Generalized Additive Modelling of Dependent Frequency and Severity Distributions for Aggregate Claims

Tingting Chen, Anthony Francis Desmond (University of Guelph), Peter Adamic* (Ave Maria University)

* Corresponding Author

ABSTRACT

This paper examines the problem of accurately estimating the expected value and variance of aggregate claims for each policyholder. Through an appropriate statistical model to estimate the pure premium, an insurer can find niche markets to operate competitively and profitably. To this end, the framework of generalized linear models (GLMs) for aggregate claims is extended to encompass a species of frequentist generalized additive models (GAMs) based on cubic penalized regression splines. The new structure could allow for the incorporation of more flexible nonlinear and/or nonparametric trend terms for the marginal claim frequency, conditional claim severity, and finally for Tweedie modelling as well. This nonparametric approach is illustrated through simulation and applied to an automobile insurance dataset. A juxtaposition of hypothesis test results, AIC values, and attendant graphical diagnostics effectively demonstrate that the GAMs under both the independent and dependent settings give a better fit than the GLM approach.

Keywords: Premium, Generalized Additive Models, Dependence, Splines, Frequency, Severity

1 Introduction

The *Pure premium*, also known as the *expected loss cost*, consists of that part of the insurance premium necessary to pay for potential losses, and, as such, is the expected value of aggregate losses per unit of loss exposure (Werner & Modlin 2016). For insurers in the property and casualty industry, in order to operate both competitively and profitably, it is essential to accurately estimate the pure premium and then adequately price the insurance premium for each class of exposures to be insured. This protects the insurer against *adverse selection*: the insurer ends up with comparatively more bad risks and loses good, potentially more profitable risks at the same time, which can lead to unsatisfactory profits and loss in market share (Dionne et al. 2001). In the context of property and casualty insurance, the covariates used to predict the pure premium are usually called *rating variables* (Jong & Heller 2008). Generalized linear models (GLMs) have been widely used by actuaries to price and classify risks, as opined in Werner & Modlin (2016). The principal constraint of GLMs is that the transformed mean of the response (i.e., the linear predictor) is only modelled in a linear and parametric form. This is not a difficult situation for qualitative or categorical variables that could be coded by dummy variables, but can be unduly restrictive for quantitative or continuous predictors which may have strongly nonlinear and/or nonparametric effects. It has been an accepted way of modelling possibly nonlinear effects by means of parametric polynomials (Goldburd et al. 2016, Werner & Modlin 2016). However, it is now well documented that low-degree polynomials are often not flexible enough to capture the variability in the data and that increasing their degree produces unstable estimates, especially for extreme values of the rating predictors (Goldburd et al. 2016, Jong & Heller 2008, Klein et al. 2015). Due to the restrictions of GLMs and polynomial fittings, more recent research has focused on relaxing the assumptions of GLMs and has given particular attention to some nonparametric approaches to replace the use of parametric polynomials, which include the following four main aspects (Jong & Heller 2008):

(1) *Allowing for a broader class of response types*: the *quasi-likelihood* approach is to only specify how the mean and variance of the response are connected to the linear predictor;

(2) *Explanatory variable transformations*: continuous predictors are transformed with smoothers such as kernels, regression splines, loess, etc.;

(3) *Mean and dispersion models*: apart from modelling the mean of the response variable, the dispersion parameter of the response is also modelled in terms of explanatory variables;

(4) *Location, scale and shape models*: models for the mean, scale and up to two shape parameters, such as skewness and kurtosis, are specified.

In the property and casualty insurance sector, there are some relevant studies conducted in the aforementioned areas. Denuit & Lang (2004) developed Bayesian Generalized Additive Models (GAMs) based on P-splines to predict the pure premium as accurately as possible for each policyholder. Verrall (1996) applied LOESS nonparametric smoothing for claims reserve modelling. Antonio & Beirlant (2008) implemented a semiparametric Bayesian model using P-splines. Boucher et al. (2017) analyzed the effect of exposure time and distance traveled on the premium calculation simultaneously using GAMs based on cubic regression splines. Yang et al. (2018) modelled the aggregate insurance claim amounts by employing a nonparametric Tweedie compound Poisson model, and by using a gradient tree-boosting algorithm. Stasinopoulos & Rigby (2007) introduced the framework of generalized additive models for location, scale and shape (GAMLSS) to fit more complex response distributions, where not only the expectation, but multiple parameters are related to the additive predictors by the employ of a suitable link function. Klein et al. (2014) and Klein et al. (2015) proposed a general class of Bayesian GAMLSS to model the count data distribution to correct the preponderance of overdispersion and zero-inflation so commonly encountered with insurance claim frequency data.

In order to price property casualty insurance contracts more precisely, a GLM approach for aggregate claims under dependence was developed by Garrido et al. (2016) to predict the pure premium and which specifically accounted for dependence between the claim frequency and severity of the aggregate claims. They found that the pure premium under the dependent assumptions is the product of a marginal mean frequency, a modified conditional mean severity and a correction term. However, this research still left an area of improvement in that it could be readily extended to the framework of GAMs. In this paper, an alternative approach for estimating insurance pure premiums which allows for nonparametric and/or strongly nonlinear trend terms in the aggregate claims models under both the independent and dependent assumptions will be introduced by modelling the marginal claim frequency, conditional claim severity and the Tweedie pure premium distribution utilizing a GAM framework.

This paper is organized as follows. Section 2 explores the limitations of GLM and polynomial fittings, and introduces the new fields of study on alternative regression techniques which could be potentially applied to property and casualty insurance pricing. In addition, it provides a review of the *GLMs for Aggregate Claims under Dependence method* introduced by Garrido et al. (2016). Details of the new nonparametric approaches to modelling pure premiums, namely, *GAMs for Aggregate Claims under Independence* and *GAMs for Aggregate Claims under Dependence* will also be provided in Section 2, including the estimation procedures, algorithm fitting methodology, model comparison techniques and diagnostics. Section 3 presents the general GAM framework with a focus on the cubic penalized regression splines by a cardinal spline basis. Section 4 validates the modelling approaches described in the previous section, and compares the model fittings between the new nonparametric approach and previously proposed parametric methods through a simulation study. Section 5 provides an application of this new approach using a real one-year vehicle insurance dataset and compares the various model fits. Finally, in Sections 6 and 7, the paper concludes with a summary of the results obtained from this research and a discussion of possible future work.

2 GLMs for Aggregate Claims

Modelling aggregate claims in the insurance industry is commonly done within the GLM framework (Werner & Modlin 2016). The traditional and standard approach is to adopt the use of GLMs for modelling both the frequency (the claim counts) and severity (the claim size) of the claims separately (Garrido et al. 2016). Then, the estimate of the pure premium is simply the product of the mean frequency and mean severity. Alternatively, Tweedie generalized linear models (Tweedie GLMs) could be employed in the analysis to model the expected loss cost directly. Note that both these two methods inherently assume that the frequency and severity components are independent.

2.1 Independence Assumptions for Aggregate Claims

The traditional and most common approach of modelling the amount paid on all claims occurring in a fixed time period for a given policyholder i in a particular class is to record the incurred loss payments and then add them up. In this case, the aggregate losses (i.e., aggregate claims) can be represented by

$$S_i = \sum_{j=1}^{N_i} Y_{ij},\ N_i = 0, 1, 2, ..., \tag{1}$$

where N_i is the number of claims incurred, or the claim frequency, and the Y_{ij} is the individual claim amount, or the claim severity. Note that $S_i = 0$ when $N_i = 0$. It is reasonable to assume that for a given policyholder, conditional on $N_i = n_i$, the individual claim severities $Y_{ij},\ j = 1, ..., N_i$ are independent and identically distributed (i.i.d.) random variables i.e., $Y_{ij} \overset{\text{iid}}{\sim} f_{Y_i}$. In other words, the individual claim size Y_{ij} have the same probability distribution and are independent of any other claim size given N_i. Consequently, S_i belongs to a *compound distribution*: a distribution formed by summing up a random number of identical random variables Y_{ij}, of which the probability density function f_{S_i} could be expressed as,

$$f_{S_i}(y_i, n_i) = f_{Y_i|N_i}(y_i|n_i) f_{N_i}(n_i), \tag{2}$$

where $f_{Y_i|N_i}$ is the conditional probability distribution of Y_i given N_i.

It is further assumed that conditional on $N_i = n_i$, the common distribution of the random variables $Y_{i1}, Y_{i2}, ..., Y_{iN_i}$ does not depend on n_i. Moreover, the distribution of N_i does not depend in any way on the values of $Y_{i1}, Y_{i2}, ..., Y_{iN_i}$. This model is known as the **independent aggregate claims model**. When the aggregate claims (under independence) is calculated by the sum of the individual losses for one period, the *collective risk model* is obtained (Klugman et al. 2012). Since $f_{Y_i|N_i}(y_i|n_i) = f_{Y_i}(y_i)$ under this independence assumption, Equation (2) can be then simplified to: $f_{S_i}(y_i, n_i) = f_{Y_i}(y_i)f_{N_i}(n_i)$. Then, in the independence setting, it is easy to obtain

$$\begin{aligned} E[S_i] = E[E[S_i|N_i]] &= E\Big\{E\Big[\sum_{j=1}^{N_i} Y_{ij}|N_i\Big]\Big\} = E\Big\{\sum_{j=1}^{N_i} E[Y_{ij}|N_i]\Big\} \\ &= E\Big[\sum_{j=1}^{N_i} E[Y_{ij}]\Big] = E[N_i E[Y_i]] = E[N_i] \times E[Y_i], \end{aligned} \tag{3}$$

and

$$\begin{aligned} Var(S_i) = E[Var(S_i|N_i)] + Var(E[S_i|N_i]) &= E[Var(\sum_{j=1}^{N_i} Y_{ij}|N_i)] + Var(E[\sum_{j=1}^{N_i} Y_{ij}|N_i]) \\ = E[N_i Var(Y_i)] + Var(N_i E[Y_i]) &= E[N_i]Var(Y_i) + Var(N_i)\{E[Y_i]\}^2. \end{aligned} \tag{4}$$

2.2 Modelling the Claim Frequency and Severity by GLMs Separately

A GLM framework is defined by three basic components to model a response variable: a systematic component, a random or error component and a link function. A *systematic* component is given by a linear predictor $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}^\top \boldsymbol{\beta}$ where $\boldsymbol{x}$ is a $p \times 1$ vector of known covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters. The response Y should come from an exponential family distribution with mean μ, for which the probability density function (PDF) takes the general form:

$$f_Y(y; \theta, \phi) = \exp\Big[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\Big], \tag{5}$$

where θ is called the *canonical parameter* and ϕ is called the *dispersion parameter* (Faraway 2016). By specifying the functions a, b and c, various members of the family could be defined. The link function, $g(\cdot)$, describes the relationship between the mean of the response μ and the linear predictor η, i.e., $\eta = g(\mu)$. In principle, g should be a monotone continuous and differentiable function, and is a *canonical link* if $\eta = g(\mu) = \theta$, the canonical parameter of the exponential family distribution. The canonical link is often the natural choice of link and is computationally convenient (Faraway 2016). When modelling claim frequency (i.e., expected claim count per unit of exposure), the most commonly used distribution is the *Poisson* distribution. When modelling the severity of claims (i.e., expected claim size per unit of exposure), the most commonly used distribution is the *gamma* distribution.

