Les Cahiers du GERAD ISSN: 0711–2440

Generalized linear models for dependent frequency and severity of insurance claims

J. Garrido, C. Genest,
J. Schulz

G–2015–138

December 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Generalized linear models for dependent frequency and severity of insurance claims

José Garrido [a]
Christian Genest [b]
Juliana Schulz [b]

[a] *Department of Mathematics and Statistics, Concordia University, Montréal (Québec) Canada, H3G 1M8*

[b] *GERAD & Department of Mathematics and Statistics, McGill University, Montréal (Québec) Canada, H3A 0B9*

jose.garrido@concordia.ca
cgenest@math.mcgill.ca
juliana.schulz@mail.mcgill.ca

December 2015

Les Cahiers du GERAD
G–2015–138

Abstract: Traditionally, claim counts and amounts are assumed to be independent in non-life insurance. This paper explores how this oft unwarranted assumption can be relaxed in a simple way while incorporating rating factors into the model. The approach consists of fitting generalized linear models to the marginal frequency and the conditional severity components of the total claim cost; dependence between them is induced by treating the number of claims as a covariate in the model for the average claim size. This model is both easy to implement and has the advantage that when Poisson counts are assumed together with a log-link for the conditional severity model, the resulting pure premium can be expressed as the product of a marginal mean frequency, a modified marginal mean severity, and an easily interpretable correction term that reflects the dependence. The approach is illustrated through simulations and applied to a Canadian automobile insurance dataset.

Key Words: Aggregate claims model, claim frequency, claim severity, dependence, exponential dispersion models, generalized linear model, loss cost.

Acknowledgments: This work was supported by the Canada Research Chairs Program, scholarships and grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN/39476–2011, RGPIN/36860–2012), the Canadian Statistical Sciences Institute, and the Fonds de recherche du Québec – Nature et technologies (2015–PR–183236).

1 Introduction

In order to price their contracts adequately, property casualty insurers need to rely on an accurate estimation of all future costs associated with the insurance protection they provide. To this end, they usually consider claim counts and amounts separately. Generalized linear models (GLMs) are useful in this context (Renshaw, 1994) because the means of the frequency and severity processes can then be expressed, through specific transforms, as linear combinations of rating variables such as age, sex, etc.

This modeling approach implicitly assumes that the claim counts and amounts are independent. The pure premium at the individual or class level is then simply the product of the two mean estimates. Alternatively, the total loss cost is sometimes modeled directly by means of the Tweedie distribution. This approach, which is reviewed by Quijano-Xacur and Garrido (2015), amounts to modeling the aggregate claims as a compound Poisson-Gamma sum. As emphasized by these authors, however, this method also implicitly assumes independence between claim counts and individual claim sizes.

In practice, claim frequency and severity are often dependent. For example, home insurance claims due to sewer backup or flooding tend to be both large and frequent in problematic neighborhoods, while claim counts and amounts are often negatively associated in collision automobile insurance because drivers who file several claims per year are typically involved in minor accidents. There is thus a need to adapt the aggregate claims model to account for potential association between claim frequency and severity.

Two general approaches have been proposed to account for dependence between frequency and severity. Frees and Wang (2006), Gschlößl and Czado (2007) and Frees et al. (2012) proceed by conditioning and use the claim count as a covariate in modeling the average claim size distribution. In contrast, Czado et al. (2012) and Krämer et al. (2013) fit marginal GLMs to the frequency and severity components and link them through a copula. See Shi et al. (2015) for a comparison of these two approaches.

In this paper, a conditional approach is used to model the dependence between the claim counts and amounts of the aggregate losses process. Proceeding along similar lines as Frees et al. (2012), we postulate GLMs both for the marginal frequency and the conditional severity components. Whereas these authors use this approach to model health care expenditures arising from panel data, we explore its merits from a ratemaking perspective for a typical property and casualty insurance portfolio.

In particular, it will be seen that when Poisson counts are assumed and a log-link is used in the conditional severity model, the pure premium resulting from this general approach can be written as the product of three terms, i.e., the marginal mean frequency, a modified marginal mean severity, and a dependence correction term. The latter is indexed by a real-valued parameter that accounts for the association between the frequency and severity components of the model. This method is especially convenient from a practitioner's point of view, as it is a simple extension to the common approach used in industry. It can be readily implemented with standard tools and makes it easy to account for dependence within the aggregate claims process.

This paper is organized as follows. In Section 2, basic facts about the independent aggregate claims model are recalled for completeness. A simple adaptation of this model which accounts for dependence between claim counts and amounts is then described in Section 3. The impact of dependence on the total losses is briefly illustrated in Section 4 through a simulation study, and Section 5 presents the highlights of an original application of this model to automobile claims data from a large Canadian insurance company. Full details of this analysis can be found in the MSc thesis of Schulz (2013), where the conditional GLM approach was considered independently. Our main findings are summarized and discussed in Section 6.

2 The independent aggregate claims model

The aggregate losses incurred by an insurer is the total amount paid out in claims over a fixed time period. For a given class of policyholders, this amount can be expressed as

$$S = \sum_{j=1}^{N} Y_j,$$

where N is the number of claims incurred, and for $j \in \{1, \ldots, N\}$, Y_j is the jth incurred claim amount. Both N and $Y_1, \ldots, Y_N$ are taken to be random and, by convention, $S \equiv 0$ when $N = 0$.

Traditionally, it is assumed that conditional on N, the individual claim amounts $Y_1, \ldots, Y_N$ are mutually independent, identically distributed, and such that their distribution does not depend on N. It is further assumed that the distribution of N does not depend on the values of the claim amounts. Letting Y denote a generic severity random variable, one then has

$$\mathrm{E}(S) = \mathrm{E}(N)\mathrm{E}(Y), \quad \mathrm{var}(S) = \mathrm{E}(N)\mathrm{var}(Y) + \mathrm{var}(N)\mathrm{E}(Y)^2.$$

When a vector $x = (x_1, \ldots, x_p)$ of covariates is available at the individual level, this information can be incorporated into the ratemaking process through separate GLMs for N and Y. This amounts to assuming that for given link functions g_N and g_Y,

$$\nu = \mathrm{E}(N \mid x) = g_N^{-1}(x\gamma), \quad \mu = \mathrm{E}(Y \mid x) = g_Y^{-1}(x\beta),$$

where β and γ are $p \times 1$ vectors of regression coefficients. Owing to the independence between frequency and severity, one then has

$$\mathrm{E}(S \mid x) = \nu\mu = g_N^{-1}(x\gamma) \times g_Y^{-1}(x\beta).$$

In particular when both marginal GLMs use a log-link, this reduces to

$$\mathrm{E}(S \mid x) = \nu\mu = \exp(x\gamma + x\beta). \tag{1}$$

This link function ensures that $\mu > 0$ and $\nu > 0$; it also has the advantage of yielding a simple rating structure reflecting the common practice that each covariate alters the baseline rate by a multiplicative factor.

