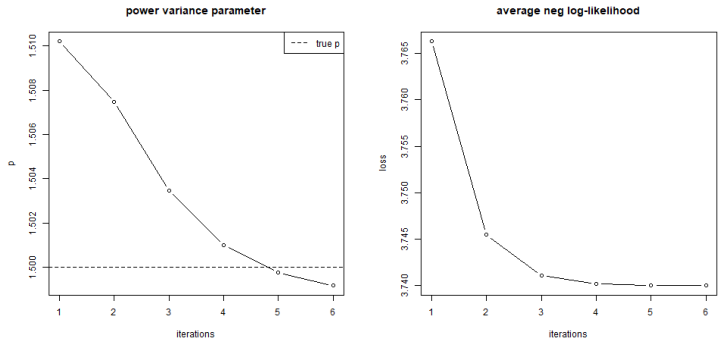



Figure 2: The trace plots of the power variance parameter and the average negative log-likelihood on the training data during the EM iterations.

We investigate the out-of-sample performance of the calibrated model on the test $n^* = 10,000$ policies. We calculate the average negative log-likelihood of Tweedie distribution (2.3) for the test data $(Y_i, \boldsymbol{x}_i, e_i)_{i=1:n^*}$,

$$\mathcal{L}(\boldsymbol{y}; \hat{\theta}) = -\frac{1}{10,000} \sum_{i=1}^{n^*=10,000} \log(f_Y(y_i; \hat{\mu}_i, \hat{\phi}_i, \hat{p})) = 3.6907,$$

which is called test loss in the following; see Table 2. We also calculate the Pearson/ Kendall/ Spearman correlations of the estimated $\hat{\mu}_i, \hat{\phi}_i$ with true $\mu_i, \phi_i$, respectively, which are listed in



Electronic copy available at: https://ssrn.com/abstract=4148316

Table 1: The estimated parameters by the EM algorithm. The shape parameter is estimated as $\hat{\alpha} = 1.0033$.

| parameters | estimate | std. error | $z$ value | $\Pr(>\|z\|)$ |
|---|---|---|---|---|
| $\beta_0$ | -1.9575 | 0.0169 | -115.5642 | 0.0000 |
| $\beta_1$ | 0.2970 | 0.0248 | 11.9609 | 0.0000 |
| $\beta_2$ | -0.3002 | 0.0248 | -12.0881 | 0.0000 |
| $\beta_3$ | -0.2798 | 0.0253 | -11.0391 | 0.0000 |
| $\beta_4$ | 0.2687 | 0.0245 | 10.9651 | 0.0000 |
| $\beta_5$ | -0.2670 | 0.0249 | -10.7005 | 0.0000 |
| $\gamma_0$ | 6.8825 | 0.0076 | 903.3589 | 0.0000 |
| $\gamma_1$ | -0.0084 | 0.0112 | -0.7521 | 0.4520 |
| $\gamma_2$ | 0.1620 | 0.0113 | 14.3333 | 0.0000 |
| $\gamma_3$ | 0.1229 | 0.0112 | 10.9297 | 0.0000 |
| $\gamma_4$ | 0.0153 | 0.0109 | 1.3988 | 0.1619 |
| $\gamma_5$ | 0.0052 | 0.0112 | 0.4648 | 0.6421 |

Table 2. The correlations are high, indicating that the EM algorithm performs well. Note that those correlations are only possible for the simulated data since the true pure premium $\mu_i$ and dispersion $\phi_i$ are needed.

We consider two competing models: a single Tweedie GLM and a Tweedie DGLM. In the single Tweedie GLM, we assume a log-linear regression function for $\mathbb{E}(Y_i)$:

$$\log \mathbb{E}(Y_i) = \delta^\top \boldsymbol{x}_i = \delta_0 + \delta_1 x_{i,1} + \delta_2 x_{i,2} + \delta_3 x_{i,3} + \delta_4 x_{i,4} + \delta_5 x_{i,5}, \tag{4.3}$$

with weight of $e_i$. The power variance parameter is estimated as 1.51 by the profile loss, as shown in Figure 3. The out-of-sample metrics are listed in Table 2. The test loss 3.7057 is larger than that in the EM approach, while the predicted pure premiums are comparable with the EM approach. In the Tweedie DGLM, two GLMs for mean and dispersion are fitted alternatively. For mean modelling, a Tweedie GLM is fitted to the observed pure premium with regression function (4.3) and weight of $e_i/\phi_i$. For dispersion modelling, a gamma GLM is fitted to the leverage-adjusted unit deviance $d_i^* = d_i/(1-h_i)$ with mean of $\mathbb{E}(d_i^*) = \exp(\xi^\top \boldsymbol{x}_i)$ and weight of $1-h_i$, where

$$d_i = 2e_i \left( y_i \frac{y_i^{1-p} - \mu_i^{1-p}}{1-p} - \frac{y_i^{2-p} - \mu_i^{2-p}}{2-p} \right)$$

is the unit deviance and $h_i$ is the leverage for the Tweedie GLM. Please refer to Smyth and Jørgensen (2002), Smyth et al. (2001) and Smyth and Verbyla (1999) for more detail on the DGLM and restricted MLE. The power variance parameter is estimated as 1.50 by the profile loss, as shown in Figure 3. The out-of-sample metrics are listed in Table 2. The test loss 3.6969 is smaller than that in the single GLM but larger than that in the EM approach. We conclude that those three modelling approaches return a comparable prediction for the pure premium, and the EM approach provides a better estimation of the dispersion than the DGLM, resulting in the smallest test loss.