2.3 Inference

Consider an individual policyholder i in the independent setting of Equation (3) with claim frequency N_i, average claim severity $\bar{Y}_i = \dfrac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$ and covariate vector $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^\top$. According to the general properties of mean and variance of random variables, $E[\bar{Y}_i] = E[Y_{ij}]$ and $Var(\bar{Y}_i) = Var(Y_{ij})/N_i$. Then, from Equations (3) and (4), it follows that

$$E[S_i|\boldsymbol{X}_i] = E[N_i|\boldsymbol{X}_i] \times E[Y_i|\boldsymbol{X}_i] = E[N_i|\boldsymbol{X}_i] \times E[\bar{Y}_i|\boldsymbol{X}_i], \tag{6}$$

and

$$\begin{aligned} Var(S_i|\boldsymbol{X}_i) &= E[N_i|\boldsymbol{X}_i]Var(Y_i|\boldsymbol{X}_i) + Var(N_i|\boldsymbol{X}_i)\{E[Y_i|\boldsymbol{X}_i]\}^2 \\ &= E[N_i|\boldsymbol{X}_i] \times n_i Var(\bar{Y}_i|\boldsymbol{X}_i) + Var(N_i|\boldsymbol{X}_i)\{E[\bar{Y}_i|\boldsymbol{X}_i]\}^2. \end{aligned} \tag{7}$$

Furthermore, if the claim frequency N_i is assumed to follow a Poisson distribution and the average claim severity $\bar{Y}_i$ is assumed to follow a gamma distribution, both of these two components could be modelled as GLMs and it is common to choose a log-link function for both marginal GLMs in insurance practice because it leads to a multiplicative rating structure (Jong & Heller 2008, Ohlsson & Johansson 2010). Let us denote the mean frequency given the covariate vector $\boldsymbol{X}_i$ by μ_{i1} $=E[N_i|\boldsymbol{X}_i]$, and the mean severity given the covariate vector $\boldsymbol{X}_i$ by μ_{i2}= $E[\bar{Y}_i|\boldsymbol{X}_i]$. Then,

$$g(E[N_i|\boldsymbol{X}_i]) = \ln(\mu_{i1}) = \eta_{i1} = \boldsymbol{X}_{i1}^{\top}\boldsymbol{\beta_1}, \tag{8}$$

and

$$g(E[\bar{Y}_i|\boldsymbol{X}_i]) = \ln(\mu_{i2}) = \eta_{i2} = \boldsymbol{X}_{i2}^{\top}\boldsymbol{\beta_2} \tag{9}$$

where g is a log-link function for both of the marginal GLMs, $\boldsymbol{X}_{i1}$, and $\boldsymbol{X}_{i2}$ are subsets of the covariate vector $\boldsymbol{X}_i$. $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are vectors of unknown regression parameters for the frequency and severity GLMs, respectively. Thus,

$$E[S_i|\boldsymbol{X}_i] = \mu_{i1} \times \mu_{i2} = \exp(\boldsymbol{X}_{i1}^{\top}\boldsymbol{\beta_1}) \times \exp(\boldsymbol{X}_{i2}^{\top}\boldsymbol{\beta_2}) = \exp(\boldsymbol{X}_{i1}^{\top}\boldsymbol{\beta_1} + \boldsymbol{X}_{i2}^{\top}\boldsymbol{\beta_2}). \tag{10}$$

For the gamma variable Y with mean μ and dispersion ϕ, if there are n mutually independent identical observations $y_1, ..., y_n$, then $\bar{Y} \sim gamma(\mu, \phi/n)$ (Song 2007). In this case, the claim size Y_{ij} is assumed to follow a gamma distribution with mean μ_{i2} and dispersion ϕ, thus $\bar{Y} \sim gamma(\mu_{i2}, \phi/n_i)$. Let $\phi/n_i = \phi_{1i}$, then Equation (7) becomes:

$$\begin{aligned} Var(S_i|\boldsymbol{X}_i) &= E[N_i|\boldsymbol{X}_i] \times n_i Var(\bar{Y}_i|\boldsymbol{X}_i) + Var(N_i|\boldsymbol{X}_i)\{E[\bar{Y}_i|\boldsymbol{X}_i]\}^2 \\ &= \mu_{i1} n_i \mu_{i2}^2 \phi_{1i} + \mu_{i1}\mu_{i2}^2 = \mu_{i1}\mu_{i2}^2(n_i\phi_{1i} + 1) = (\phi + 1)\exp(\boldsymbol{X}_{i1}^{\top}\boldsymbol{\beta_1} + 2\boldsymbol{X}_{i2}^{\top}\boldsymbol{\beta_2}). \end{aligned} \tag{11}$$

The parameter vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are typically estimated by using a maximum likelihood approach. This also suggests that *Iteratively reweighted least squares* (IRWLS) could be used to derive estimates, of which the full details can be found in McCullagh & Nelder (1989).

2.4 Tweedie Modelling

For a distribution in the exponential family, recall that the mean and variance can be expressed as $E[Y] = \mu = b'(\theta)$ and $Var(Y) = b''(\theta)\phi$. The variance function $V(\mu) = b''(\theta)$ describes how the variance is connected to the mean, which can be expressed by $V(\mu) = \mu^p$ for the Tweedie model, $p = 0, 1, 2, 3$ corresponding to the Gaussian, Overdispersed Poisson, gamma and the inverse Gaussian distributions, respectively. Any choice of p, except for the open interval $(0, 1)$, defines a *Tweedie distribution*. Hence, a Tweedie distribution can be denoted by $Tw(p, \mu, \phi)$ with variance function exponent p, mean μ and dispersion parameter ϕ. When values of p are in the interval $[1, 2]$, the Tweedie distribution becomes a compound Poisson-gamma distribution, which can model the expected loss costs directly if the aggregate claims are assumed to be the sum of a Poisson number of gamma distributed variables (a Compound Poisson-gamma sum). When $p \in [1, 2]$, the Tweedie distribution $Tw(p, \mu, \phi)$ corresponds to a compound Poisson-gamma distribution with the mean of the Poisson $\dfrac{\mu^{2-p}}{(2-p)\phi}$ and the gamma variables having shape parameter $\dfrac{2-p}{p-1}$ and scale parameter $\phi(p-1)\mu^{p-1}$. This alternative approach to modelling the aggregate claims also implicitly makes the assumption of independence between the frequency and severity. In addition, the mean and variance formulae of pure premiums modelled by Tweedie distributions are the same as predicted by the well-known collective risk model, as shown in Equation (6) and (7). Full details on how to model insurance claims data using Tweedie distributions were first provided by Jørgensen & Paes de Souza (1994).

Quijano et al. (2015) compared the advantages and disadvantages of the Tweedie GLMs for the pure premium and the approach of modelling the claim frequency and severity separately under the independent setting. They concluded that, on one hand, the Tweedie GLM should be preferred if it explains the data equally well as the other method due to principle of parsimony for the number of parameters used; Tweedie allows for a positive probability of claim frequency $N = 0$ and so claim severity $Y = 0$ to model excess zeroes in the insurance claims; the non-optimality of parameter estimations when modelling the mean of claim frequency and severity separately was also mentioned. However, on the other hand, the use of Tweedie is limited since its mean increases with its variance while the variance of the aggregate claims under the approach of modelling the claim frequency and severity separately is not strictly monotone as the mean differs. This problem with the Tweedie GLMs could be fixed by extending the framework of fitting a single Tweedie's compound Poisson model to a structure of double generalized linear models (i.e., also modelling the dispersion parameter), which was first introduced for insurance claims data by Smyth & Jørgensen (2002).

2.5 GLMs for Aggregate Claims Under Dependence

Although the independent aggregate claims model is simple and easy to implement, it ignores the possible underlying correlation between the frequency and severity component. In reality, claim frequency and severity are often dependent. Note that in terms of different insurance coverages and for different lines of business, the relationship between the claim frequency and severity would be different. For example, consider automobile and home insurance. In accident benefit automobile insurance, the amount of claims and the number of claims over a given period of time are usually negatively correlated because drivers who submit several claims per year are typically involved in minor accidents. On the other hand home insurance claims due to a catastrophe like flooding tend to be both large and frequent in problematic regions. Thus, the independence model approach can lead to inaccurate results and there is a need to adapt the aggregate claims model to account for potential association between claim frequency and severity.

2.6 Dependent Assumptions for Aggregate Claims

The assumptions of the independent aggregate claims model are as follows:

(1) conditional on $N_i = n_i$, the individual claim severities $Y_{ij}, j = 1, ..., N_i$ are independent and identically distributed (i.i.d.) random variables, i.e., $Y_{ij} \overset{\text{iid}}{\sim} f_{Y_i}$;

(2) conditional on $N_i = n_i$, the common distribution of the random variables $Y_{i1}, Y_{i2}, ..., Y_{iN_i}$ does not depend on n_i and, moreover, the distribution of N_i does not depend in any way on the values of $Y_1, Y_2, ..., Y_{N_i}$.

By relaxing the assumption of independence between the claim size and claim count, we will continue to assume (1) but not (2), since individual severities are now assumed to depend on the claim count N_i, which together define the dependent aggregate claims model.

There are two general multivariate modelling approaches to account for the dependence between the claim frequency and severity for a typical correlated property and casualty insurance portfolio: a *copula* approach and a *conditional* approach. The copula approach involves linking the marginal frequency and severity GLMs through a copula and constructing a bivariate Poisson-gamma GLM with dependence (Czado et al. 2012, Shi et al. 2015); while the conditional approach is to include the claim count as a covariate in the marginal average severity model, which was first introduced by Garrido et al. (2016). A comparison between these two methods, and summary of the details, can be found in Shi et al. (2015).

As usual, let $\boldsymbol{X}_i = (X_{i1}, ..., X_{ip})^\top$ denote the set of covariate values defining a given class. Without any assumption of independence between the frequency and severity, Equation (2) cannot be simplified any further since $f_{Y_i}(y_i) \neq f_{Y_i|N_i}(y_i|n_i)$, and so the first moment of the aggregate claims cannot simply be written in terms of the marginal means of the frequency and severity components, as shown below:

$$E[S_i|\boldsymbol{X}_i] = E[N_i\bar{Y}_i|\boldsymbol{X}_i] = E[E[N_i\bar{Y}_i|\boldsymbol{X}_i, N_i]|\boldsymbol{X}_i] = E[N_iE[\bar{Y}_i|\boldsymbol{X}_i, N_i]|\boldsymbol{X}_i] \neq E[N_i|\boldsymbol{X}_i]E[\bar{Y}_i|\boldsymbol{X}_i]. \tag{12}$$

Again, a Poisson frequency distribution and a gamma severity distribution are assumed and log-link functions are used in the GLMs. In the dependent model, the marginal frequency GLM is modelled the same way as in the independent model. Parameter μ_{i1} still denotes the mean frequency given the covariates $E[N_i|\boldsymbol{X}_i]$, and now let μ_{i2}^D denote the mean severity given both the covariate vector and the frequency $E[\bar{Y}_i|\boldsymbol{X}_i, N_i]$. Here $E[\bar{Y}_i|\boldsymbol{X}_i, N_i]$ is a function of both N_i and $\boldsymbol{X}_i$, which can be defined through a conditional GLM using a log-link as:

$$g(E[\bar{Y}_i|\boldsymbol{X}_i, N_i]) = \ln(\mu_{i2}^D) = \tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2 + N_i\beta_N, \tag{13}$$

where $\beta_N \in \mathbb{R}$ introduces a degree of dependence between claim counts and amounts and the regression parameters $\tilde{\beta}_2$ are different than the regression parameters in Equation (9). Similarly, the covariates $\tilde{\boldsymbol{X}}_{i2}$ are not necessarily the same as $\boldsymbol{X}_{i2}$. Equation (13) implies that the conditional mean severity is given by

$$E[\bar{Y}_i|\boldsymbol{X}_i, N_i] = \exp(\tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2 + N_i\beta_N) = \exp(\tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2)\exp(N_i\beta_N) = \tilde{\mu}_{i2}\exp(N_i\beta_N), \tag{14}$$

where we denote $\exp(\tilde{\boldsymbol{X}}_{i2}^{\top}\tilde{\boldsymbol{\beta}}_2)$ as $\tilde{\mu}_{i2}$. Also note that for a random variable Z, the *moment generating function* (MGF) is defined as $M_Z(t) = E[e^{tZ}]$. Then, the mean aggregate claims is given by (Schulz 2013):

$$\begin{aligned} E[S_i|\boldsymbol{X}_i] &= E[N_i E(\bar{Y}_i|\boldsymbol{X}_i, N_i)|\boldsymbol{X}_i] = E[N_i \tilde{\mu}_{i2}\exp(N_i\beta_N)|\boldsymbol{X}_i] \\ &= \tilde{\mu}_{i2}E[N_i\exp(N_i\beta_N)|\boldsymbol{X}_i] = \tilde{\mu}_{i2}E\left[\frac{\partial}{\partial\beta_N}\exp(N_i\beta_N)|\boldsymbol{X}_i\right] \\ &= \tilde{\mu}_{i2}\frac{\partial}{\partial\beta_N}E[\exp(N_i\beta_N)|\boldsymbol{X}_i] = \tilde{\mu}_{i2}\frac{\partial}{\partial\beta_N}M_{N_i|\boldsymbol{X}_i}(\beta_N), \end{aligned} \tag{15}$$

where M is the MGF of N_i. For a Poisson response $N_i \sim Pois(\mu_{i1})$, $M_N(t)$ is given by $M(t;\mu_{i1}) = \exp\{\mu_{i1}\{\exp(t)-1\}\}, t \in \mathbb{R}$. So in this case, $M_{N_i|X_i}(\beta_N) = \exp\{\mu_{i1}(e^{\beta_N}-1)\}$, for $\beta_N \in \mathbb{R}$ (Schulz 2013). Then, it follows that

$$\begin{aligned} E[S_i|\boldsymbol{X}_i] &= \tilde{\mu}_{i2}\frac{\partial}{\partial\beta_N}M_{N_i|X_i}(\beta_N) = \tilde{\mu}_{i2}\frac{\partial}{\partial\beta_N}\exp\{\mu_{i1}(e^{\beta_N}-1)\} \\ &= \tilde{\mu}_{i2}\exp\{\mu_{i1}(e^{\beta_N}-1)\}\mu_{i1}\exp(\beta_N) = \tilde{\mu}_{i2}\mu_{i1}\exp\{\mu_{i1}(e^{\beta_N}-1)+\beta_N\}. \end{aligned} \tag{16}$$

Now the pure premium can be written as the product of three terms: the marginal mean frequency, a modified marginal mean severity, and a dependence correction term, i.e., $\exp\{\mu_{i1}(e^{\beta_N}-1)+\beta_N\}$. Table 1 shows the $E[S_i|\boldsymbol{X}_i]$ results for different N_i distributions (Garrido et al. 2016).