Suppose in addition that the marginal distributions of the aggregate claims model are from the exponential dispersion family (EDF). Then

$$\mathrm{var}(S \mid x) = \nu\phi V_Y(\mu) + \psi V_N(\nu)\mu^2,$$

where V_N and V_Y are the variance functions for the frequency and severity components with dispersion parameters ψ and ϕ, respectively.

Example 2.1 Suppose N is Poisson with mean $\lambda > 0$, denoted $N \sim \mathcal{P}(\lambda)$. Suppose also that Y is Gamma with mean $\alpha\delta > 0$ and variance $\alpha\delta^2$, denoted $Y \sim \mathcal{G}(\alpha, \delta)$. Then S has a compound Poisson-Gamma distribution and

$$\mathrm{E}(S) = \lambda\alpha\delta, \quad \mathrm{var}(S) = \lambda\alpha(\alpha + 1)\delta^2.$$

In terms of the canonical parametrizations for these specific EDFs, one has $V_Y(\mu) = \mu^2$, $V_N(\nu) = \nu$, $\psi = 1$ and $\phi > 0$, so that

$$\mathrm{var}(S \mid x) = \nu\mu^2 + \phi\nu\mu^2 = \nu\mu^2(\phi + 1). \tag{2}$$

2.1 Inference

Suppose that claims data are available for m different classes of policyholders. Assume that class $i \in \{1, \ldots, m\}$ has N_i claims and let $x_i = (x_{i1}, \ldots, x_{ip})$ be the vector of rating variables (with the convention that $x_{i1} \equiv 1$). For each claim occurrence $j \in \{1, \ldots, N_i\}$, let Y_{ij} denote the individual claim amount.

When independent GLMs are assumed for the claim counts and amounts, the vector parameters of the frequency and severity components can be estimated separately. Suppose in particular that $N_i \sim \mathcal{P}(\nu_i)$ and $Y_{ij} \sim \mathcal{G}(\mu_i, \phi)$, where the canonical parametrizations of the two EDFs are used. Further assume log-link functions in the GLMs for the claim frequency and severity components so that, for all $i \in \{1, \ldots, m\}$,

$$\nu_i = \exp(x_i\gamma), \quad \mu_i = \exp(x_i\beta),$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ and $\gamma = (\gamma_1, \ldots, \gamma_p)^\top$. The systems of p score equations for β and γ are then given, for all $k \in \{1, \ldots, p\}$, by

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{1}{\phi} \frac{x_{ik}}{\mu_i} (y_{ij} - \mu_i) = 0, \quad \sum_{i=1}^{m} x_{ik}(n_i - \nu_i) = 0,$$

where n_i and y_{ij} are the observed values of N_i and Y_{ij}, respectively. The simple form of these score equations makes the computation of the maximum likelihood estimates, $\hat{\beta}$ and $\hat{\gamma}$, a straightforward task.

Standard asymptotic results imply that, as $n \to \infty$,

$$\hat{\beta} \approx \mathcal{N}\left(\beta, I_\beta^{-1}\right), \quad \hat{\gamma} \approx \mathcal{N}\left(\gamma, I_\gamma^{-1}\right),$$

where I_β and I_γ are $p \times p$ Fisher information matrices, for which consistent estimators are respectively given by

$$I_{\hat{\beta}} = \frac{1}{\phi}\, x x^\top, \quad I_{\hat{\gamma}} = x \operatorname{diag}(\hat{\nu}_1, \dots, \hat{\nu}_m)\, x^\top$$

in terms of the $p \times m$ design matrix x with ith column $x_i^\top$ for all $i \in \{1, \dots, m\}$. Note that in $I_{\hat{\beta}}$, the dispersion parameter ϕ is generally unknown but can be estimated, e.g., by

$$\hat{\phi} = \frac{1}{m-p} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\frac{Y_{ij} - \hat{\mu}_i}{\hat{\mu}_i} \right)^2 .$$

3 The dependent aggregate claims model

Although it is convenient to assume independent GLMs for claim counts and amounts, these variables are often associated in practice. Proceeding along similar lines as Frees et al. (2012), we consider here a conditional approach in which the GLM for the severity component of the aggregate claims models is allowed to depend on the claim count.

For a given class of policyholders, we continue to assume that conditional on N, the individual claim amounts $Y_1, \dots, Y_N$ are mutually independent and identically distributed, and that the distribution of N itself does not depend on the values of the claim amounts. To account for dependence, however, the mean of the severity distribution is allowed to depend on N.

As before, let $x = (x_1, \dots, x_p)$ denote the set of covariates defining a given class. GLMs for N and $\bar{Y} = S/N = (Y_1 + \dots + Y_N)/N$ incorporating the effect of x are considered (the latter when $N > 0$ only). To be specific, assume that for given link functions g_N and g_Y,

$$\nu = \mathrm{E}(N \mid x) = g_N^{-1}(x\gamma), \quad \mu_\theta = \mathrm{E}(\bar{Y} \mid N, x) = g_Y^{-1}(x\beta + \theta N),$$

where β and γ are $p \times 1$ vectors of regression coefficients, and $\theta \in \mathbb{R}$ induces a degree of dependence between claim counts and amounts. Thus unless $\theta = 0$,

$$\mathrm{E}(S \mid x) = \mathrm{E}\{N \mathrm{E}(\bar{Y} \mid N, x) \mid x\} \neq \mathrm{E}(N \mid x)\mathrm{E}(Y \mid x).$$

In particular, when g_Y is a log-link, one gets both

$$\mu_\theta = \exp(x\beta + \theta N) \equiv \mu \exp(\theta N)$$

and

$$\mathrm{E}(S \mid x) = \mathrm{E}\left\{N \mu \exp(\theta N) \mid x\right\} = \mu \frac{\partial}{\partial \theta} M_N(\theta),$$

where $M_N(\theta \mid x)$ is the moment generating function of N. When $\theta = 0$, one has $\mu_\theta = \mu_0 = \mu$ and the model then reduces to the case of independence. Furthermore when $N \sim \mathcal{P}(\nu)$, then for all $\theta \in \mathbb{R}$, one has

$$M_N(\theta) = \exp\{\nu(e^\theta - 1)\},$$

in which case

$$\mathrm{E}(S \mid x) = \nu \mu \exp\left\{\nu(e^\theta - 1) + \theta\right\}. \tag{3}$$

In comparing Equations (1) and (3), one can see that the only difference is the multiplicative factor $\exp\{\nu(e^\theta - 1) + \theta\}$, which can be regarded as a correction term for dependence.