As a benchmark, we conduct a Poisson-gamma modelling given $N$ is observed. The regression functions are assumed as the same as (4.2). The out-of-sample metrics are listed in Table 2,



Electronic copy available at: https://ssrn.com/abstract=4148316

Table 2: The test loss and three types of correlation of the estimated pure premium (or the estimated dispersion) with the true pure premium (or the true dispersion) in case 1.

| model | test loss | correlation with true mean | | | correlation with true dispersion | | |
|---|---|---|---|---|---|---|---|
| | | Pearson | Kendall | Spearman | Pearson | Kendall | Spearman |
| EM | **3.6907** | 0.8818 | **0.7398** | 0.8776 | **0.9317** | **0.8531** | **0.9699** |
| single GLM | 3.7057 | 0.8818 | 0.7335 | **0.8795** | NA | NA | NA |
| DGLM | 3.6969 | **0.8822** | 0.7368 | 0.8790 | 0.8798 | 0.698 | 0.8733 |
| Poisson-gamma | 3.6908 | 0.8820 | 0.7399 | 0.8777 | 0.9287 | 0.8565 | 0.9700 |
| true | 3.6887 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

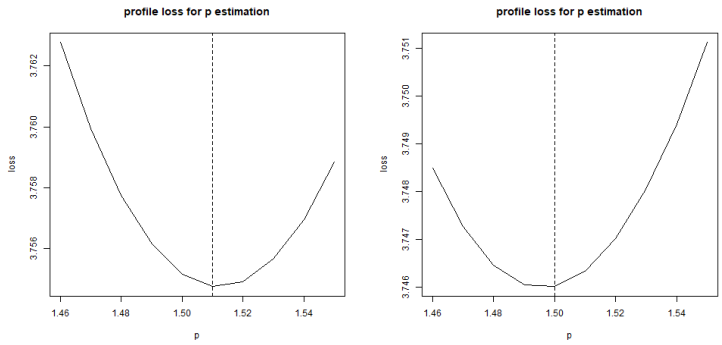



Figure 3: The profile loss plots for estimating power variance parameter in the single GLM and the DGLM in case 1.

which are comparable to those from the EM approach. Thus, our proposed model-fitting method can reasonably recover the unobserved number of claims by augmenting data with synthetic policies and suitable weights. For comparison, the true test loss is calculated as 3.6887, which is undoubtedly smaller than those based on estimated parameters.

#### 4.1.2 Case 2: Small exposure

Suppose that exposure $e_i$ is generated from a uniform distribution in $(0, 2)$. The proportion of zero claims is 87.91%, much higher than that in case 1 of large exposure. We anticipate that the DGLM might not perform well in this case. Since fewer policies have made claims, we obtain an $n^+ = 14,836$ augmented data set smaller than in case 1 .

We implement the same four modelling approaches as in case 1. In the EM approach, the sum of conditional probabilities is $10,000$, justifying the selection of $M = 5$. The power variance parameter is estimated as 1.50. In the single GLM and DGLM, the power variance parameter is estimated as 1.51 and 1.37, respectively; see Figure 4.

The out-of-sample metrics are summarized in Table 3. The EM approach is much better than the two competing models in terms of all the out-of-sample metrics. The EM approach is only slightly worse than the Poisson-gamma model, which uses more data than the EM approach. The single GLM is better than the DGLM in terms of test loss and pure premium prediction. As we anticipated, the DGLM cannot model the dispersion appropriately since the saddlepoint approximation is not accurate in the case of a large proportion of zero claims. Indeed, the DGLM estimates the dispersion in a wrong direction, leading to worse test loss and



Electronic copy available at: https://ssrn.com/abstract=4148316

pure premium prediction than the single GLM.

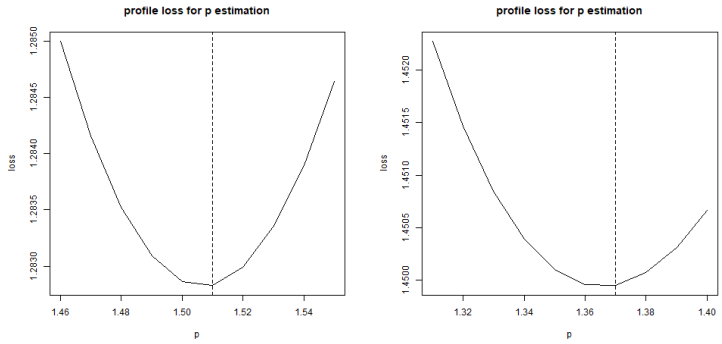



Figure 4: The profile loss plots for estimating power variance parameter in the single GLM and the DGLM in case 2.

Table 3: The test loss and three types of correlation of the estimated pure premium (or the estimated dispersion) with the true pure premium (or the true dispersion) in case 2.

| model | test loss | correlation with true mean | | | correlation with true dispersion | | |
|---|---|---|---|---|---|---|---|
| | | Pearson | Kendall | Spearman | Pearson | Kendall | Spearman |
| EM | **1.1638** | **0.8528** | **0.6842** | **0.8474** | **0.9354** | **0.8361** | **0.9645** |
| single GLM | 1.1678 | 0.8458 | 0.6728 | 0.8400 | NA | NA | NA |
| DGLM | 1.3481 | 0.8426 | 0.6681 | 0.8362 | -0.652 | -0.5079 | -0.6987 |
| Poisson-gamma | 1.1637 | 0.8535 | 0.6854 | 0.8484 | 0.9355 | 0.8238 | 0.9584 |
| true | 1.1629 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

## 4.2 Real data example

We analyse the `swmotorcycle` data in the `CASdatasets` R package. The data contains comprehensive insurance for motorcycles covering loss due to theft, fire or vandalism rather than collision. The data is aggregated on each policy from 1994 to 1998. This data has been analysed by Ohlsson and Johansson (2010) and Delong et al. (2021). Although the number of claims is available in this data set, we do not use this information in the following analysis.