Table 1: $E[S_i|\boldsymbol{X}_i]$ for Other Distributions of N_i by a Log-link in Severity GLM

Frequency Distribution	$E[N_i\|\boldsymbol{X}_i]$	$E[S_i\|\boldsymbol{X}_i]$
Binomial(m,p)	mp	$\tilde{\mu}_{i2}mpe^{\beta_N}\{(1-p)+pe^{\beta_N}\}^{m-1}$
Negative Binomial(r,p)	$r(1-p)/p$	$\tilde{\mu}_{i2}\{r(1-p)/p\}e^{\beta_N}[p/(1-(1-p)e^{\beta_N})]^{r+1}$
Zero Inflated Poisson (π,λ)	$(1-\pi)\lambda$	$\tilde{\mu}_{i2}(1-\pi)\lambda\exp\{\lambda(e^{\beta_N}-1)+\beta_N\}$

If the individual claim size, conditional on the number of claims $Y_{ij}|N_i$, has a gamma distribution with mean μ and dispersion parameter ϕ, then the average claim amount $\bar{Y}|N_i$ also follows a gamma distribution with mean μ and dispersion ϕ/N_i (Song 2007). So, in this dependent setting, $\bar{Y}_i|N_i$ follows a gamma distribution with mean μ_{i2}^D and dispersion $\phi_{2i} = \phi'/n_i$, where $Y_{ij}|N_i \sim (\mu_{i2}^D, \phi')$. Then the variance of the aggregate claims under the auspices of dependence is as follows:

$$\begin{aligned} Var(S_i|\boldsymbol{X}_i) &= Var(E[S_i|\boldsymbol{X}_i, N_i]|\boldsymbol{X}_i) + E[Var(S_i|\boldsymbol{X}_i, N_i)|\boldsymbol{X}_i] \\ &= \mu_{i1}(\tilde{\mu}_{i2})^2\Big\{\mu_{i1}\exp\{\mu_{i1}(e^{2\beta_N}-1)+4\beta_N\} + \\ &(\phi'+1)\exp\{\mu_{i1}(e^{2\beta_N}-1)+2\beta_N\} - \mu_{i1}\exp\{2\mu_{i1}(e^{\beta_N}-1)+2\beta_N\}\Big\}, \end{aligned} \tag{17}$$

which was originally derived by Schulz (2013). Please see Appendix A for the full derivation. Therefore, when a log-link is used in conditional claim severity models, the closed form formulae for both the mean and variance of the dependent aggregate losses can be determined.

3 GAMs for Aggregate Claims

Just as GLMs follow from linear models, additive models could be readily extended to GAMs. A GAM is an additive extension of the family of GLMs (Hastie & Tibshirani 1990). GAMs extends traditional GLMs by allowing the linear predictor to depend linearly on unknown smooth functions. In GAMs, the linear predictor would be replaced by an additive sum of parametric fits and some smooth functions to predict the expected value of the response, and the response follows an exponential family distribution or just has a known mean-variance relationship.

3.1 The Basic GAM Model

Assume that the observations $y_i's$ are i.i.d. with some exponential family distribution, $i = 1, ..., n$. When a GAM relates an individual response y_i to some predictors $\mathbf{x}_i$, the mathematical expression is given by

$$g\{E(y_i)\} = g(\mu_i) = \eta(\boldsymbol{x}_i) = \boldsymbol{x}_i^{*\top}\boldsymbol{\beta}^* + \sum_{j=1}^{p} f_j(x_{ij}), \tag{18}$$

where g is a known link function, μ_i is the mean of the response y_i, and $\eta(\boldsymbol{x}_i)$ is the linear predictor in which there is an additive model. The $\boldsymbol{x}_i^{*\top}\boldsymbol{\beta}^*$ term is parametric and the $f_j's$ are traditionally modelled using nonparametric terms (i.e., smooth functions) that are estimated from data. The $x_{ij}'s$ are continuous covariates which have effects on the mean of y_i and $(\boldsymbol{x}_i^{*\top}, x_{i1}, ..., x_{ij})^\top \in \boldsymbol{x}_i$. Hence, the advantages of GAMs are quite obvious. It makes the linear predictor $\eta = X\beta$ more flexible since it can not only incorporate factors and linear terms, but can also include the smoothers f, which can vary, e.g., loess, kernels, penalized splines etc. So the range of potential fits to the data is much larger than GLMs. In addition, it will automatically fit a nonparametric f_j for each predictor x_j to the data points locally without manually trying out different transformations, which can potentially lead to more precise predictions when facing complicated nonlinear and/or nonparametric trends.

As discussed earlier, the closed form formulae of mean and variance of the dependent aggregate losses can be obtained, as long as a log-link is applied in the conditional claim severity model. So for the GAM fittings, the same condition and formulae could be applied as in the GLM case. However, the claim counts N_i covariate can not be smoothed in the marginal dependent severity GAM, otherwise the closed forms of $E[S_i|\mathbf{X}_i]$ and $Var(S_i|\mathbf{X}_i)$ are no longer accessible (see Equations (14), (15), (16) and (45).)

For the `gam` function from the `mgcv` library, there are two arguments namely, `Tweedie` and `tw`, designed for fitting the Tweedie families and restricted to variance function powers between 1 and 2. `tw` is for use when the variance function exponent parameter p is to be estimated by likelihood methods during fitting. `Tweedie` is for use with fixed p. Notice that Tweedie distributions with automatic estimation of the extra parameter p are only available with REML or ML smoothing parameter estimation. And these estimation methods tend to work better than prediction error criteria with selection penalties such as GCV, due to a reduced tendency to undersmooth (Wood 2017). For the Tweedie modelings, the expected aggregate losses and the standard error of the prediction can be directly obtained from the `predict` function in R, which makes it easier to get the prediction interval of the estimated mean aggregate claims. As discussed earlier, the Tweedie GAM also inherently assumes that the frequency and severity components are independent. For a Tweedie GAM, the expected aggregate claims $E[\bar{S}_i|\mathbf{X}_i]$ can be defined through a GAM using a log-link function as:

$$g(E[S_i|\boldsymbol{X}_i]) = \ln(\mu_{is}) = \boldsymbol{X}_{is}^{*\top}\boldsymbol{\beta}_s^* + \sum_{j=1}^{q} f_{js}(x_{ij}), \tag{19}$$

where $\boldsymbol{X}_s^*$ is the design matrix for parametric factor and quantitative predictors, and f_{js} represent nonparametric smooth functions. Also, $(\boldsymbol{X}_{is}^{*\top}, X_{i1}, ..., X_{iq})^\top \in \boldsymbol{X}_i$ is the same set of predictors as considered in the alternative separately modelling of the claim frequency and severity.

3.2 Controlling Smoothness by Penalizing Wiggliness

Consider a response variable Y with n observations whose mean $E(Y|X = x)$ is fitted by a cubic spline. Note that except for using a truncated power series basis, there are many alternative equivalent ways to represent a cubic spline by using different basis expansions. So the estimate of each y fitted by a smoother f with a series of basis functions $B(x)$ could be written as:

$$\hat{y} = \hat{f}(x) = \sum_{j=1}^{q} \hat{\beta}_j B_j(x), \tag{20}$$

where $B_j(x)$ could be, for example, the cardinal spline basis or the B-spline basis. Let us denote the response $\boldsymbol{y} = (y_1, ..., y_n)^\top$, the parameter vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^\top$ as usual and $\boldsymbol{B}$ is the $q \times n$ design matrix for which each column contains all the basis functions $(B_{1i}(x), ..., B_{qi}(x))^\top$ for the corresponding observation $i = 1, ..., n$ (also assuming that these basis functions have at least two integrable derivatives). The basic idea behind the roughness penalty approach is to make a necessary compromise between the smoothness and wiggliness of a model fit, i.e., controlling the model's smoothness by adding a roughness penalty to the least squares fitting. Rather than fitting the model by minimizing $||\boldsymbol{y} - \boldsymbol{B}^\top \boldsymbol{\beta}||^2$, for example, it could be fitted by minimizing

$$||\boldsymbol{y} - \boldsymbol{B}^\top \boldsymbol{\beta}||^2 + \lambda \sum_{j=2}^{k-1} \{f(\xi_{j-1}) - 2f(\xi_j) + f(\xi_{j+1})\}^2, \tag{21}$$

where k is the number of knots with ordered interior knots denoted by $\xi_1 < \xi_2 < ... < \xi_k$. Here the summation term is a sum of squared second differences of the function at the knots, which measures the wiggliness and which coarsely approximates $\int [f^{''}(x)]^2 dx$, i.e., the integrated squared second derivative penalty (Eilers & Marx 1996). $\int [f^{''}(x)]^2 dx$ is regarded as a *roughness penalty.* When f is a straight line, the penalty is zero. The penalty term will be low if f is a smooth curve. By contrast, if the model fit f is very wiggly, then the penalty will be high.

There are a number of options available for automatic smoothing parameter estimation. The most popular method of automatically selecting the λ is the *generalized cross-validation* (GCV) (Ruppert et al. 2003):

$$\begin{aligned} GCV(\lambda) &= \sum_{i=1}^{n} \Big(\frac{\{(I - \boldsymbol{S}_\lambda)y\}_i}{1 - n^{-1} tr(\boldsymbol{S}_\lambda)} \Big)^2 \\ &= \frac{RSS(\lambda)}{1 - n^{-1} tr(\boldsymbol{S}_\lambda)}^2, \;\; 0 \le \lambda \le \infty. \end{aligned} \tag{22}$$

where RSS is the *residual sum of squares* and $\boldsymbol{S}_\lambda$ is the smoother matrix associated with $\hat{f}$. Also, $\hat{f}(x_i; \lambda) = \sum_{j=1}^{n} S_{\lambda,ij} y_j$ ($S_{\lambda,ij}$ is the (i, j) entry of $\boldsymbol{S}_\lambda$) and $tr(\boldsymbol{S}_\lambda)$ is the trace of the matrix $\boldsymbol{S}_\lambda$, which will be further discussed in section 2.3.4. GCV is actually an approximation to the *cross-validation* (CV) criterion, which is defined by

$$CV(\lambda) = \sum_{i=1}^{n} \{y_i - \hat{f}_{-i}(x_i; \lambda)\}^2, \;\; 0 \le \lambda \le \infty \tag{23}$$

where $\hat{f}_{-i}$ indicates that point (x_i, y_i) is left out of the fit. The λ that minimizes this criterion over $\lambda \ge 0$ would be picked, which is against the wiggly fits that $\hat{f}$ gives. However, the CV is not computationally efficient so a simplified approximation of it, namely GCV, has been widely used instead (Faraway 2016).

4 Simulation Study

4.1 Preliminaries

In order to validate the new modelling approach of GAMs for aggregate claims, a simulation study was carried out using an imitative portfolio including $m = 10000$ cases per iteration. The effect of dependence between the claim frequency and severity, as well as a nonparametric and/or strongly nonlinear predictor in the aggregate claims model were quantified. The GAM and GLM model fits under both the independent and dependent assumptions were also compared. In this simulation study, six aggregate claims models, namely, independent GLMs (frequency GLM + independent severity GLM), Tweedie GLM, dependent GLMs (frequency GLM + dependent severity GLM), independent GAMs (frequency GAM + independent severity GAM), Tweedie GAM and dependent GAMs (frequency GAM + dependent severity GAM), were considered. All of these models were fitted to the same set of 10000-case simulated claim frequency and severity, with 1000 iterations, in order to be compared consistently.

Simulating the Claim Frequency Dataset

For each individual $i \in \{1, ..., m\}$ where $m = 10000$, a claim count N_{fi} was simulated from a Poisson distribution:

$$N_{fi} | X_{i1}, X_{i2} \sim Poisson(\mu_{fi}), \tag{24}$$

where $\mu_{fi} = e^{\eta_{fi}}$ and the linear predictor is given by

$$\eta_{fi} = 1 + 0.5x_{i1} + sin(x_{i2}). \tag{25}$$

Denote the parametric coefficients of the intercept and X_1 predictor, $\beta_{f0} = 1$ and $\beta_{f1} = 0.5$, respectively. A single predictor variable X_{i1} was generated from a folded standard normal (the half standard normal) distribution using *acceptance sampling* (Schilling & Neubauer 2009). The function $sin(x_{i2})$ with the domain [0,4] for another predictor X_{i2} was chosen arbitrarily since the purpose of this simulation study is just to test whether the new GAM approach performs well and compare its performance to different approaches. So this test function mimics a 'nonparametric' function form and introduces a nonlinear term to the linear predictor η_{fi}. A total of 10000 values of X_{i2} were generated from a [0,4] continuous uniform distribution and a small amount of noise (0.01) was added to each $sin(x_{i2})$ by the R `jitter` function. As usual, assume the claim counts N_{fi} follows a Poisson distribution and a log-link is used to fit both the GLM and GAM. Then, the mean of the claim counts μ_{fi} can be represented by $\mu_{fi} = \exp(\eta_{fi})$. The individual claim count n_{fi} was generated from a Poisson distribution by using the `rpois` function in R for each corresponding μ_{fi}.