Example 3.1 Suppose that $N \sim \mathcal{P}(\nu)$ and $\bar{Y} \mid N \sim \mathcal{G}(\mu_\theta, \phi_\theta/N)$ whenever $N > 0$. A simple calculation then shows that

$$\operatorname{var}(S \mid X) = \nu\mu^2 \left[\nu \exp\left\{\nu(e^{2\theta} - 1) + 4\theta\right\} - \nu \exp\left\{2\nu(e^{\theta} - 1) + 2\theta\right\} \right. \\ \left. +(\phi_\theta + 1) \exp\left\{\nu(e^{2\theta} - 1) + 2\theta\right\}\right].$$

In particular if $\theta = 0$, then $\phi_0 = \phi$, in which case the variance simplifies to

$$\operatorname{var}(S \mid X) = \nu\mu^2(\phi + 1),$$

which matches the expression (2) in the independent model, as it should.

3.1 Inference

Assume as before that for class $i \in \{1, \ldots, m\}$, N_i claims have been observed and that all policyholders in the class share the same vector $x_i = (x_{i1}, \ldots, x_{ip})$ of rating variables. Let S_i be the total claim size and write $\bar{Y}_i = S_i/N_i$ whenever $N_i > 0$. In view of the GLM structures postulated for the frequency and severity components of the aggregate claims, the means of $\bar{Y}_i|N_i$ and N_i can then be expressed as

$$\nu_i = \exp(x_i\gamma), \quad \mu_{\theta i} = \exp(x_i\beta + n_i\theta)$$

using parameters $\gamma = (\gamma_1, \ldots, \gamma_p)^\top$, $\beta = (\beta_1, \ldots, \beta_p)^\top$, and $\theta \in \mathbb{R}$. Further denote by f_N and $f_{\bar{Y}|N}$ the marginal frequency and conditional severity densities, respectively. The joint log-likelihood is then of the form

$$\ell(\gamma, \beta, \theta) = \sum_{i=1}^{m} \ell_N(\gamma; n_i) + \sum_{i=1}^{m} \ell_{\bar{Y}|N}(\beta, \theta; \bar{y}_i \mid n_i).$$

When $N_i \sim \mathcal{P}(\nu_i)$ and $\bar{Y}_i \mid N_i \sim \mathcal{G}(\mu_{\theta i}, \phi_\theta/N_i)$, the systems of p score equations for β and γ are respectively given, for all $k \in \{1, \ldots, p\}$, by

$$\sum_{i=1}^{m} \frac{n_i}{\phi_\theta} \frac{x_{ik}}{\mu_{\theta i}} (\bar{y}_i - \mu_{\theta i}) = 0, \quad \sum_{i=1}^{m} x_{ik}(n_i - \nu_i) = 0$$

and the additional equation that allows one to determine θ uniquely is

$$\sum_{i=1}^{m} \frac{n_i}{\phi_\theta} \frac{n_i}{\mu_{\theta i}} (y_i - \mu_{\theta i}) = 0.$$

The solutions to these $2p + 1$ equations are denoted $\check{\beta}$, $\check{\gamma}$, and $\check{\theta}$. Owing to the separable nature of the likelihood, the score equations for the regression parameter γ are identical to those corresponding to the independent model. Therefore, $\check{\gamma} = \hat{\gamma}$ and, as $n \to \infty$,

$$\check{\gamma} \approx \mathcal{N}\left(\gamma, I_\gamma^{-1}\right),$$

where the $p \times p$ Fisher information matrix I_γ can be estimated consistently in the same way as before, i.e.,

$$I_{\check{\gamma}} = I_{\hat{\gamma}} = x \operatorname{diag}(\hat{\nu}_1, \ldots, \hat{\nu}_m)\, x^\top.$$

However, the maximum likelihood estimate for the severity parameter β is now based on a conditional likelihood, as is the estimation of the dependence parameter θ. The asymptotic behavior of conditional maximum likelihood estimators is derived, e.g., by Andersen (1970) under broad regularity assumptions. Letting $\beta_\theta = (\beta^\top, \theta)^\top \in \mathbb{R}^{p+1}$ and denoting its estimate by $\check{\beta}_\theta$, one can see that, as $n \to \infty$,

$$\check{\beta}_\theta \approx \mathcal{N}(\beta_\theta, I_{\beta\theta}^{-1}),$$

where the $(p + 1) \times (p + 1)$ Fisher information matrix $I_{\beta\theta}$ can be estimated consistently by

$$I_{\check{\beta}\check{\theta}} = \frac{1}{\phi_\theta}\, x_n \operatorname{diag}(n_1, \ldots, n_m)\, x_n^\top,$$

where x_n is the $(p + 1) \times m$ design matrix with ith column $(x_{i1}, \ldots, x_{ip}, n_i)^\top$ for $i \in \{1, \ldots, m\}$.

Remark 3.2 When the individual claims are mutually independent conditional on $N > 0$, the assumption that $\bar{Y} \mid N \sim \mathcal{G}(\mu_\theta, \phi_\theta/N)$ is equivalent to $Y_1, \ldots, Y_N$ forming a random sample from $\mathcal{G}(\mu_\theta, \phi_\theta)$. In particular, the score equations derived for the regression parameters β are the same whether the individual claim amounts Y_{ij} or the average claim amounts $\bar{Y}_i$ are used, viz.

$$\sum_{i=1}^{m} \frac{1}{\phi_\theta} \frac{x_{ik}}{\mu_{\theta i}} \sum_{j=1}^{n_i} (y_{ij} - \mu_{\theta i}) = \sum_{i=1}^{m} \frac{1}{\phi_\theta} \frac{x_{ik}}{\mu_{\theta i}} n_i (\bar{y}_i - \mu_{\theta i}) = 0,$$

where n_i, y_{ij} and $\bar{y}_i$ are the observed values of N_i, Y_{ij} and $\bar{Y}_i$, respectively.

4 Simulation study

In order to illustrate the effect of dependence in the aggregate claims model, a small simulation study was carried out using a fictive portfolio involving $m = 10{,}000$ classes. For each class $i \in \{1, \ldots, m\}$, a single predictor variable x_i was generated from a folded standard Normal distribution, and a claim count N_i was generated from a Poisson distribution, viz.

$$N_i \mid x_i \sim \mathcal{P}(\nu_i), \quad \nu_i = \exp(1 + x_i/2).$$

In 239 cases, it turned out that $N_i = 0$, and hence S_i was set equal to 0. In the remaining 9761 cases, an average claim size $\bar{Y}_i$ was generated from a Gamma distribution with a pre-specified degree of dependence θ, viz.

$$\bar{Y}_i \mid N_i, x_i \sim \mathcal{G}(\mu_{\theta i}, \phi_\theta), \quad \mu_{\theta i} = \exp(5 + 5x_i + \theta N_i), \quad \phi_\theta = 2.$$

In total, 101 equally spaced values of θ were considered in the interval $[-.5, .5]$. Both dependent and independent GLMs with log-links were then fitted to the data. The main results are summarized in Figures 1–2.