### 4.2.1 Data description and preprocessing

In this data, there are six risk factors available including `OwnerAge, Gender, Area, RiskClass, VehAge` and `BonusClass`, and two loss variables available including `ClaimNb` and `ClaimAmount`. We briefly explain the meaning of each variable and how to preprocess them; we also refer to the appendix of Wüthrich and Merz (2022). The age of driver `OwnerAge` is censored at 70 due to sparsity beyond that age. The gender of driver `Gender` has two levels: `Female` and `Male`. There are seven Swedish geographical areas `Area`, and we merge areas 5-7 since there are few exposures in those areas. Seven ordered motorcycle classes `RiskClass` are obtained from the so-called EV ratio defined as (Engine power in KW $\times 100$)/(Vehicle weight in kg + 75kg). We merge levels 6 and 7 due to sparsity in level 7. The age of motorcycle `VehAge` is censored at 20



Electronic copy available at: https://ssrn.com/abstract=4148316

due to few exposures beyond. There are seven ordered bonus-malus classes `BonusClass` with an entry-level of 1. Each policy's total exposure `Exposure` ranges from 0.0027 (one day) to 31.3397 years. The number of claims `ClaimNb` on each policy ranges from 0 to 2. The total claim costs `ClaimAmount` on each policy ranges from 0 to 365,347 Swedish crowns (SEK).

After data cleaning, there are $n = 62,036$ policies with an average exposure of 1.0441 years, among which 98.94% policies have not made any claims. The empirical claim frequency is $\sum_{i=1}^n N_i / \sum_{i=1}^n e_i = 1.05\%$. The empirical average claim size is $\sum_{i=1}^n S_i / \sum_{i=1}^n N_i = 24,641.35$ SEK. The empirical pure premium is $\sum_{i=1}^n S_i / \sum_{i=1}^n e_i = 259.83$ SEK. We assume that the number of claims is not observed, and we only observe the occurrence of claims denoted by $O_i = \mathbb{1}_{(0,\infty)}(Y_i)$. The empirical claim occurrence rate is $\sum_{i=1}^n O_i / \sum_{i=1}^n e_i = 1.01\%$. The empirical average total claim amount is $\sum_{i=1}^n S_i / \sum_{i=1}^n O_i = 25,655.55$ SEK. The empirical pure premium can still be obtained as 259.83 SEK without knowing the number of claims. Note that the empirical average total claim amount is close to the empirical average claim size since very few policies make more than one claim.

We conduct a preliminary analysis of the marginal effects of covariates on the claim occurrence rate and the total claim amount; see Figures 10 and 11 in the appendix. It shows that `OwnerAge` has a non-monotone effect on the claim occurrence rate, while `Area, RiskClass, VehAge` has a roughly monotone effect on the claim occurrence rate. `Gender` has a minor effect on the claim occurrence rate, and the effect of `BonusClass` is also weak since `BonusClass` relates to the collision claims rather than the comprehensive coverage considered in this data. We have a similar observation for the total claim amount. Therefore, in the following analysis, we include `OwnerAge, OwnerAge`$^2$`, Area, RiskClass, VehAge, BonusClass` as continuous variables and `Gender` as a binary variable in the linear predictor. We will not perform a generalization analysis since only 656 policies have made claims. Remark that although `Area` is a categorical variable, we can interpret its levels as ordinal and map them to integers.

### 4.2.2 Comparison of different modelling approaches

As a benchmark, we fit a Poisson-gamma model to the number of claims $N_i$ and the observed pure premium $Y_i$ given $N_i$. The regression functions are specified as follows:

$$\begin{aligned}
\log \mathbb{E}(N_i) &= \log e_i + \log \lambda(\boldsymbol{x}_i; \beta) \\
&= \log e_i + \beta^\top \boldsymbol{x}_i \\
&= \log e_i + \beta_0 + \beta_1 \texttt{OwnerAge}_i + \beta_2 \texttt{OwnerAge}_i^2 + \beta_3 \texttt{Area}_i + \\
&\quad \beta_4 \texttt{RiskClass}_i + \beta_5 \texttt{VehAge}_i + \beta_6 \texttt{BonusClass}_i + \beta_7 \texttt{Gender}_i, \\
\log \mathbb{E}(Y_i) &= \log N_i - \log e_i + \log \tau(\boldsymbol{x}_i; \gamma) \\
&= \log N_i - \log e_i + \gamma^\top \boldsymbol{x}_i \\
&= \log N_i - \log e_i + \gamma_0 + \gamma_1 \texttt{OwnerAge}_i + \gamma_2 \texttt{OwnerAge}_i^2 + \gamma_3 \texttt{Area}_i + \\
&\quad \gamma_4 \texttt{RiskClass}_i + \gamma_5 \texttt{VehAge}_i + \gamma_6 \texttt{BonusClass}_i + \gamma_7 \texttt{Gender}_i.
\end{aligned} \tag{4.4}$$

The regression coefficients $\beta, \gamma$ are estimated in two separate GLMs. The MLE of shape parameter is calculated as 0.6870; thus, the power variance parameter is estimated as 1.5928. The average negative log-likelihood of Tweedie distribution (2.3) is 0.1724.

Without knowing the number of claims $N_i$, we implement the proposed model-fitting method with the same regression functions as (4.4). We obtain an $n^+ = 64,660$ augmented data set



Electronic copy available at: https://ssrn.com/abstract=4148316

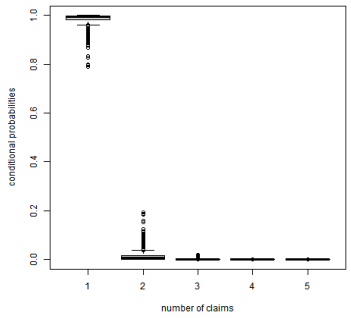



Figure 5: The boxplot of conditional probabilities $f_{N|Y=y_i}(m_i)$ for positive $y_i > 0$.

containing $n_0 = 61,380$ policies without claims and $5 \times (n - n_0) = 3,280$ synthetic policies with claims. The EM algorithm converges in 3 iterations. The sum of conditional probabilities $f_{N|Y=y_i}(m_i)$ on the augmented data is $62,036$ justifying the selection of $M = 5$. We draw the boxplot of conditional probabilities $f_{N|Y=y_i}(m_i)$ for positive $y_i$ w.r.t. $m_i \in \{1, \ldots, M = 5\}$ in Figure 5. It shows that given $Y > 0$, there is a very high chance of only 1 claim occurring, which is parallel to the observed number of claims. The estimated parameters are listed in Tables 4 and 5. As indicated in the preliminary analysis, `Gender` and `BonusClass` are less significant than the other covariates in both Poisson and gamma GLMs. The shape parameter is estimated as 0.6856; thus, the power variance parameter is estimated as 1.5933, which is quite close to that in the benchmark Poisson-gamma model. The average negative log-likelihood is 0.1724, the same as that in the benchmark Poisson-gamma model. The correlation of estimated pure premiums with those estimated from the benchmark Poisson-gamma model is nearly one (also for the dispersion), as shown in Table 6. Therefore, our proposed method performs as well as the benchmark Poisson-gamma model even without knowing the number of claims.