Simulating the Claim Severity Dataset

By generating $m = 10000$ claim frequency n_{fi}, given a simulated positive claim count n_{fi}, an average claim severity $\bar{Y}_i = (Y_{i1} + ... + Y_{in_{fi}})/n_{fi}$ was generated by simulating n_{fi} individual claim sizes from a gamma distribution with a pre-specified degree of dependence β_N and a known nonlinear function. As such, the individual claim amount is expressed as:

$$Y_{ij}|N_{fi}, X_{i1}, X_{i2} \sim gamma(\mu_{yi}, \phi), \tag{26}$$

where $j = 1, ..., n_{fi}$, $\phi = 0.1$ and $\mu_{yi} = e^{\eta_{yi}}$ by a log-link function. The predictor was set to

$$\eta_{yi} = 1 + 2x_{i1} + cos(x_{i2}) + sin(x_{i2}) + \beta_N n_{fi}. \tag{27}$$

Denote the parametric coefficients of the intercept and X_1 predictor, $\beta_{y0} = 1$ and $\beta_{y1} = 2$. One hundred equally spaced β_N in the interval of $[-0.2, 0.2]$ were examined. Notice that if this interval was wider, the fitting algorithms for the independent severity GLMs and GAMs in R tended not to converge.

When using a penalized regression spline, the smoothing parameter selection criterion does all of the work of model selection, by associating penalities with the otherwise unpenalized model terms and estimating the associated smoothing parameters. Neither the accurate choice of the number of knots, k, nor the precise selection of knot locations, plays a vital role in determining the model fit. Given the large sample size of $m = 10000$, it is reasonable to pick twenty evenly-spaced knots for the cubic penalized regression spline of X_2 for both the frequency GAM and severity GAMs. The maximum basis dimension k was selected to be 20 accordingly.

4.2 Comparing the GLM and GAM Frequency Simulations

Table 2 gives the average values of the estimates of β_{f0} and β_{f1} and their standard errors from 1000 iterations. The $\hat{\beta}_{f0}$ of the frequency GAM is clearly closer to the true value and more efficient than that of the frequency GLM since it has a smaller average standard error. The estimates of β_{f0} for the GAM fitting maintained around 1.414. This value deviates 0.414 unit from the true value of β_{f0} and compensates for the disparity between the nonlinear estimate of the smoothed X_2 and true $\texttt{sin}(x_2)$ function, which will be further discussed below. The frequency GLM and GAM have almost identical average $\hat{\beta}_{f1}$ and standard errors of the estimate, which suggests the two models perform equally well in predicting the parametric coefficient of X_1. In addition, after over 1000 resamplings and refittings, there is no adjusted R^2 of the frequency GAM larger, or AIC value of it smaller, than the corresponding frequency GLM. Table 3 provides the mean of the AIC and adjusted R^2. The smaller average AIC and adjusted R^2 confirm that the GAM fitting is superior to the GLM.

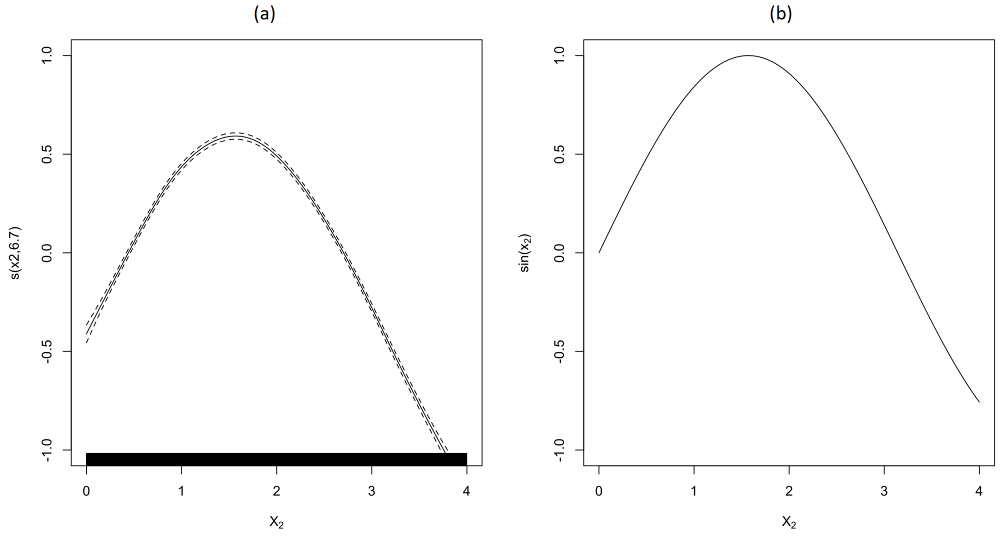
Table 2: Average Estimates of β_{f0} and β_{f1} and their Standard Errors from 1000 Iterations

	Mean of $\hat{\beta}_{f0}$	Mean of se($\hat{\beta}_{f0}$)	Mean of $\hat{\beta}_{f1}$	Mean of se($\hat{\beta}_{f1}$)
Frequency GLM	1.9002	0.0087	0.4998	0.0053
Frequency GAM	1.4141	0.0069	0.4998	0.0053
Note: $\beta_{f0} = 1$ and $\beta_{f1} = 0.5$.				

Table 3: Mean of Adjusted R^2 and AIC from 1000 Iterations

	Mean of Adjusted R^2	Mean of AIC
Frequency GLM	35.5 %	56800.2
Frequency GAM	70.9 %	46106.1

Figure 1: Comparison of (a) the Frequency GAM Fitting of X_2 at the 95% Confidence Band and (b) the True $\sin(X_2)$ Function



For a random selected claim frequency dataset with $m = 10000$ cases, the left panel of Figure 1 shows the estimated effect of X_2 in the frequency GAM as a solid curve and the 95% Bayesian credible interval of the fitting as dashed lines. The number 6.7 in the vertical axis label represents the effective degrees of freedom of the smoother, which relays the shape of the fit. One degree of freedom corresponds to a linear shape. The higher the degrees of freedom, the more curved our fitting is. The total effective degrees of freedom of this GAM fitting is the sum of 6.7 plus 1 degree of freedom for each of the model intercept and X_1, thereby giving df_{fit} =8.7. The right panel of Figure 1 depicts the shape of the sine function defined in Equation (25). As shown in the plots, the relationship between X_2 and its transformed function predicted by the frequency GAM has a similar shape as the pre-defined true sine function. However, there is an approximate 0.4 unit vertical shift in the GAM fitting of X_2 from the true `sin(x)` function. This discrepancy is possibly due to the choice of the spline basis functions, which would be compensated by adding the same value to the estimated intercept.

Table 4 indicates that the estimated β_{f0} of the frequency GAM is 1.4168, which is 0.4168 units higher than the true value of $\beta_{f0} = 1$, which compensates for the vertical shift of the fitting of X_2 as expected. A

Table 4: Regression Parameter Estimates - Frequency GLM and GAM

Model	Regression Parameter	Estimate	Standard Error	t-value	p-value
GLM	Intercept	1.9153	0.0086	221.67	$< 2 \cdot 10^{-16}$
	X_1	0.4960	0.0054	91.50	$< 2 \cdot 10^{-16}$
	X_2	-0.2004	0.0033	-61.03	$< 2 \cdot 10^{-16}$
GAM	Intercept	1.4168	0.0068	207.50	$< 2 \cdot 10^{-16}$
	X_1	0.5048	0.0054	93.24	$< 2 \cdot 10^{-16}$

Note: $\beta_{f0} = 1$ and $\beta_{f1} = 0.5$.

Figure 2: GAM Fitting of X_2 and 0.416825 Unit Vertical Down-Shift of the True $\sin(x_2)$ function

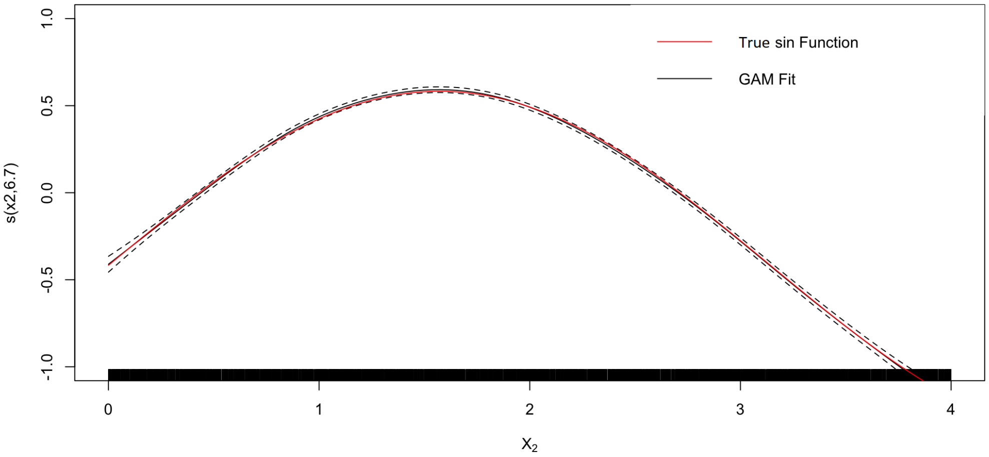
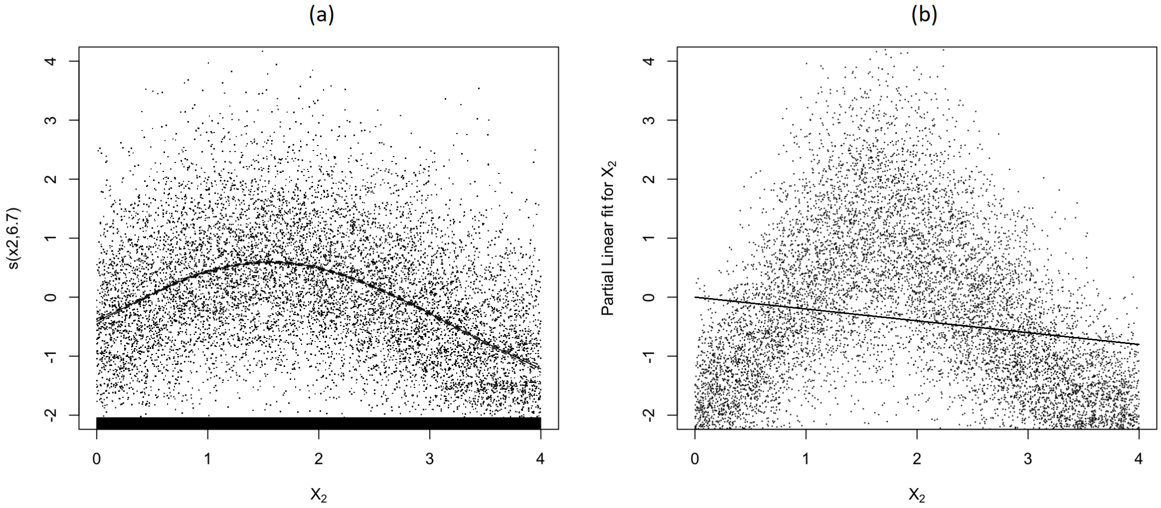


Figure 3: Partial Residual Plots of X_2 for (a) the Frequency GAM and (b) the GLM Fit



comparison of 0.4168 unit vertical down-shift of the true `sin` function and the GAM fitting of X_2 shown in Figure 2 verifies the conjecture we made before. $\hat{\beta}_{f0}$ of the frequency GLM is given by 1.9153, which deviates from the true value more severely than the frequency GAM. However, both the β_{f1} estimates for the frequency GLM and GAM are close to the true value of $\beta_{f1} = 0.5$. The parameter estimate of X_2 in the frequency GLM is given by -0.2004. Figure 3 shows the respective partial residuals plots of X_2 for the GAM and GLM fittings. By using the penalized spline smoothing in GAM, there seems to be a good transformation of the predictor X_2 as the fitted curve correctly captured the trend of X_2 and the partial residuals are evenly distributed around the fit, while the GLM fitting of predictor X_2 appears to be inadequate since the pattern is not well described by the fitted line. To examine whether there is a nonlinear trend for X_2, i.e., whether the smoothing of X_2 is necessary, the p-value can be computed from approximate F-statistics by the R `anova` function. The extreme value of the tail probability against $H_0 : X_2$ *is linear* (p-value $= 2.2 \times 10^{-16}$) indicates there is a need to smooth X_2. The evidence from a parametric bootstrap p-value$=0$, after repeating 2000 times, supports the conclusion we found here. With 1000 iterations, the mean of the bootstrapping p-values is equal to 0. This result also indicates the frequency GAMs provide better fits for the simulated dataset than the frequency GLMs.

Therefore, we can conclude that the frequency GAM significantly improved the frequency model fit compared with the parametric approach under this simulation scenario.