Figure 1 shows how the estimates of β_0 and β_1 are affected by the level θ of dependence. In the top row, the true parameter values are displayed in red and their estimates under the independent model (Model I) and the dependent model (Model D) are shown in green and blue, respectively. It is clear from the graphs that the estimates of β_0 and β_1 based on Model I become increasingly biased as $|\theta|$ moves away from 0. The standard error of these estimates is also larger under Model I except in the neighborhood of $\theta = 0$. At $\theta = 0$, $\hat{\beta}_0$ and $\hat{\beta}_1$ are clearly more efficient than $\check{\beta}_0$ and $\check{\beta}_1$, because θ then acts as a nuisance parameter in the (then superfluous) Model D.

Figure 2 shows the effect of dependence on the mean (left) and variance (right) of the aggregate claims. The left panel shows, as a function of θ, the average percent difference (APD) in the expected aggregate claims under Models I and D, viz.

$$\mathrm{APD}(\theta) = 100 \times \frac{1}{m} \sum_{i=1}^{m} \mathrm{APD}_i(\theta),$$

where

$$\mathrm{APD}_i(\theta) = \frac{\check{\nu}_i \check{\mu}_i \exp\{\check{\nu}_i (e^{\check{\theta}} - 1) + \check{\theta}\}}{\hat{\nu}_i \hat{\mu}_i} - 1$$

and, for all $i \in \{1, \ldots, m\}$, $\check{\nu}_i = \hat{\nu}_i$ because $\check{\gamma} = \hat{\gamma}$. This graph shows that in Model I, the expected total loss tends to be over (under) estimated in the presence of negative (positive) dependence between the claim counts and amounts. Moreover, the use of Model I generally leads to an over or under estimation of the variance of the total loss, according as $\theta < 0$ or $\theta > 0$. This is shown in the right panel of Figure 2, which displays the average percent difference in the variance of the aggregate claims under Models I and D.

5 Application

To illustrate the feasibility of the proposed approach, Schulz (2013) used dependent GLMs to analyze automobile insurance data from a large Canadian insurance company. This section provides an overview of her conclusions.

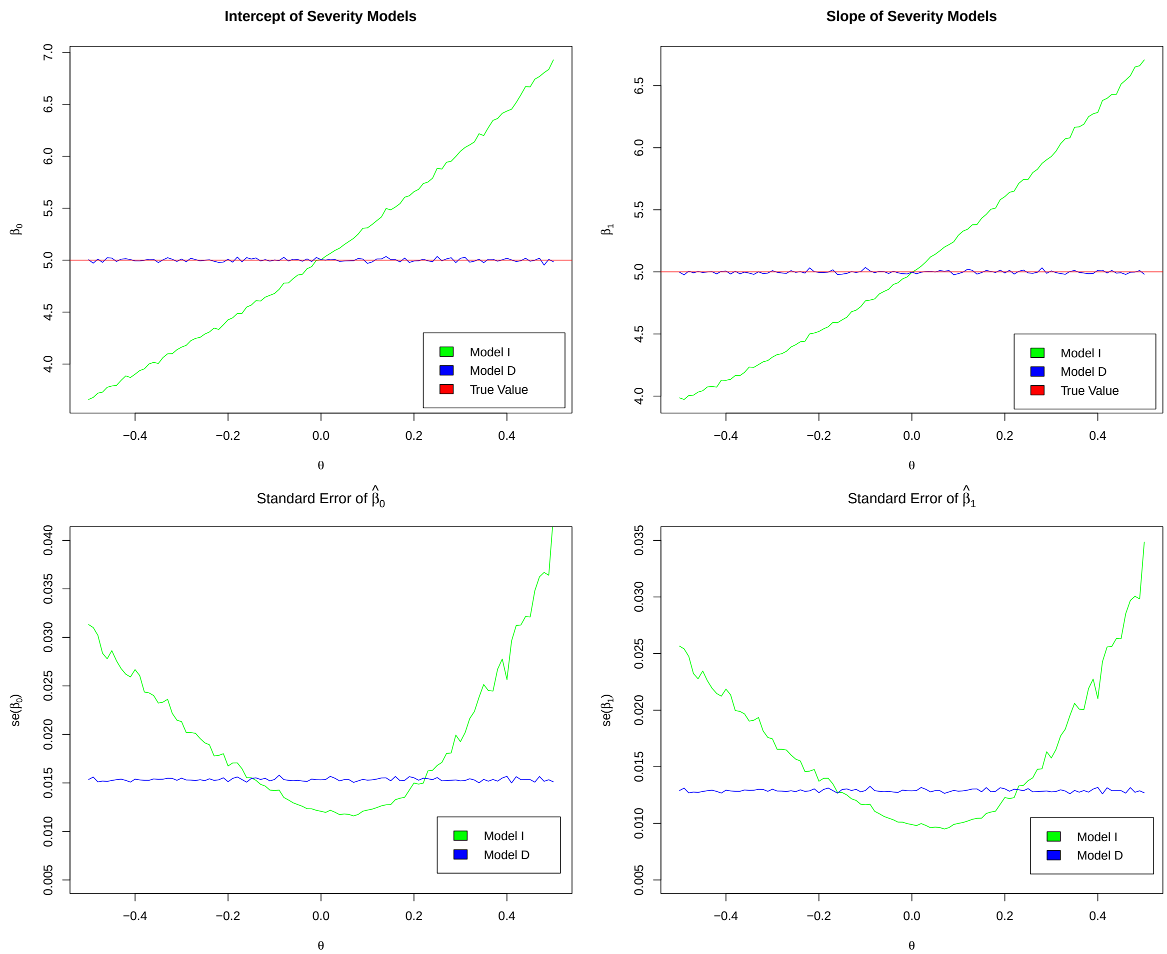

Figure 1: Effect of dependence level θ on the estimate of β_0 (upper left) and β_1 (upper right) and their standard errors (lower left and right, respectively).