Table 4: The estimated parameters $\hat{\beta}$ in the Poisson GLM by the EM algorithm.

| variables | estimate | std. error | $z$ value | $\Pr(>\|z\|)$ |
|---|---|---|---|---|
| Intercept | 1.2622 | 0.4145 | 3.0452 | 0.0023 |
| `OwnerAge` | -0.2080 | 0.0213 | -9.7708 | 0.0000 |
| $\texttt{OwnerAge}^2$ | 0.0019 | 0.0003 | 7.1541 | 0.0000 |
| `Area` | -0.4547 | 0.0321 | -14.1589 | 0.0000 |
| `RiskClass` | 0.1850 | 0.0295 | 6.2806 | 0.0000 |
| `VehAge` | -0.0806 | 0.0070 | -11.5810 | 0.0000 |
| `BonusClass` | 0.0503 | 0.0184 | 2.7353 | 0.0062 |
| `GenderMale` | 0.3093 | 0.1360 | 2.2743 | 0.0229 |

For comparison, a single Tweedie GLM is fitted to the observed pure premium with regression function

$$\begin{aligned}\log \mathbb{E}(Y) =& \delta^\top \boldsymbol{x}_i \\ =& \delta_0 + \delta_1 \texttt{OwnerAge}_i + \delta_2 \texttt{OwnerAge}_i^2 + \delta_3 \texttt{Area}_i + \\ & \delta_4 \texttt{RiskClass}_i + \delta_5 \texttt{VehAge}_i + \delta_6 \texttt{BonusClass}_i + \delta_7 \texttt{Gender}_i.\end{aligned} \tag{4.5}$$

The estimated parameters are listed in Table 7. There are only two significant covariates `Area`



Electronic copy available at: https://ssrn.com/abstract=4148316

Table 5: The estimated parameters $\hat{\gamma}$ in the gamma GLM by the EM algorithm.

| variables | estimate | std. error | $z$ value | $\Pr(>\|z\|)$ |
|---|---|---|---|---|
| Intercept | 8.4504 | 0.2616 | 32.3018 | 0.0000 |
| `OwnerAge` | 0.1290 | 0.0134 | 9.5987 | 0.0000 |
| `OwnerAge`$^2$ | -0.0017 | 0.0002 | -10.1191 | 0.0000 |
| `Area` | -0.0673 | 0.0189 | -3.5641 | 0.0004 |
| `RiskClass` | 0.0132 | 0.0155 | 0.8548 | 0.3927 |
| `VehAge` | -0.0889 | 0.0041 | -21.8308 | 0.0000 |
| `BonusClass` | -0.0016 | 0.0110 | -0.1430 | 0.8863 |
| `GenderMale` | 0.1635 | 0.0807 | 2.0251 | 0.0429 |

Table 6: The training loss and three types of correlation of the estimated pure premium (or the dispersion) with those estimated from the benchmark Poisson-gamma model.

| model | training loss | mean correlation | | | dispersion correlation | | |
|---|---|---|---|---|---|---|---|
| | | Pearson | Kendall | Spearman | Pearson | Kendall | Spearman |
| EM | **0.1724** | **1.0000** | **0.9975** | **1.0000** | **0.9986** | **0.9719** | **0.9989** |
| single GLM | 0.1908 | 0.9739 | 0.8979 | 0.9859 | NA | NA | NA |
| Poisson-gamma | 0.1724 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

and `VehAge` in the model. This may be due to the opposite effects of covariates on the frequency and severity; see the estimated parameters for `OwnerAge` and `OwnerAge`$^2$ in Tables 4 and 5. The power variance parameter is estimated as 1.6300. We point out that the `glm` function with `tweedie` family in R does not converge for $p \geq 1.64$. We have not investigated the reason since we now have a more stable EM approach. The average negative log-likelihood is 0.1908, higher than that in the EM approach. Table 6 shows that the correlations of estimated pure premiums among the three modelling approaches are very high because the dispersion has a secondary effect on the mean estimation, and we use the same parametric regression function for mean.

Table 7: The estimated parameters $\hat{\delta}$ in the single Tweedie GLM.

| variables | estimate | std. error | $t$ value | $\Pr(>\|t\|)$ |
|---|---|---|---|---|
| Intercept | 10.6777 | 1.7435 | 6.1244 | 0.0000 |
| `OwnerAge` | -0.1035 | 0.0851 | -1.2158 | 0.2241 |
| `OwnerAge`$^2$ | 0.0005 | 0.0010 | 0.4357 | 0.6630 |
| `Area` | -0.5063 | 0.1275 | -3.9718 | 0.0001 |
| `RiskClass` | 0.0865 | 0.1133 | 0.7639 | 0.4449 |
| `VehAge` | -0.1453 | 0.0247 | -5.8803 | 0.0000 |
| `BonusClass` | 0.0858 | 0.0679 | 1.2623 | 0.2068 |
| `GenderMale` | 0.0429 | 0.4734 | 0.0905 | 0.9279 |

The DGLM approach does not work for this real data, at least by implementing the `glm` function with the `tweedie` family in R. We have tried several power variance parameters $p \in (1.55, 1.65)$, however, alternatively fitting two GLMs does not decrease the loss. We refrain from investigating the reason. As a conclusion, this real data example shows that the proposed EM



Electronic copy available at: https://ssrn.com/abstract=4148316

approach is more stable than the Tweedie GLM approach.