4.3 Comparing Conditional Severity Models

With the simulated claim frequency dataset held fixed, the severity dataset was regenerated 1000 times and the models were correspondingly refitted 1000 times. The mean of the estimations was obtained to represent the estimate of the parametric coefficients under each β_N. Figure 4 displays how the estimates of β_{y0} and β_{y1} are affected by the dependence level β_N. Notice that if the claim frequency dataset was also regenerated each time, the curvatures would become more wiggly and the fitting algorithm of the independent severity GLM in R would tend to not converge as $|\beta_N| \to 0.2$. Nevertheless, it is wise to hold the claim frequency dataset fixed since the aim here is to compare the dependent effect β_N built into the claim severity models.

As the absolute value of β_N departs away from 0, the estimates of β_{y0} and β_{y1} based on the independent severity GLM and GAM become increasingly biased as expected, due to the missing claim count N_f predictor in the independent severity models. For both of the dependent severity models, the coefficients $\hat{\beta}_{y0}$ and $\hat{\beta}_{y1}$ were constant. $\hat{\beta}_{y1}$ of the dependent severity GAM is centered around the true value 2 while $\hat{\beta}_{y1}$ of the dependent severity GLM is centered around 1.8, which presents an approximate 0.2 unit of consistent bias. Both of the $\hat{\beta}_{y0}$ for the dependent severity models show biases. However, $\hat{\beta}_{y0}$ for the dependent severity GAM is about 1.3 and much closer to the true value 1 than the GLM case with around the value of 1.7. As discussed in the frequency part above, the choice of the spline basis functions would affect the estimated intercept value, which will be compensated by the same amount of value deducted from the smoothing of X_2 (i.e., a vertical down-shift from the true function).

Figure 5 shows how the standard errors of $\hat{\beta}_{y0}$ and $\hat{\beta}_{y1}$ change with the value of β_N for the four marginal severity models. As $|\beta_N|$ moves away from 0, the standard errors of $\hat{\beta}_{y0}$ and $\hat{\beta}_{y1}$ for the independent severity models are increasingly larger than the dependent severity models. However, in the neighborhood of $\beta_N = 0$, the independent severity models have smaller standard errors than their dependent counterparts. These patterns are also identified by Garrido et al. (2016). In addition, the estimates of β_{y0} and β_{y1} under the GAM structure clearly have smaller standard errors than those under the parametric GLM structure for both of the independent and dependent cases.

Figure 6 illustrates how the AIC for each of the four severity models differs as β_N changes. As β_N increases, the mean AIC value shows an overall upward trend for all of the four models. However, the dependent severity GAM invariably has the lowest AIC. Also, the AIC criterion is always in favor of the GAM structure over the parametric GLM approach no matter the independent or dependent case. Notice that when β_N ranges from -0.2 to -0.1, the AIC values for the independent GLM and independent GAM are very similar to each other; otherwise, the two independent severity models have quite different AIC values. This is probably because they have similar model fittings for X_2 when $\beta_N \in [-0.2, -0.1]$, which is demonstrated by Figures 4.10, 4.11 and 4.12 below.

Figures 7, 8 and 9 show the partial residual plots of X_2 for the other three model fits at three different levels of β_N, -0.2, 0 and 0.2, respectively. In order to compare the model fits, the same scale has been used on all the plots. As shown clearly, all of the plots, except for the independent GAMs when $\beta_N = 0$ (which is equivalent to dependent severity GAMs in this situation) indicate a lack of fit and therefore that the true

Figure 4: Effect of the Dependence Level β_N on the Estimates of the Parametric Coefficients β_{y0} and β_{y1} for the Four Different Severity Models: (a) Intercept; (b) Slope

Note: $\beta_{y0} = 1$ and $\beta_{y1} = 2$

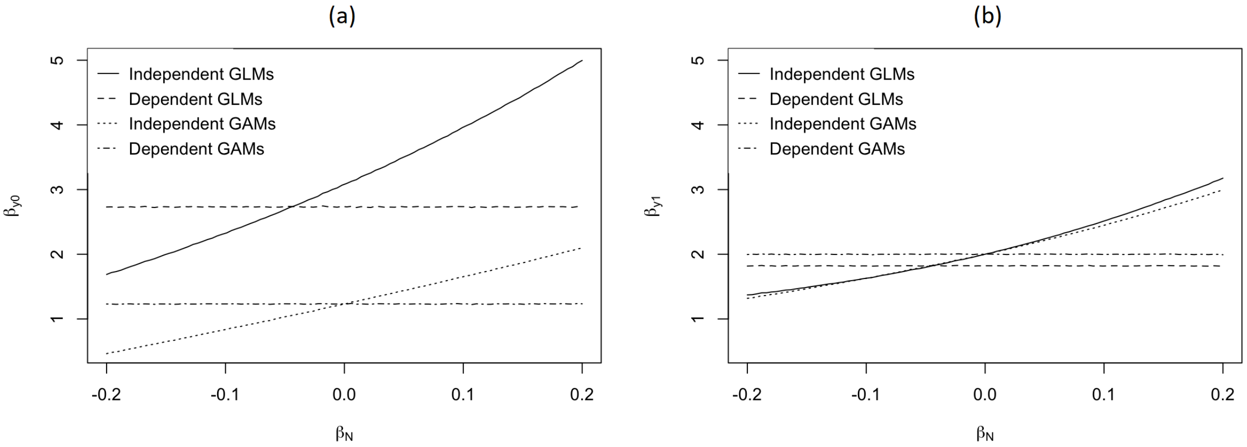
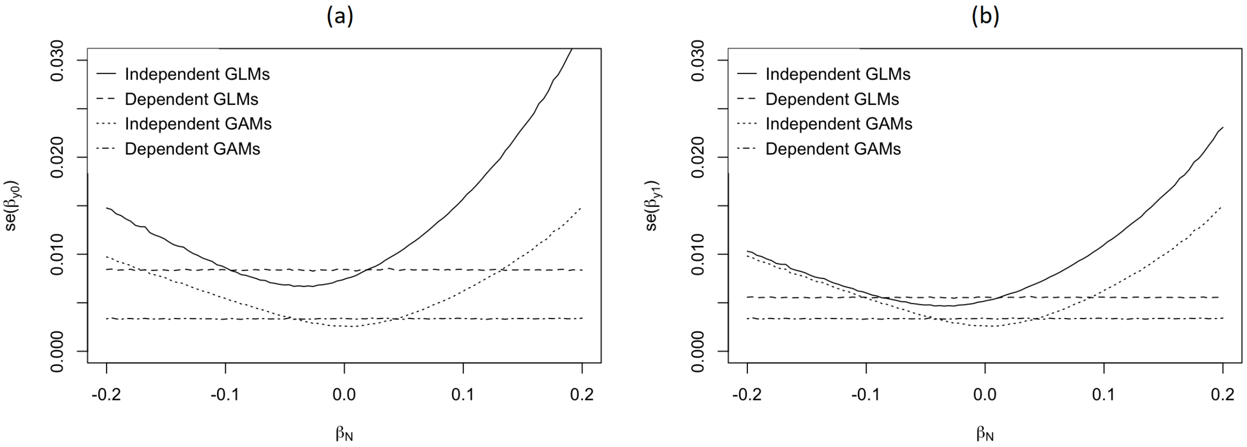


Figure 5: Effect of the Dependence Level β_N on the Estimates of the Parametric Coefficients β_{y0} and β_{y1} for the Four Different Severity Models: (a) Intercept; (b) Slope



curvature is not well described by the fitted line. When $\beta_N = -0.2$, the independent severity GLMs and GAMs do have similar fits for X_2, which corresponds to the curvatures we found in Figure 6.

In addition, the mean of the parametric bootstrap p-values (with 2000 repetitions) to test whether the smoothing of X_2 is needed over the 1000 iterations is equal to 0 for all β_N under both the independent and dependent case. Hence, we can conclude that the claim severity GAMs show better model fits than their parametric counterparts under both the independent and dependent assumptions.

4.4 Comparing the Tweedie GLM and GAM Simulations

Since there is no GCV criterion available to predict the variance function power p for the Tweedie GAMs, we could use REML to estimate p and then use the estimated p to fit the Tweedie distributions again by GCV in order to compare with the other model fits easily. When $\beta_N = 0$, the pure premiums were fitted

Figure 6: Effect of β_N on AIC of the Conditional Severity Models

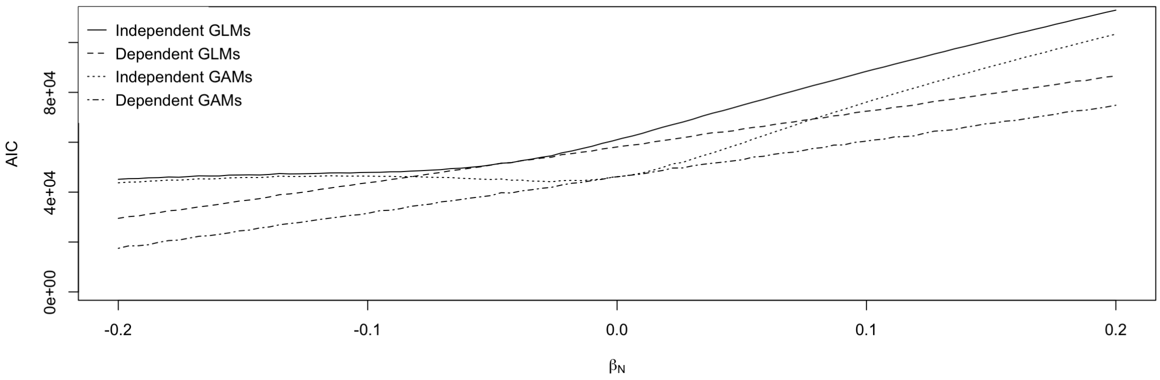
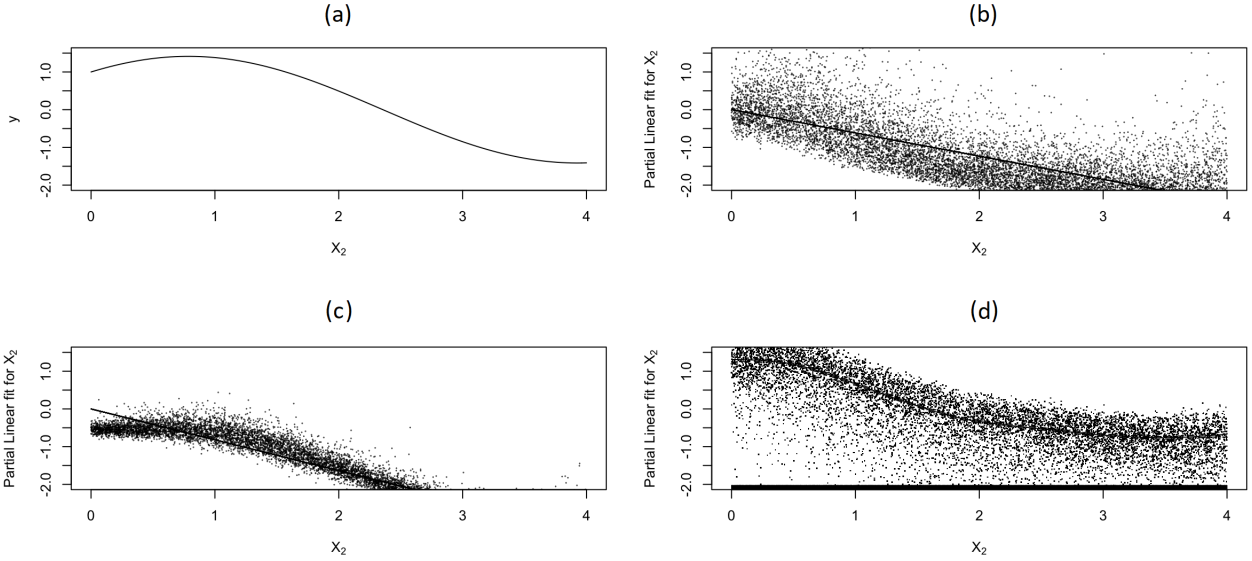


Figure 7: The Fits of (a) the True $\sin(X_2) + \cos(X_2)$ Function; and (b) by Independent GLMs, (c) Dependent GLMs, and (d) Independent GAMs, when $\beta_N = -0.2$



by Tweedie GLMs and GAMs using both REML and GCV to do the model selections. The product of the simulated claim frequency and severity, i.e., the pure premium, is our response.

By using the same simulation technique as described earlier, the whole claim frequency and severity dataset was regenerated 1000 times with $\beta_N = 0$. The Tweedie GLM and GAM were correspondingly refitted 1000 times. X_1 and X_2 were treated as rating variables, and X_2 was smoothed in the Tweedie GAMs. As shown in Table 5, the Tweedie GAMs clearly have smaller average standard errors of the estimation for both the intercept and the X_1 predictor, and thus more efficient. The REML and GCV criterion also give very similar estimations for both the Tweedie GLMs and GAMs, which confirms that the Tweedie model fittings are not significantly affected by using different model selection criteria, which might be used interchangeably. Table 6 provides the mean of AIC and estimated p. The lower AIC for the Tweedie GAMs fitted by both REML and GCV confirms that the Tweedie GAM is superior to the Tweedie GLM for fitting this simulated dataset when $\beta_N = 0$.