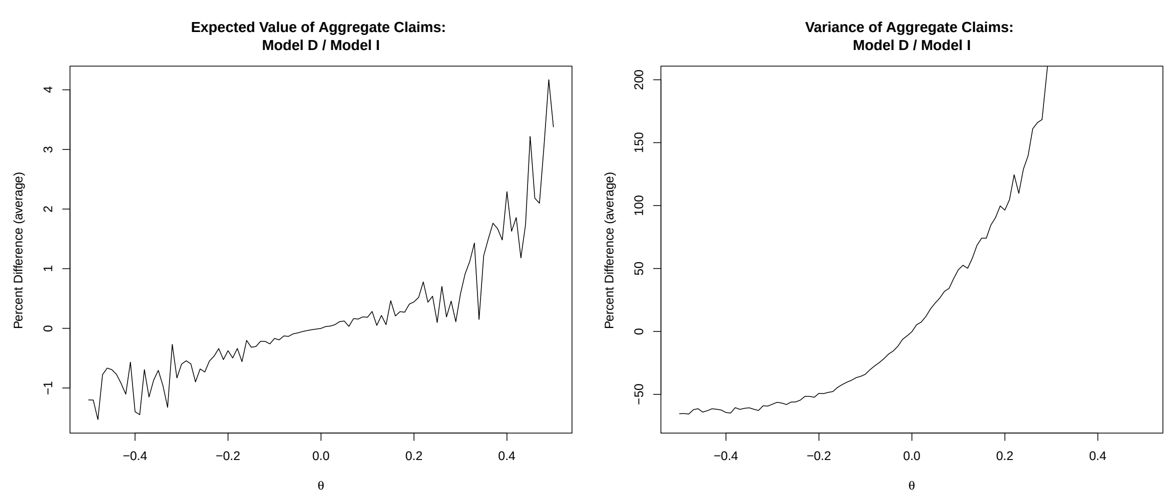

Figure 2: Average over all classes of the percent difference in the mean (left) and variance (right) of the aggregate claims as a function of dependence θ.

5.1 Data description

The analysis was carried out on the company's collision claim experience for 2003–08, inclusively. The data consist of claim counts, claim amounts and several rating variables for each policyholder having at least two weeks of exposure. Specifically, the deductible (x_1), the driver's age (x_2), gender (x_3), and marital status (x_4) were available, as well as the number of years the driver has been licensed (x_5), the number of years the policy has been with the company (x_6), the vehicle type (x_7) and the vehicle age (x_8). The model was fitted using data for the years 2003–05, and its predictive ability was assessed using the 2006–08 data.

The data for 2003–05 comprises $m = 799{,}877$ observations, each of which consists of the total claim experience for a specific combination of rating variables and a given calendar year at the individual policyholder level. These observations are treated as distinct classes in the insurance portfolio. Positive claim counts were observed for 18,895 (or 2.36%) of these classes.

Table 1 gives the number f_k of classes whose claim count is equal to $k \in \{0, 1, 2, 3\}$, along with the corresponding average claim amount, i.e.,

$$\bar{\bar{Y}}_k = \frac{1}{f_k} \sum_{i=1}^{m} \bar{Y}_i \times \mathbf{1}(N_i = k),$$

where $\mathbf{1}(A)$ denotes the indicator function of the set A. The fact that $\bar{\bar{Y}}_k$ decreases as k increases suggests negative association between frequency and severity. Note in passing that given the nature of the data, this negative association is not captured by measures of dependence such as Pearson's correlation (r_P) or Spearman's ρ: setting $\bar{Y}_i = 0$ whenever $N_i = 0$, one finds

$$r_P(N, \bar{Y}) = 0.6814, \quad \rho(N, \bar{Y}) = 0.9999$$

which rather suggest a strong positive association. This is misleading, however, given that nearly 98% of the observations are zeros. When the calculation is restricted to observations with positive claim counts, one finds

$$r_P(N, \bar{Y}) = -0.0170, \quad \rho(N, \bar{Y}) = 0.0045,$$

but this is again misleading to the extent that among policies having a positive claim count, 98% incur only one claim.

Table 1: Distribution of claim count and average amounts for 2003–05.

Claim Count	**Frequency**	**Percent**	**Average Amount (CDN \$)**
0	780,982	97.64	—
1	18,584	2.32	4,757.34
2	301	0.04	4,120.52
3	10	0.00	3,600.08
Total	799,877	100.00	4,746.59

In the following analysis, gender (x_3), marital status (x_4) and vehicle type (x_7) were treated as factor covariates. Table 2 shows the correlation between the other five variables, which were treated as continuous covariates in the models. While the driver's age (x_2) and his/her number of years licensed (x_5) are highly correlated, both predictors turned out to be significant. Multicollinearity was not deemed to be an issue.

5.2 Data modeling strategy

The dependent and independent GLMs described in Sections 2 and 3 were fitted using the `glm` function in R with log-links. A Poisson model was assumed for the marginal frequency GLM while the distribution of the severity was taken to be Gamma. The latter choice is supported by Figure 3.

Table 2: Continuous covariates correlation matrix.

	x_1	x_2	x_5	x_6	x_8
x_1	1.000	−0.086	−0.077	−0.081	−0.160
x_2	−0.086	1.000	0.842	0.338	0.071
x_5	−0.077	0.842	1.000	0.384	0.040
x_6	−0.081	0.338	0.384	1.000	0.087
x_8	−0.160	0.071	0.040	0.087	1.000

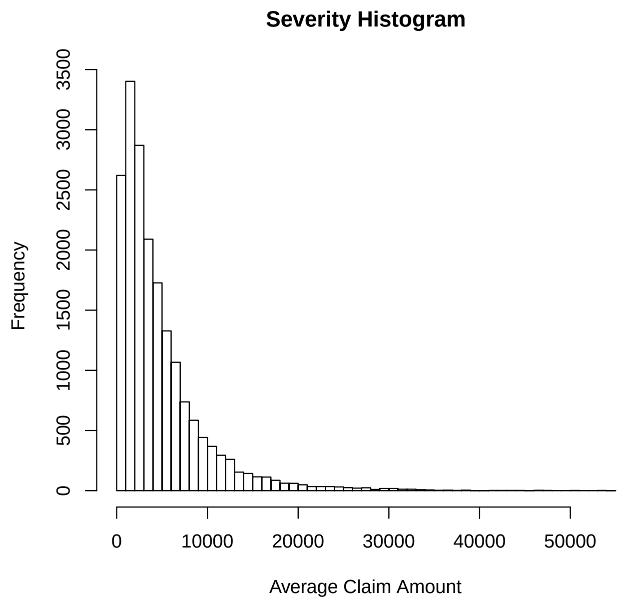

Figure 3: Average severity histogram for 2003–05.

In the frequency model, an exposure offset was used in order to take into account each policy's exposure to risk. For class $i \in \{1, \dots, m\}$ with exposure variable t_i, the GLM for N_i was defined by

$$\ln(\nu_i) = \ln(t_i) + x\gamma,$$

so that on the mean scale one has $\nu_i = t_i e^{x\gamma}$, allowing $e^{x\gamma}$ to be interpreted as a yearly expected claim count. In the severity models, the observed claim counts n_i were used as weights, given that $\bar{Y}_i \sim \mathcal{G}(\mu_i, \phi/N_i)$ when $N_i > 0$.