# 5 Conclusions

This paper aims to improve model fitting of the marginal Tweedie's compound Poisson model for pure premium when number of claims is unavailable. Currently, the single GLM and the DGLM are based on the GLM framework since Tweedie distribution belongs to the exponential distribution family. However, both models are not satisfactory theoretically or practically.

We propose a new model-fitting method using the EM algorithm. It is equivalent to iteratively re-weighted Poisson-gamma regressions on an augmented data set. Thus, the proposed method can be easily implemented. The proposed method overcomes three weaknesses of the single GLM and the DGLM. First, the proposed method performs well under either small or large proportions of zero claims. Second, the power variance parameter is estimated directly in the EM iterations. Third, the proposed method is more stable, as shown in the real data example. The numerical examples also illustrate that the dispersion has a secondary effect on the mean estimation. So if one is only interested in the pure premium prediction, any of the three models should be sufficient. However, if one is also interested in the dispersion of future loss, our proposed method provides the most accurate dispersion estimation.



Electronic copy available at: https://ssrn.com/abstract=4148316

# A Figures

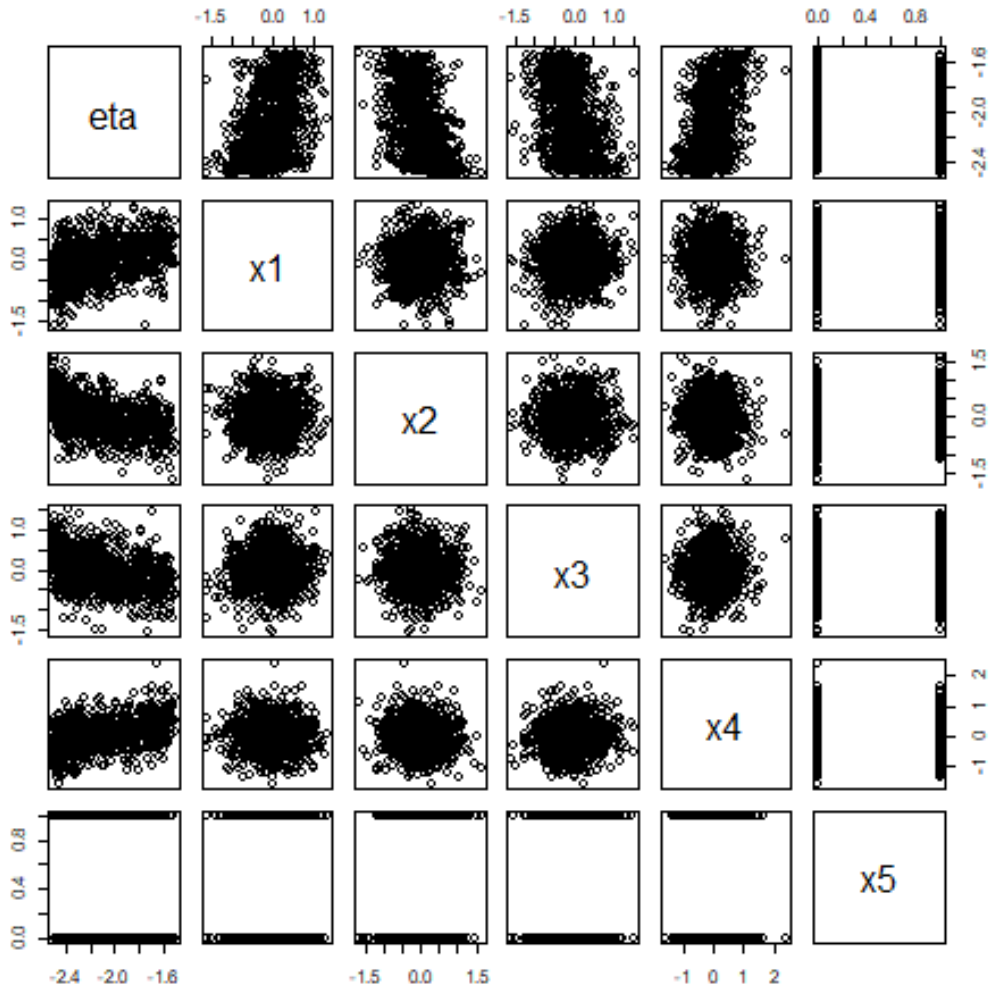



Figure 6: The pair plot for the linear predictor $\eta_i$.