For one randomly generated dataset, Figures 10 and 11 present the partial residuals plots of X_2 fitted by the Tweedie GLM and GAM when $\beta_N = 0$. As shown in the plots, Tweedie GAMs give us a more reasonable fit of X_2 since the partial residuals seem to capture the main trend of X_2; this trend is not well described by

Figure 8: The Fits of (a) the True $\sin(X_2) + \cos(X_2)$ Function; and (b) by Independent GLMs, (c) Dependent GLMs, and (d) Independent GAMs, when $\beta_N = 0$

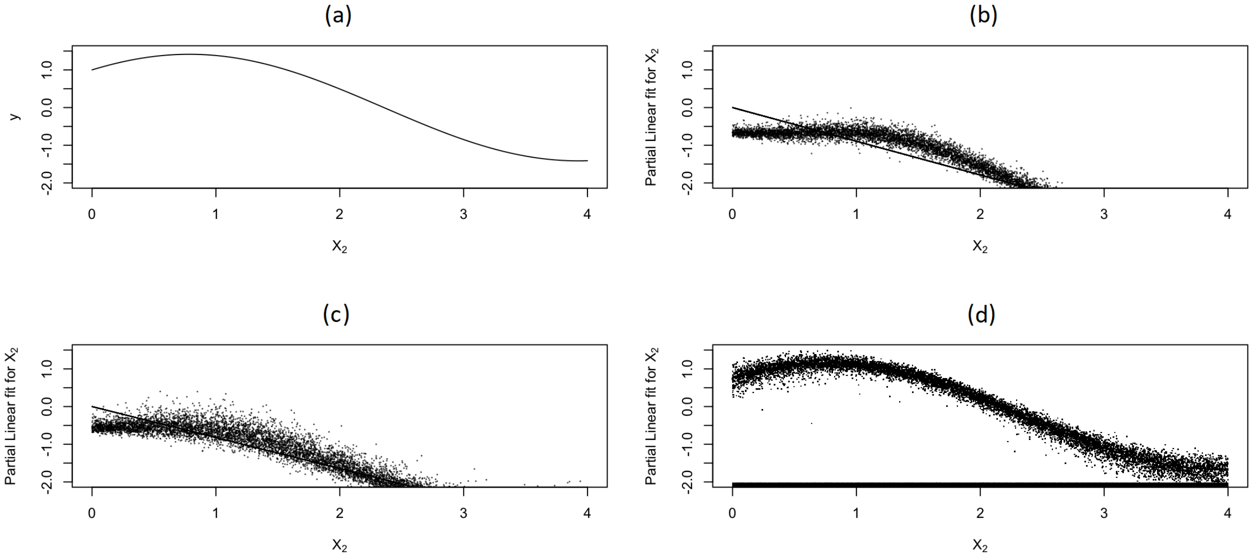
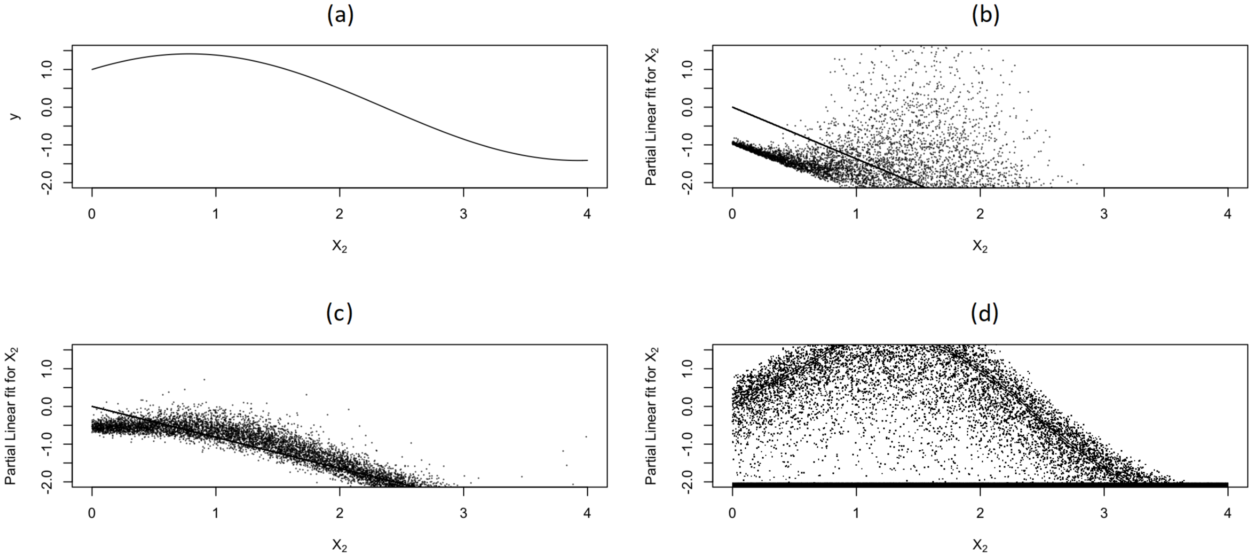


Figure 9: The Fits of (a) the True $\sin(X_2) + \cos(X_2)$ Function; and (b) by Independent GLMs, (c) Dependent GLMs, and (d) Independent GAMs, when $\beta_N = 0.2$



the partial residuals from Tweedie GLMs. However, even for the Tweedie GAMs, the partial residuals are not evenly distributed around the fitted line: the residuals below the line are more spread out than those above the model fitting, which suggests there is lack of fit for the Tweedie modellings. The possible reason for this is that the use of Tweedie has the limitation that the variance of the modelled response will only increase with its mean. Hence, double generalized linear models or zero-inflated Tweedie might be needed to modify the Tweedie fittings, as we alluded to earlier.

The residual plots for the Tweedie GLMs (not shown here to save space) show a tapering pattern while the deviance residuals for GAMs distribute more compactly and are more evenly spaced around zero, which

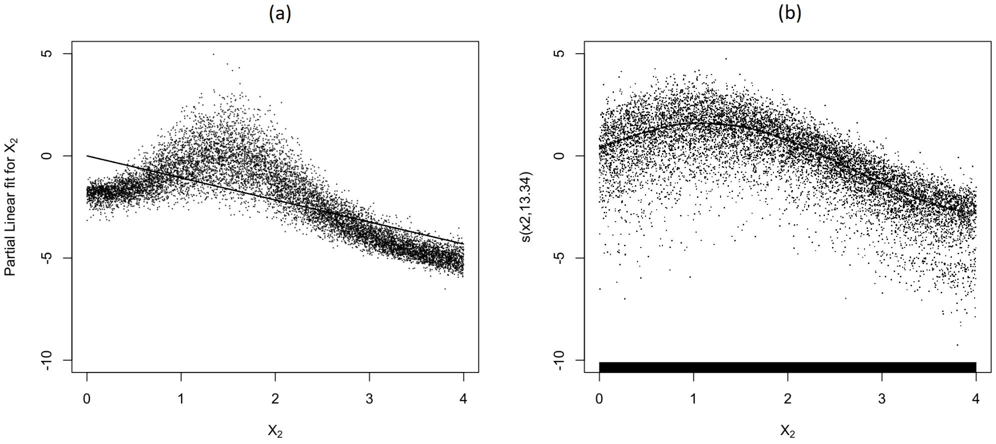
Table 5: Mean of the Regression Parameter Estimates for the Tweedie Modellings for $\beta_N = 0$

	GLM by REML	GLM by GCV	GAM by REML	GAM by GCV
Intercept $\hat{\beta}_{s0}$	5.0099	4.8965	2.6376	2.6378
se($\hat{\beta}_{s0}$)	0.0181	0.0171	0.0079	0.0074
$\hat{\beta}_{s1}$ for X_1	2.5000	2.5000	2.4999	2.4999
se($\hat{\beta}_{s1}$)	0.0111	0.0099	0.0050	0.0046
$\hat{\beta}_{s2}$ for X_2	-1.0718	-1.0032	-	-
se($\hat{\beta}_{s2}$)	0.0069	0.0066	-	-

Table 6: Mean of AIC and p by REML and GCV for Tweedie Models and $\beta_N = 0$

	AIC(REML)	AIC(GCV)	p
Tweedie GLM	117754.2	118334.4	1.70
Tweedie GAM	104851.6	104920.1	1.59

Figure 10: Partial Residuals Plots of X_2 when the Tweedie's are fit by REML: (a) GLM, (b) GAM



indicates the Tweedie GAMs provide a more adequate fit. In addition, the approximate F-test gives an extreme p-value=2.2×10^{-16} to reject the null hypothesis that smoothing X_2 is not needed in the Tweedie modellings.

When β_N varies from -0.2 to 0.2, the Tweedie GAMs and GLMs provide similar parametric estimations of X_1 but quite differing estimations of the intercept term. However, for both of the terms and both of the models, as the $|\beta_N|$ is away from 0, their estimations become increasingly biased. Figure 12 shows how the standard errors of $\hat{\beta}_{s0}$ and $\hat{\beta}_{s1}$ change with the level of β_N. As the $|\beta_N|$ tends to be larger, the standard errors of $\hat{\beta}_{s0}$ and $\hat{\beta}_{s1}$ for the Tweedie models are larger; and notice that the Tweedie GAMs always have lower standard errors for both of the terms than the Tweedie GLMs, which indicates the nonparametric approach is more efficient here. Furthermore, the Tweedie GAM always has a lower mean AIC value than the Tweedie GLM. And finally, the mean of the bootstrapping p-values to test where smoothing of X_2 is needed is equal to 0 under each β_N, which suggests the Tweedie GAMs provide more reasonable model fits. Therefore, we may safely infer that the Tweedie GAM gives a better fit compared with the Tweedie GLM regardless of the introduced dependence level between the claim frequency and severity.

Figure 11: Partial Residuals Plots of X_2 when the Tweedie's are fit by GCV: (a) GLM, (b) GAM

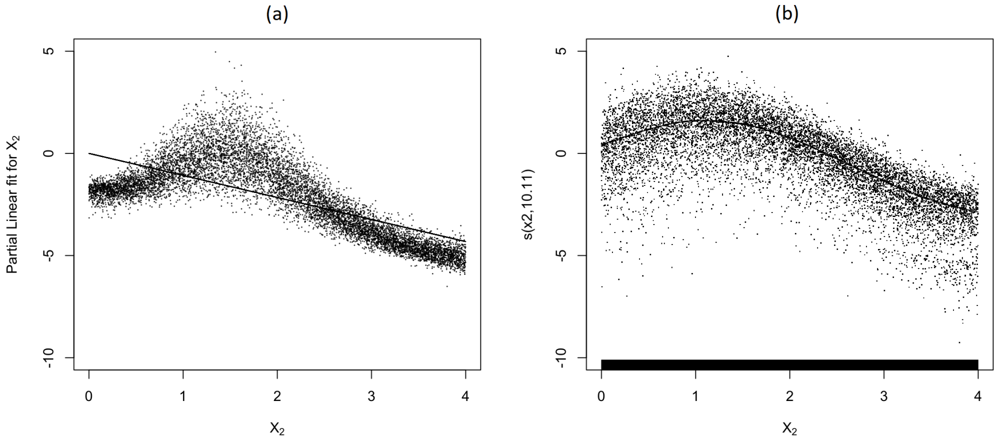
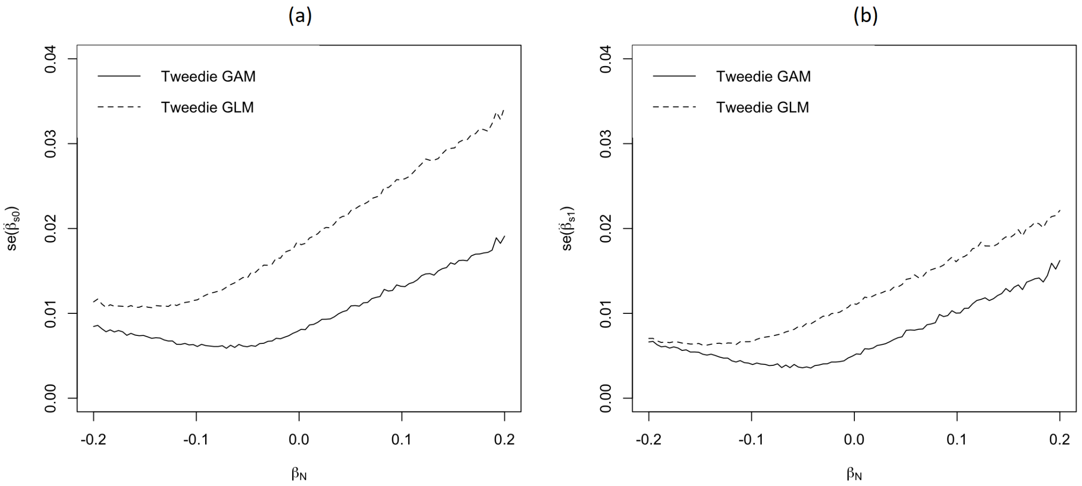