Note that the distribution of the aggregate losses S_i is not necessarily from the EDF. For this reason, model fit and adequacy was checked using a modified version of the standard deviance residuals. For each $i \in \{1, \dots, m\}$, the ith contribution to the deviance was defined by

$$D_i^* = 2\{\ell_{\bar{Y}}(\tilde{\beta}; \bar{y}_i) + \ell_N(\tilde{\gamma}; n_i)\} - 2\{\ell_{\bar{Y}}(\hat{\beta}; \bar{y}_i) + \ell_N(\hat{\gamma}; n_i)\}$$

in the independent model (Model I) and as

$$D_i^* = 2\{\ell_{\bar{Y}|N}(\tilde{\beta}, \tilde{\theta}; \bar{y}_i|n_i) + \ell_N(\tilde{\gamma}; n_i)\} - 2\{\ell_{\bar{Y}|N}(\check{\beta}, \check{\theta}; \bar{y}_i|n_i) + \ell_N(\hat{\gamma}; n_i)\}$$

in the dependent model (Model D), where in both cases, tildes refer to estimates under the full model. The ith deviance residual is then defined as

$$\operatorname{sign}\{s_i - \mathrm{E}_I(S_i \mid x)\}\sqrt{D_i^*} \quad \text{or} \quad \operatorname{sign}\{s_i - \mathrm{E}_D(S_i \mid x)\}\sqrt{D_i^*} \tag{4}$$

according as the predicted value of S_i is obtained using Model I or D.

5.3 Fitted models

Models I and D share the same frequency component by design. All covariates except gender (x_3) were significant at the 0.1% level. Using an analysis of deviance for nested models, the best fitting main-effect model was thus

$$\nu = t\exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_4 x_4 + \gamma_5 x_5 + \gamma_6 x_6 + \gamma_7 x_7 + \gamma_8 x_8).$$

For Poisson counts, the dispersion parameter ψ should be 1. In the sample, however, it was found that $\hat{\psi} = 0.182$ based on the estimate $D(n;\hat{\nu})/m - p_N$, where D denotes the deviance and p_N is the number of non-zero regression components of γ. This points to underdispersion in the model.

The same problem is reflected in Figure 4, which exhibits two distinct sets of deviance residuals. The residuals greater than 2 correspond to the observations with at least one claim. Accordingly, the Poisson assumption is unduly liberal for these specific data. While a zero-inflated Poisson distribution might provide a better fit, it was decided to ignore this option as this shortcoming affects both Models I and D, and the main focus of the present work is on their comparison. Moreover, the Poisson assumption is often favored out of convenience in actuarial practice.

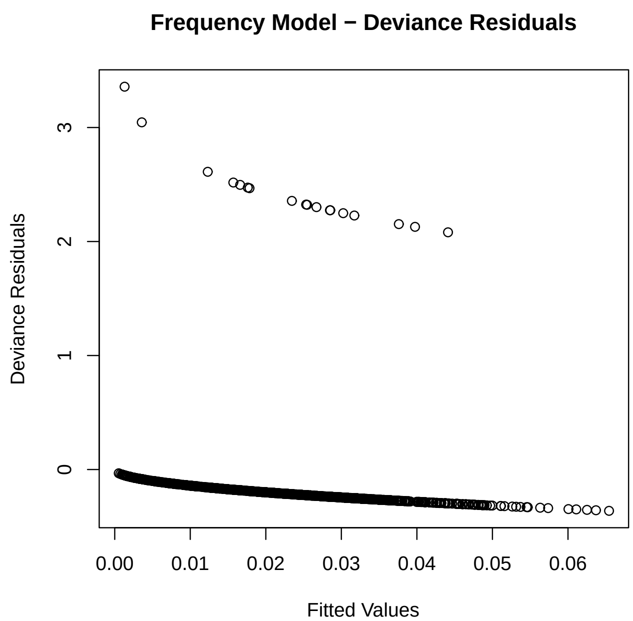

Figure 4: Frequency model deviance residuals.

The severity components of Models I and D were both found to depend on deductible (x_1), number of years licensed (x_5), number of years with the company (x_7), and vehicle age (x_8), all at the 0.1% significance level. Under the assumption of independence, the best fitting main-effect model for the severity component was thus

$$\mu = \exp(\beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8),$$

while in the dependent model, the best fit was obtained by taking

$$\mu_\theta = \mu e^{\theta N} = \exp(\beta_0 + \beta_1 x_1 + \beta_5 x_5 + \beta_7 x_7 + \beta_8 x_8 + \theta N).$$

More importantly, however, the estimate $\check{\theta} = -0.1397$ of the dependence parameter is significantly different from 0. Based on the statistic

$$\frac{\check{\theta}}{\sqrt{\mathrm{var}(\check{\theta})}} = -\frac{0.1397}{0.0388} = -4.1368,$$

which is asymptotically Normal, the hypothesis $\mathcal{H}_0 : \theta = 0$ is rejected at any reasonable level, leading to the conclusion that the assumption of independence between claim counts and amounts is flawed.

Plots of the deviance residuals for the severity component of Models I and D are given in Figure 5. They appear to be adequate. Not surprisingly, the separation between observations with and without claims observed in the frequency residuals is also apparent in Figure 6, which exhibits the deviance residuals for the aggregate loss S as defined in (4).

5.4 Model comparisons

In order to compare Models I and D, it is natural to focus on the effect of dependence on the expected total loss cost, $\mathrm{E}(S \mid x)$, respectively given by

$$\mathrm{E}_I(S \mid x) = \nu\mu \quad \text{and} \quad \mathrm{E}_D(S \mid x) = \nu\mu \exp\{\nu(e^\theta - 1) + \theta\}.$$