Electronic copy available at: https://ssrn.com/abstract=4148316

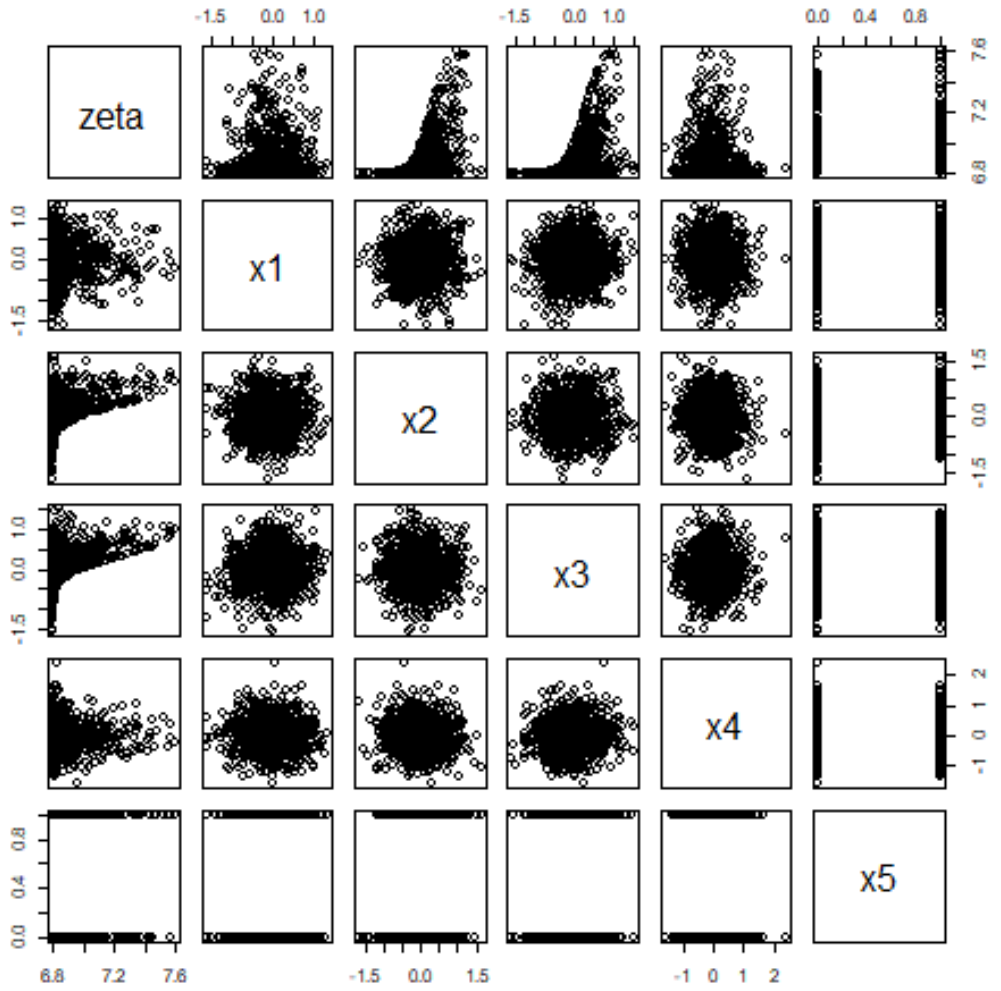



Figure 7: The pair plot for the linear predictor $\zeta_i$.



Electronic copy available at: https://ssrn.com/abstract=4148316

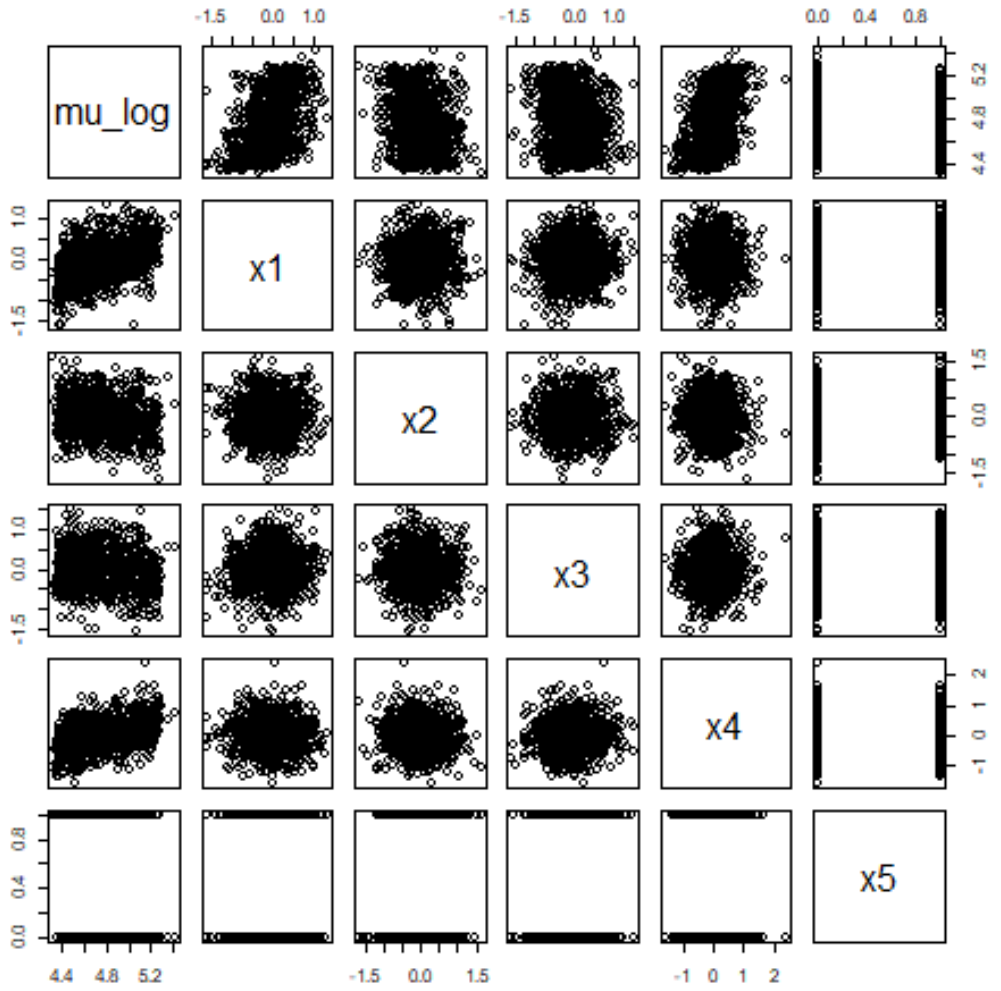



Figure 8: The pair plot for the logged $\mu_i$.