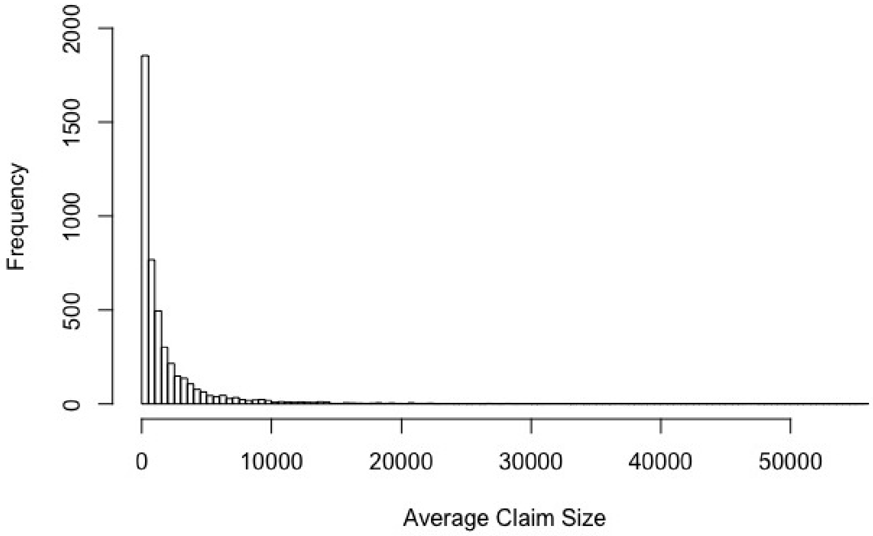
Figure 12: Standard Errors of (a) $\hat{\beta}_{s0}$ and (b) $\hat{\beta}_{s1}$ using Dependent β_N
Note that the Tweedie's are fitted by REML



5 Insurance Claims Study

In this section, the frequentist GAMs allowing for nonparametric and/or strongly nonlinear trend terms in aggregate claims under both the independent and dependent assumption are applied to a real insurance dataset and a hold-out dataset (where we test the model fits derived from the training set). The performance of different aggregate claims models (i.e. independent GLMs, dependent GLMs, independent GAMs and dependent GAMs) are compared.

Figure 13: Claim Severity Histogram for the Positive Claim Counts



5.1 Data Description

The `dataCar` dataset is retrieved from the R package `insuranceData`. A collection of 67856 one-year auto insurance policies taken out in 2004 or 2005 were recorded on 10 variables (Jong & Heller 2008). The variables in the original dataset are listed in Table 7. Notice that there are only a limited number of rating variables available for this dataset, whereas typical industry data are far more comprehensive. Since all policies were insured for a full year, but not necessarily provided at the same point in time, the 'Exposure' variable in this dataset measures the proportion of the whole one year coverage provided at the specific time point when data was collected. For example, an exposure unit of 0.5 implies the individual policyholder was insured for half a year. Naturally, as the exposure increases, the number of claims and pure premium will increase proportionally. As such, the claim frequency and pure premium components in the following GLMs and GAMs need to be divided by the 'Exposure' variable.

Table 7: Variables in the Vehicle Insurance Dataset

Variable	**Range**
Vehicle value	\$0-\$350,000 (recorded in \$10,000s)
Exposure	[0,1]
Claim occurrence	0(no), 1(yes)
Number of claims	0, 1, 2,...
Average claim size	\$0-\$56,000 (\$0 if no claim)
Vehicle body type	bus, convertible, coupe, hatchback, hardtop, motorized caravan/combi, minibus, panel van, roadster, sedan, station wagon, truck, utility
Vehicle age	1(new), 2, 3, 4
Gender	male, female
Area of residence	A, B, C, D, E, F
Age band of policyholders	1(youngest), 2, 3, 4, 5, 6

Table 8: Claim Count Distribution

Claim Count	**Frequency**	**Percentage**
0	63,232	93.19%
1	4,333	6.39%
2	271	0.40%
3	18	0.02%
4	2	0.00%
Total	67,856	100%

There are 4624 (6.8%) out of the 67856 policies which had positive claim counts (at least one claim). The frequency table of 'Number of claims' is given in Table 8. The 'Number of claims' ranges from 0 to 4 with a substantial number of zero claims (93.2%). The excess of zeros suggests that the claim frequency distribution assumption of a Poisson response might not be adequate.

Table 9: Mean of Claim Severity Grouped by Claim Counts

Claim Count	**Mean of Claim Severity**
0	$0
1	$1947
2	$2945
3	$4024
4	$4439

Given that there exists at least one claim, the average claim severity ranged from $200 to $55922. The distribution of claim size for the positive claim counts is presented in Figure 13. And Table 9 shows that, given a loss had been incurred, the average claim size seemed to be larger as the number of claims increases. So this implies a positive correlation between the claim frequency and claim severity can be expected.

5.2 Modelling the Data

In terms of the claim frequency, claim severity, and pure premium modelling for this dataset, the 'Number of claims', 'Average claim size' and 'Pure premium' (i.e., 'Number of claims' × 'Average claim size') are naturally viewed as the response variables, respectively. 'Vehicle value' is treated as a continuous rating variable; 'Vehicle age' and 'Age band of policyholders' are viewed as discrete ordinal variables; while 'Vehicle body type', 'Gender' and 'Area of residence' are considered as factors. In addition, the 'Number of claims' is also fitted as a continuous predictor in the dependent marginal severity GLMs and GAMs.

To be clear, the purpose of this study is not to try and achieve a best model fit to the dataset, but rather to quantify the effect of nonparametric trend terms and dependence in aggregate claims by using and comparing different modelling methods. Hence, a Poisson frequency distribution and a gamma severity distribution will still be assumed for the data when fitting all of the following aggregate claims models. Also, we will not worry about the complication of adding interaction terms to the models, again for said reasons above.

Since we have such a plenitude of data, we can safely partition the data into two componentss: two-thirds will be used as a training set to fit the models and the remaining one-third will be reserved as a hold-out test set to further test and validate the model fits derived from the training set.

5.3 GLMs for Aggregate Claims Under Independence

The `glm` function in R was used to model and analyze the dataset. Note that there are 81 policyholders with a 'Convertible' in the 'vehicle body type', of which three policyholders have average claim sizes of $530,

$6126 and $233, while the remaining 78 policyholders did not have any claims at all. Also, there are only 48 cases with a 'Bus' and 27 policyholders having a 'Roadster' car, of which the sample size is too small to achieve a hold-out test setting. Besides, the standard errors might very well be inflated for the three categories in the 'Vehicle body' predictor. Thus, these three types in the 'Vehicle body' predictor, namely, 'Bus' 'Convertible' and 'Roadster' were eliminated from the analysis.

Following the independence assumptions and modelling techniques set for in Sections 2 and 3, the method of modelling the marginal frequency and severity components separately under independent GLMs as well as the Tweedie GLM approach were both employed to fit the dataset. As mentioned earlier, the 'exposure' adjustment is needed in order to analyze the data in a consistent manner. The `offset` function in R could be used for the frequency and pure premium models, which divides the 'Number of claims' and 'Pure premium' by the amount of 'Exposure' respectively. Note that it is not necessary to use an offset for the severity models since we are modelling the average loss amount per claim occurrence, which is not proportional to the exposure variable.

In the frequency and pure premium models, the weights were taken to be 1 for each observation. However, in the severity models, the claim counts were used as weights. Recall from Subsection 2.6, when we model the average claim amounts, we have that the average claim size $\bar{Y}_i \sim gamma(\mu_{i2}, \phi/n_i)$. That is, we have $a_i(\phi) = \phi/n_i$, and hence the claim frequency n_i must be used as weights in the following GLM and GAM models.

5.3.1 Frequency GLM under Independence

Let us denote the exposure variable by e_i. Then, according to Equation (8), the Poisson GLM for $E[N_i|\boldsymbol{X}_i]$ (the mean of 'Number of claims' N_i given the covariate vector $\boldsymbol{X}_i$) with an exposure offset is given by:

$$\ln\left(\frac{E[N_i|\boldsymbol{X}_i]}{e_i}\right) = \ln\left(\frac{\mu_{i1}}{e_i}\right) = \eta_{i1} = \boldsymbol{X}_{i1}^{\top}\boldsymbol{\beta_1}. \tag{28}$$

In using an 'offset' for N_i, we are essentially including the exposure variable as a fixed effect with regression coefficient equal to 1. The full main effects model for $N_i|\boldsymbol{X}_i$ is as follows:

$$\begin{aligned}\ln(E[N_i|\boldsymbol{X}_i]) = \ln(\texttt{Exposure}) + \texttt{Vehicle value} + \texttt{Vehicle body type} + \texttt{Gender}\\ + \texttt{Area of residence} + \texttt{Age band of policyholder}.\end{aligned} \tag{29}$$

By using the `step` function in R (i.e., using the AIC criterion to determine the best subset of rating variables), we get the final selected frequency GLM by the stepwise procedure:

$$\begin{aligned}\ln(E[N_i|\boldsymbol{X}_i]) = \ln(\texttt{Exposure}) + \texttt{Vehicle value} + \texttt{Vehicle body type}\\ + \texttt{Age band of policyholder},\end{aligned} \tag{30}$$

where the 'Gender' and 'Area of residence' variables were removed from the model fit. Table 10 provides the estimated coefficients and standard errors for the frequency GLM.

5.3.2 Marginal Independent Severity GLM

Following Equation (9), the marginal severity GLM under the independent setting can be expressed as:

$$\ln(E[\bar{Y}_i|\boldsymbol{X}_i]) = \ln(\mu_{i2}) = \eta_{i2} = \boldsymbol{X}_{i2}^{\top}\boldsymbol{\beta_2}. \tag{31}$$

The full main effects model for $\bar{Y}_i|\boldsymbol{X}_i$ is given by:

$$\begin{aligned}\ln(E[\bar{Y}_i|\boldsymbol{X}_i]) = \texttt{Vehicle value} + \texttt{Vehicle body type} + \texttt{Gender}\\ + \texttt{Area of residence} + \texttt{Age band of policyholder}.\end{aligned} \tag{32}$$

By using the `step` function as in the frequency case, the best subset of predictors is given by ('Vehicle value', 'Gender', 'Area of residence', 'Age band of policyholder') which removes the 'Vehicle body type' variable and contains only four predictors. This reduced final model has the minimum AIC equal to 52766 and can be written as:

$$\begin{aligned}\ln(E[\bar{Y}_i|\boldsymbol{X}_i]) = {} & \texttt{Vehicle value} + \texttt{Gender} + \texttt{Area of residence} \\ & + \texttt{Age band of policyholder}.\end{aligned} \tag{33}$$

Table 13 provides the estimated coefficients and standard errors for the independent severity GLM.

5.3.3 Tweedie GLM for the Pure Premium

As part of a Tweedie GLM fitting for the pure premium, the variance function exponent parameter p can be estimated by likelihood methods. The R `mgcv` package has an implementation of it, which uses the `gam` function instead of `glm`, but otherwise the syntax is familiar (Faraway 2016). Notice that a Tweedie distribution with automatic estimation of the parameters p is only available with REML or ML smoothing parameter estimation (Wood 2017). A log-link function was specified to this model fit. The full main effects model for the pure premium, given the covariates, is as follows:

$$\begin{aligned}\ln(E[S_i|\boldsymbol{X}_i]) = {} & \ln(\texttt{Exposure}) + \texttt{Vehicle value} + \texttt{Vehicle body type} + \texttt{Gender} \\ & + \texttt{Area of residence} + \texttt{Age band of policyholder}.\end{aligned} \tag{34}$$

By comparing the AIC scores of different model fits, the final Tweedie GLM with the minimum AIC=99787.78 is given by

$$\begin{aligned}\ln(E[S_i|\boldsymbol{X}_i]) = {} & \ln(\texttt{Exposure}) + \texttt{Gender} + \texttt{Area of residence} \\ & + \texttt{Age band of policyholder},\end{aligned} \tag{35}$$

where the estimated p is equal to 1.578 and the score of REML equal to 38113.

5.4 GLMs for Aggregate Claims Under Dependence

5.4.1 Marginal Dependent Severity GLM

In the dependent setting, the marginal frequency GLM is modelled the same way as in the independent setting. So the final model for the mean claim counts in the dependent setting is same as the model shown in Equation (30).