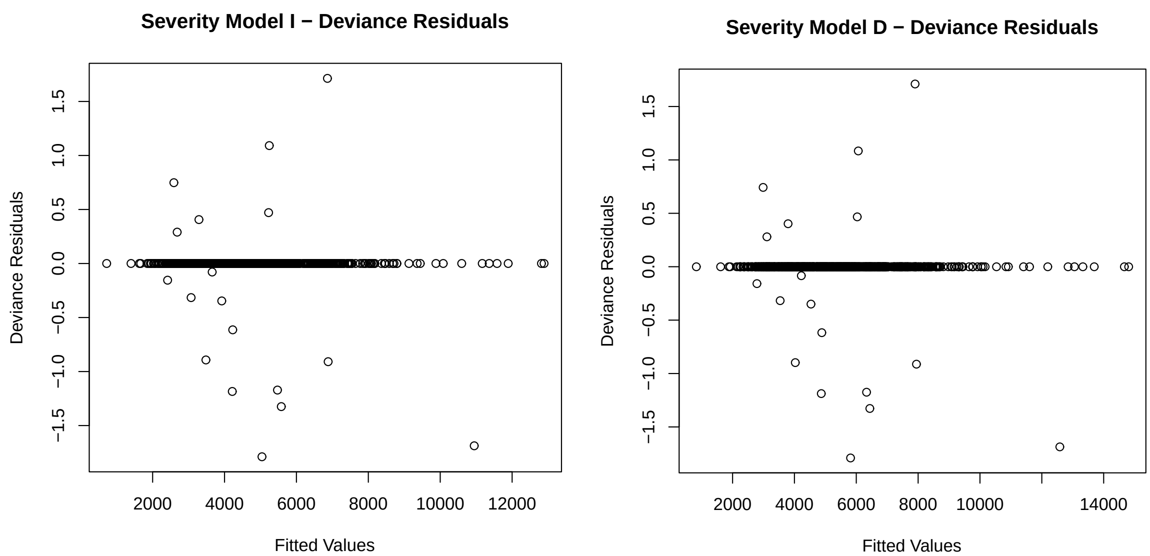

Figure 5: Severity model deviance residuals (left: Model I, right: Model D).

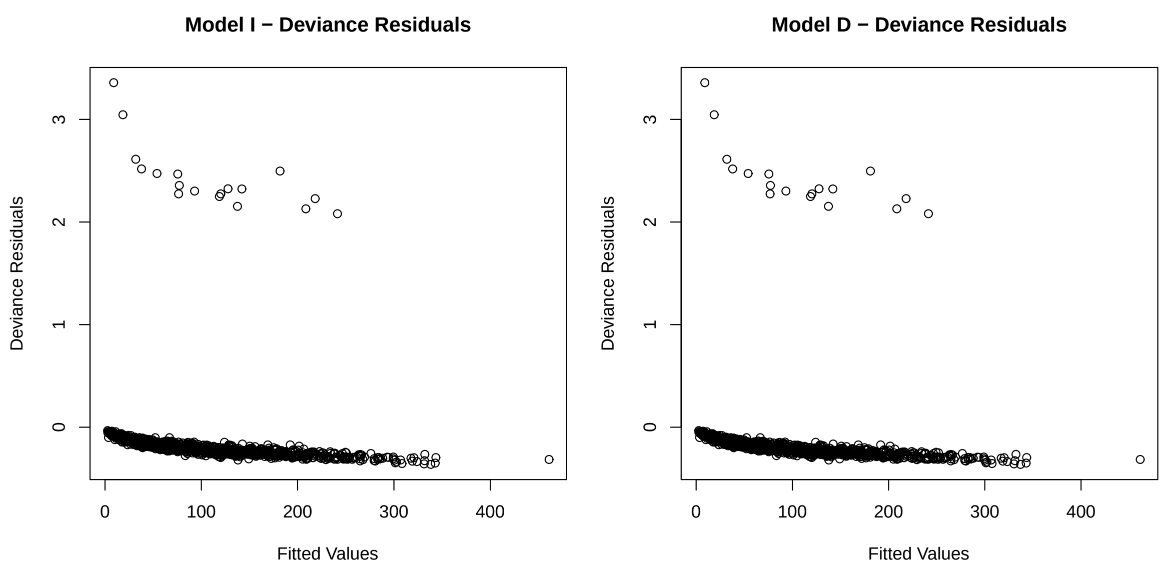

Figure 6: Aggregate loss deviance residuals (left: Model I, right: Model D).

In theory, dependence between claim counts and amounts is reflected only through the correction term $\exp\{\nu(e^{\theta}-1)+\theta\}$. In the present case, the fact that $\check{\theta} = -.1397$ would imply that this multiplicative factor will be smaller than 1, leading to a discount applicable to all policies.

However, it must be borne in mind that the estimates $\hat{\mu}$ and $\check{\mu}$ of the mean severity also differ under the two models, due to the absence/presence of the claim count N as a covariate in the severity GLM. Accordingly, an increase in any of the regression parameters in the dependent severity model could be offset by the correction term, while those coefficients that produced a decrease will be further decreased by the correction term.

For the data at hand, the impact of dependence on the expected aggregate losses turned out to be minimal. It was found that $\mathrm{APD}(\check{\theta}) = 0.1037\%$, which corresponds to a slight increase on average, although individual values of $\mathrm{APD}_i(\check{\theta})$ ranged from -1.622% to 1.361%. Interestingly, dependence had a more substantial effect on the class-level variance $\mathrm{var}(S_i \mid X_i)$, which varied from -3.417% to $+2.514\%$ and averaged -0.285% overall.

5.5 Hold-out dataset

Models I and D were used to predict claim counts and amounts as a function of the class rating variables for the policies in the hold-out dataset, consisting of 728,281 observations, 15,023 of which had positive claim counts. Predicted values for $S_i = N_i \times \bar{Y}_i$ were derived from these two models for every class i in the company's collision claim experience for the years 2006–08, inclusively. These predictions were then compared to the actual aggregate losses.

Deviance residuals computed under both models are shown in Figures 7–9. As expected, underdispersion in the Poisson model caused the same pattern to occur in the frequency and aggregate loss deviance residuals as in the modeling dataset. It was found that $\mathrm{APD}(\check{\theta}) = 0.1249\%$, while the maximum percent difference was 1.318 % and the minimum was -1.656%. In terms of the variance of the aggregate claim amount, the dependent model caused a change of -0.1886% on average; the largest positive and negative impacts were $+2.430\%$ and -3.516%, respectively.

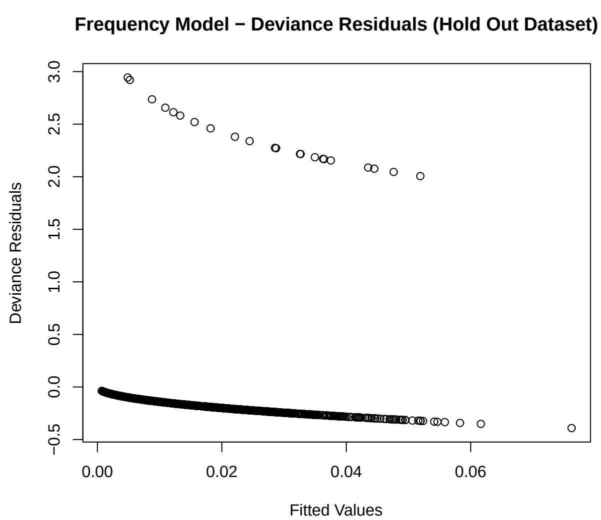

Figure 7: Frequency model deviance residuals for the hold-out dataset.