Electronic copy available at: https://ssrn.com/abstract=4148316

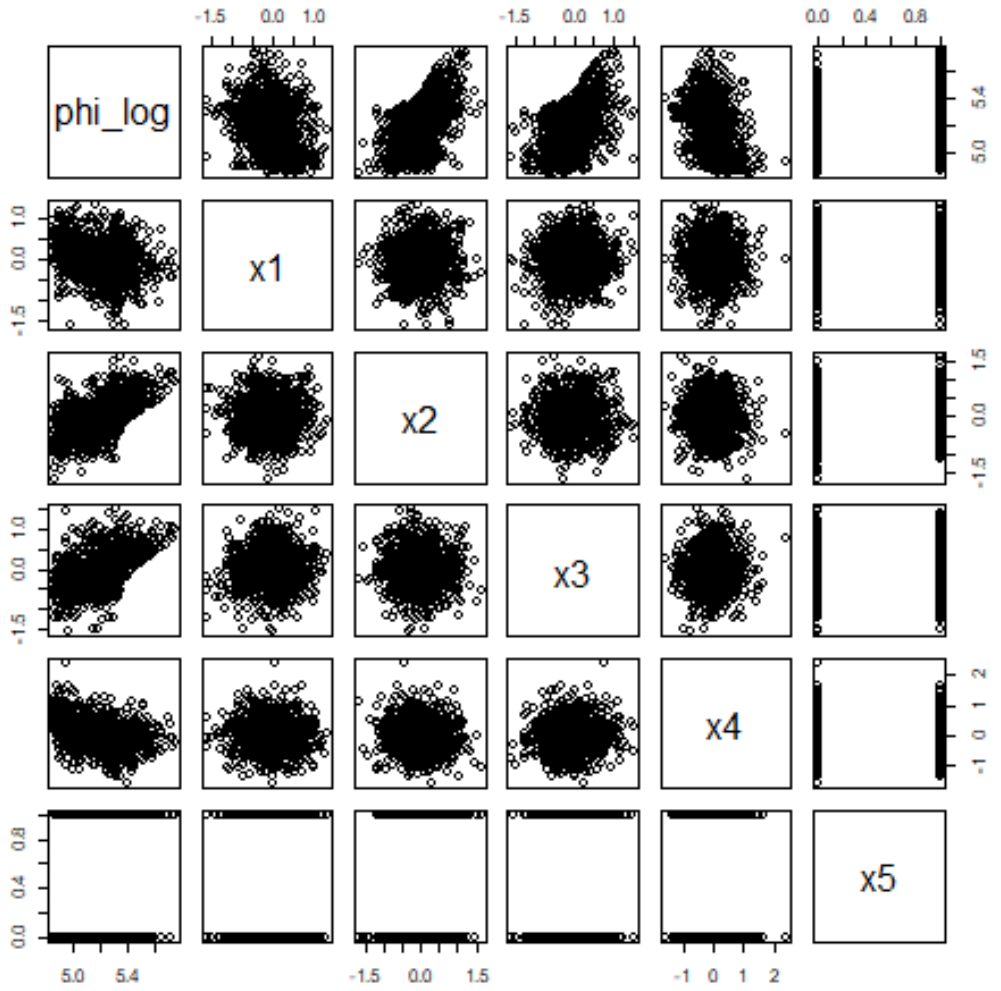



Figure 9: The pair plot for the logged $\phi_i$.



Electronic copy available at: https://ssrn.com/abstract=4148316

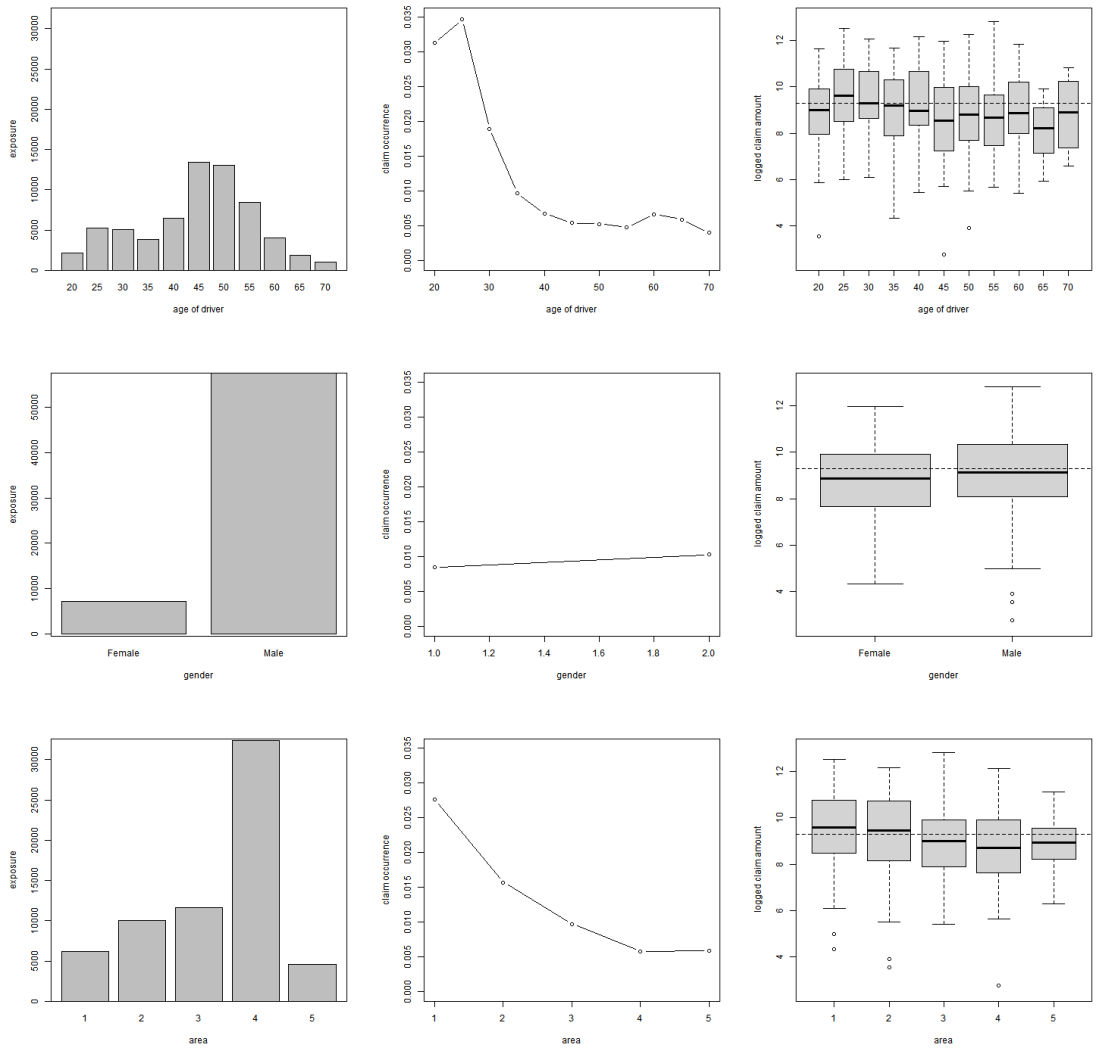



Figure 10: The histogram of exposure (left), the empirical claim occurrence rate (middle), and the boxplot of logged total claim amount (right) w.r.t. `OwnerAge, Gender, Area`.