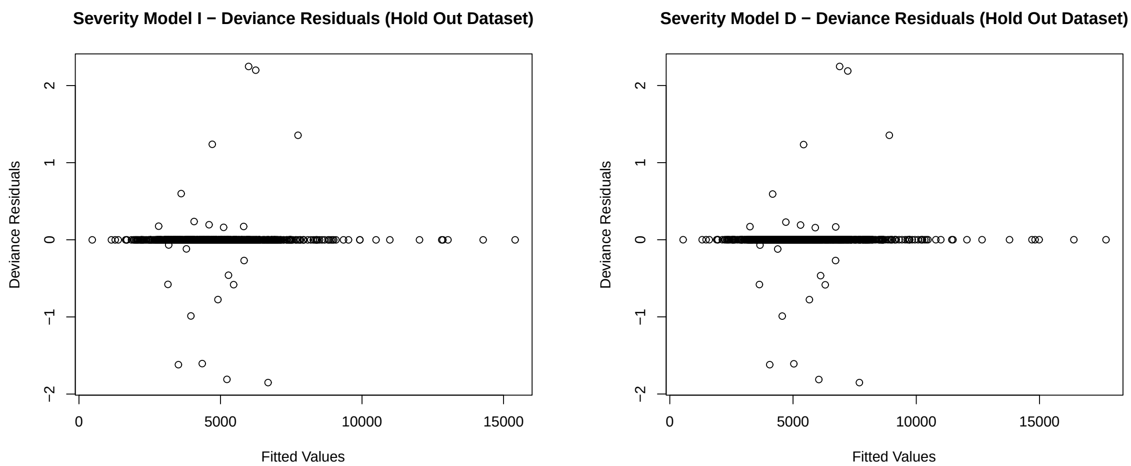

Figure 8: Severity model deviance residuals for the hold-out dataset (left: Model I, right: Model D).

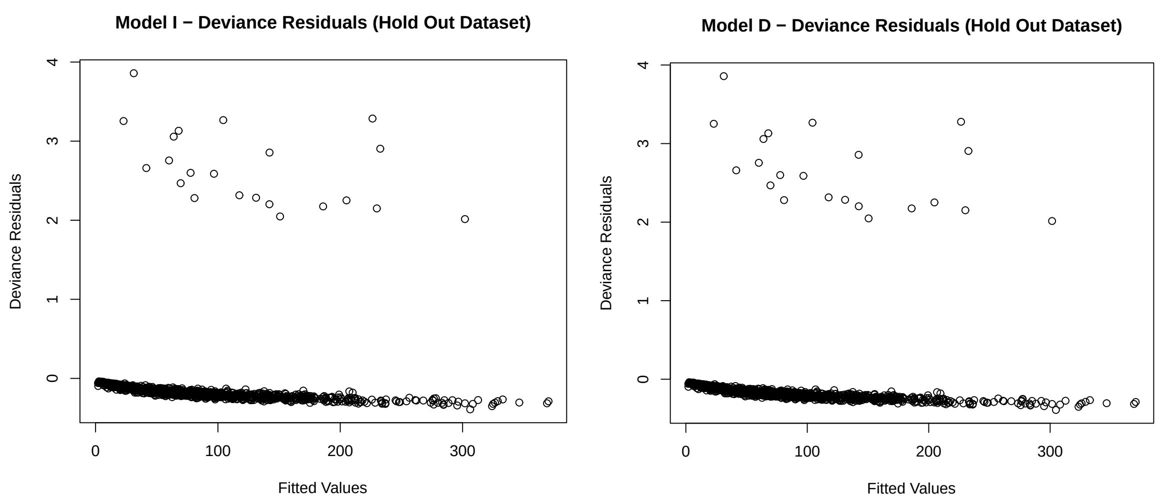

Figure 9: Aggregate loss deviance residuals for the hold-out dataset under Model I (left) and Model D (right).

6 Summary and discussion

In modeling aggregate claims, it is typically assumed that the claim counts and amounts are independent. Although this assumption allows for a simple representation of the total loss amount and justifies a separate analysis of the frequency and severity components, it is often unwarranted.

This paper describes a simple way of accounting for possible dependence between the claim counts and amounts while incorporating the effects of rating variables. This approach involves defining generalized linear models for the marginal frequency and the conditional severity components of the aggregate claims process. By including the claim count as a covariate in the conditional model for the claim amounts, a simple extension of the independent model is obtained.

In the special case of Poisson claim counts, the proposed model results in a pure premium whose form is particularly appealing: it is simply the product of a marginal mean frequency, a modified marginal mean severity, and a correction term that is easily interpretable through a dependence parameter. This formulation also allows for a straightforward comparison of the dependent and independent models, as illustrated herein through simulations.

In the data application presented here, the effect of dependence was found to be small. This is due to the weak negative association observed between claim counts and amounts for this specific automobile insurance coverage. However, ignoring dependence could have much more serious repercussions for other car insurance coverages or lines of business exhibiting a stronger (positive or negative) association between the claim frequency and severity.

While the general approach proposed here is valid irrespective of the choice of distribution for the frequency and severity components of the aggregate claims process, it is particularly convenient in the case of Poisson counts. As is well known, this assumption is often questionable in the case of property and casualty insurance data. Given that the dependent and independent models share the same frequency component, this limitation could be safely ignored for comparative purposes, as was done in Section 5. In the absolute, however, this is an issue that should be addressed and for which no entirely satisfactory solution has yet emerged.

References

Andersen, E.B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. Journal of the Royal Statistical Society, Series B, 32:283–301.

Czado, C., Kastenmeier, R., Brechmann, E.C., and Min, A. (2012). A mixed copula model for insurance claims and claim sizes. Scandinavian Actuarial Journal, 4:278–305.

Frees, E.W., Gao, J., and Rosenberg, M.A. (2012). Predicting the frequency and amount of health care expenditures. North American Actuarial Journal,15:377–392.

Frees, E.W., and Wang, P. (2006). Copula credibility for aggregate loss models. Insurance: Mathematics and Economics 38:360–373.

Gschlößl, S., and Czado, C. (2007). Spatial modeling of claim frequency and claim size in non-life insurance. Scandinavian Actuarial Journal, 3:202–225.

Krämer, N., Brechmann, E.C., Silvestrini, D., and Czado, C. (2013). Total loss estimation using copula-based regression models. Insurance: Mathematics and Economics, 53:829–839.

Quijano-Xacur, O.A., and Garrido J. (2015). Generalised linear models for aggregate claims: To Tweedie or not? European Actuarial Journal, 5:181–202.

Renshaw, A.E. (1994). Modelling the claims process in the presence of covariates. ASTIN Bulletin, 24:265–285.

Schulz, J. (2013). Generalized Linear Models for a Dependent Aggregate Claims Model. Master's thesis, Concordia University, Montréal, Canada, http://spectrum.library.concordia.ca/977691/1/Schulz_MSc_F2013.pdf.

Shi, P., Feng, X., and Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. Insurance: Mathematics and Economics, 64:417–428.