Electronic copy available at: https://ssrn.com/abstract=4148316

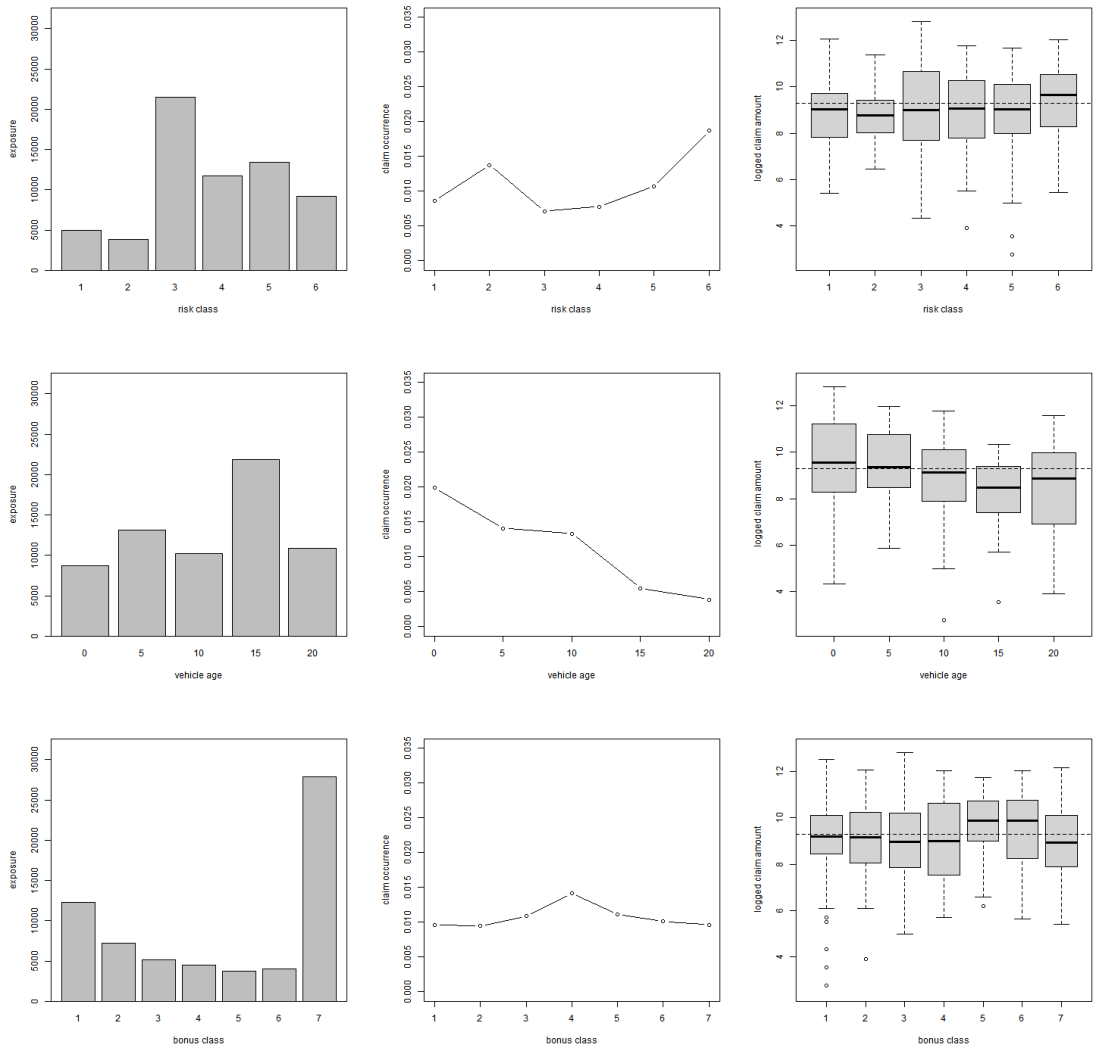



Figure 11: The histogram of exposure (left), the empirical claim occurrence rate (middle), and the boxplot of logged total claim amount (right) w.r.t. `RiskClass, VehAge, BonusClass`.



Electronic copy available at: https://ssrn.com/abstract=4148316

## Acknowledgment

Guangyuan Gao gratefully acknowledges the financial support from the National Natural Science Foundation of China (71901207).

## References

Delong, Ł., Lindholm, M. and Wüthrich, M. V. (2021). Making Tweedie's compound Poisson model more accessible, *European Actuarial Journal* **11**(1): 185–226.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1): 1–22.

Jørgensen, B. (1997). *The Theory of Dispersion Models*, CRC Press.

Jørgensen, B. and De Souza, M. C. P. (1994). Fitting Tweedie's compound Poisson model to insurance claims data, *Scandinavian Actuarial Journal* **1994**(1): 69–93.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function, *Biometrika* **74**(2): 221–232.

Ohlsson, E. and Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*, Vol. 2, Springer.

Smyth, G. K. (1989). Generalized linear models with varying dispersion, *Journal of the Royal Statistical Society: Series B (Methodological)* **51**(1): 47–60.

Smyth, G. K., Huele, A. F. and Verbyla, A. P. (2001). Exact and approximate REML for heteroscedastic regression, *Statistical modelling* **1**(3): 161–175.

Smyth, G. K. and Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling, *ASTIN Bulletin: The Journal of the IAA* **32**(1): 143–157.

Smyth, G. K. and Verbyla, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models, *Environmetrics: The official journal of the International Environmetrics Society* **10**(6): 695–709.

Tweedie, M. C. et al. (1984). An index which distinguishes between some important exponential families, *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, Vol. 579, pp. 579–604.

Wüthrich, M. V. and Merz, M. (2022). Statistical foundations of actuarial learning and its applications, *Available at SSRN 3822407* .



Electronic copy available at: https://ssrn.com/abstract=4148316