5.4.2 Marginal Dependent Severity GLM

According to the dependent assumptions and modelling techniques outlined in Section 2, the gamma GLM for the average claim size given the covariate vector is given by

$$\ln(E[\bar{Y}_i|\boldsymbol{X}_i, N_i]) = \ln(\tilde{\mu}_{i2}) + N_i\beta_N = \tilde{\boldsymbol{X}}_{i2}^{\top}\tilde{\boldsymbol{\beta}}_2 + \beta_N N_i. \tag{36}$$

The full main effect dependent severity model is as follows:

$$\begin{aligned}\ln(E[\bar{Y}_i|\boldsymbol{X}_i, N_i]) = {} & \texttt{Vehicle value} + \texttt{Vehicle body type} + \texttt{Gender} \\ & + \texttt{Area of residence} + \texttt{Age band of policyholder} + N_i.\end{aligned} \tag{37}$$

By the AIC model selection criterion, the final model is given by

$$\begin{aligned}\ln(E[\bar{Y}_i|\boldsymbol{X}_i, N_i]) = {} & \texttt{Vehicle value} + \texttt{Gender} + \texttt{Area of residence} \\ & + \texttt{Age band of policyholder} + N_i.\end{aligned} \tag{38}$$

5.5 GAMs for Aggregate Claims Under Independence

For both the marginal frequency and severity model, a GAM with cubic penalized regression splines smoothing of the continuous covariate 'Vehicle value' was applied to the model fittings. The smoothing parameter λ in each of the models was obtained by GCV in this analysis. Twenty evenly-spaced knots seemed to be a good choice for both cubic penalized regression splines of the 'Vehicle value' term in each marginal model respectively, due to the large sample size. The maximum basis dimension k was chosen to be 20 for both of the frequency model and the severity model, which is a bit higher than the estimated degrees of freedom of the smoothed predictor 'Vehicle value'.

5.5.1 Frequency GAM under Independence

The full main effects frequency GAM with the smoothed predictor ‘Vehicle value’ is as follows:

$$\begin{aligned}\ln(E[N_i|\boldsymbol{X}_i]) = \ln(&\texttt{Exposure}) + \texttt{Vehicle body type} + \texttt{Gender}\\ &+ \texttt{Area of residence} + \texttt{Age band of policyholder}\\ &+ S_f(\texttt{Vehicle value}),\end{aligned} \tag{39}$$

where $S_f(\cdot)$ stands for the smoothed function for the ‘Vehicle value’ term. Since there is no `step.gam` function available in R package `mgcv`, model selection had to be done by dropping each predictor term in order and comparing the AIC scores of them. The best model with the lowest AIC value 23195.86 is given by

$$\begin{aligned}\ln(E[N_i|\boldsymbol{X}_i]) = \ln(&\texttt{Exposure}) + \texttt{Vehicle body type} + \texttt{Age band of policyholder}\\ &+ S_f(\texttt{Vehicle value}),\end{aligned} \tag{40}$$

where the same set of predictors as kept in the frequency GLM case (see Equation (30)).

5.5.2 Marginal Independent Severity GAM

With the ‘Vehicle value’ rating variable smoothed in the independent severity model, and after the AIC model selection process, the final independent severity GAM is determined to be

$$\begin{aligned}\ln(E[\bar{Y}_i|\boldsymbol{X}_i]) = \;&\texttt{Gender} + \texttt{Area of residence} + \texttt{Age band of policyholder}\\ &+ S_y(\texttt{Vehicle value}),\end{aligned} \tag{41}$$

where the same set of predictors are kept as in the independent severity GLM case (see Equation (33)).

5.5.3 Tweedie GAM

With the ‘Vehicle value’ predictor smoothed, the full Tweedie GAM containing all the main effects is given by

$$\begin{aligned}\ln(E[S_i|\boldsymbol{X}_i]) = \ln(&\texttt{Exposure}) + \texttt{Vehicle body type} + \texttt{Gender}\\ &+ \texttt{Area of residence} + \texttt{Age band of policyholder}\\ &+ S_s(\texttt{Vehicle value}),\end{aligned} \tag{42}$$

where $S_s(\cdot)$ stands for the smooth function for the ‘Vehicle value’. The maximum basis dimension k was chosen to be 20 for the Tweedie GAM, which is a bit higher than was estimated. By the AIC model selection, the final model is the same as Equation (35) in the Tweedie GLM case, which means that the ‘Tweedie GLM’ and ‘Tweedie GAM’ are identical in this case, since they both exclude the ‘Vehicle value’ term and keep the same group of rating variables.

5.6 GAMs for Aggregate Claims Under Dependence

5.6.1 Frequency GAM under Dependence

In the dependent setting, the marginal frequency GAM is modelled the same way as in the independent setting (see Equation (40)).

5.6.2 Marginal Dependent Severity GAM

By using the AIC model selection criterion, the final marginal dependent severity GAM with the minimum AIC is given by

$$\begin{aligned}\ln(E[\bar{Y}_i|\boldsymbol{X}_i, N_i]) = \;&\texttt{Gender} + \texttt{Area of residence} + \texttt{Age band of policyholder}\\ &+ N_i + S_y(\texttt{Vehicle value}),\end{aligned} \tag{43}$$

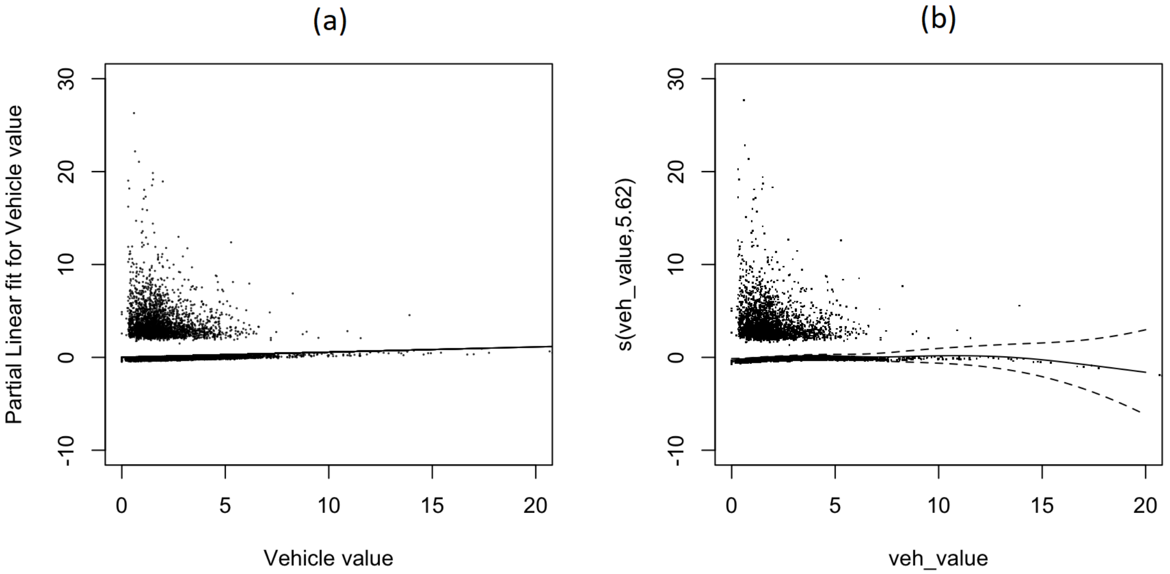
where this final model keeps the identical set of predictors as in the dependent severity GLM (see Equation (38)).

5.7 Model Comparison of the Frequency GLM and GAM

Figure 14 gives the partial residuals plots for both the GLM (left panel) and GAM (right panel) fittings. The parameter estimate of 'Vehicle value' in the frequency GLM is 0.05626, which is significant at the 0.01 level. Through utilizing cubic penalized spline smoothing for the frequency GAM model, the estimated effect of the 'Vehicle value' is shown as the solid curve, with the dashed lines standing for 95% Bayesian credible intervals, and with the numbers on the vertical axis representing the degrees of freedom for the smooth. Approximately one degree of freedom corresponds to a nearly straight line, which estimates the effect of the predictor. In the frequency GAM, the effect of 'Vehicle value' was estimated as a smooth curve with 5.62 degrees of freedom, which corresponds to a substantially curved shape of fitting, as depicted in the right panel of Figure 14. It is clearly evident that these were inadequate fits for the frequency models since the fitted curves do not capture the trend of 'Vehicle value' and the partial residuals are not evenly distributed around the fittings. Also note that the partial residuals that are significantly above the 0 level correspond to those records with claim frequency of 1 or greater.

The regression parameter estimates for the frequency GLM and GAM are shown in Tables 10 and 11. A comparison shows that the coefficient estimates are all very similar, as are their accompanying standard errors. Notice that although most of the coefficient estimates are close to zero, on the mean scale, there will actually ensue a considerable influence on the expected value since a log-link function is be being used. Residual plots for the claim frequency responses are not overly helpful in assessing the goodness of fit in this case.

Figure 14: Partial Residual Plots of `Vehicle Value` for the (a) Frequency GLM and (b) the GAM Fit



In addition to the graphical diagnostics, hypothesis tests are also used to test whether there is an improvement of fit from the frequency GLM to GAM given the same set of predictors kept in the final models. The approximate F-statistic of 3.3923 gives a p-value equal to 0.005858, which indicates there is significant evidence to reject the null hypothesis H_0 : *the smoothing of the 'Vehicle value' term is not necessary or there is just a linear trend for the 'Vehicle value' term.* So there is a need to smooth the 'Vehicle value' covariate and the frequency GAM provides a more reasonable fit. Based on 2000 bootstrap samples, the parametric bootstrap p-value is given by 0.004, which confirms what we found in the approximate F-test. Also, by a lower AIC value and higher percentage of adjusted R^2 for the frequency GAM, we can conclude that the GAM performs better than the GLM fitting in predicting the claim frequency in this case study.

Table 10: Regression Parameter Estimates - Frequency GLM

Regression Parameter	Estimate	Standard Error	z-value	p-value
Intercept	-1.0648	0.4497	-2.37	0.02
Vehicle Value	0.0563	0.0152	3.70	$2.1 \cdot 10^{-4}$
Vehicle Body COUPE	-0.2309	0.4687	-0.49	0.62
Vehicle Body HBACK	-0.6125	0.4485	-1.37	0.17
Vehicle Body HDTOP	-0.4998	0.4590	-1.09	0.28
Vehicle Body MCARA	-0.3165	0.5711	-0.55	0.58
Vehicle Body MIBUS	-0.5313	0.4781	-1.11	0.27
Vehicle Body PANVN	-0.5681	0.4725	-1.20	0.23
Vehicle Body SEDAN	-0.5828	0.4483	-1.30	0.19
Vehicle Body STNWG	-0.6019	0.4490	-1.34	0.18
Vehicle Body TRUCK	-0.6276	0.4600	-1.36	0.17
Vehicle Body UTE	-0.8660	0.4537	-1.91	0.06
Age Band	-0.0888	0.0126	-7.05	$1.8 \cdot 10^{-12}$

Table 11: Regression Parameter Estimates - Frequency GAM

Regression Parameter	Estimate	Standard Error	t-value	p-value
Intercept	-0.9490	0.4551	-2.09	0.04
Vehicle Body COUPE	-0.21704	0.4752	-0.46	0.65
Vehicle Body HBACK	-0.6207	0.4549	-1.36	0.17
Vehicle Body HDTOP	-0.5565	0.46540	-1.20	0.23
Vehicle Body MCARA	-0.3080	0.5783	-0.53	0.59
Vehicle Body MIBUS	-0.5726	0.4846	-1.18	0.24
Vehicle Body PANVN	-0.5837	0.4789	-1.22	0.22
Vehicle Body SEDAN	-0.6011	0.4545	-1.32	0.19
Vehicle Body STNWG	-0.6448	0.4551	-1.42	0.16
Vehicle Body TRUCK	-0.6713	0.4663	-1.44	0.15
Vehicle Body UTE	-0.9083	0.4600	-1.97	0.05
Age Band	-0.0874	0.0127	-6.88	$6.0 \cdot 10^{-12}$

5.8 Comparison of the Conditional Severity Models

Tables 12- 15 provide the details on the estimated parameters and standard errors obtained for the four severity models. As seen from the tables, all of the parameter estimates are very low for each model and the parameter estimations do not change appreciably from one model to another.

For the dependent severity GLM, the estimated coefficient for the frequency covariate $\hat{\beta}_N$ was 0.35213. By performing a Wald test for the N_i predictor, where $H_0 : \beta_N = 0$ vs $H_a : \beta_N \neq 0$, we obtain:

$$\frac{\hat{\beta}_N - \beta_N}{\sqrt{Var(\hat{\beta}_N)}} = \frac{0.35213 - 0}{0.11306} = 3.115. \tag{44}$$

The z-score 3.115 leads to a very small p-value(< 0.01), which indicates the number of claims covariate is strongly significant and therefore there is evidence that a dependent severity model should be used in modelling the aggregate claims. In addition, an approximate F-statistic wass also used to test the number of claims covariate. Since the severity gamma GLM has a free dispersion parameter, F-tests instead of χ^2 tests should be applied here. The F-test conducted using the `drop1` function in R also yields an extreme p-value < 0.01, which confirms that accounting for dependence can improve the model fit and that the independent

Table 12: Regression Parameter Estimates - Independent GLM Severity Model

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	7.6874	0.1071	71.78	$< 2 \cdot 10^{-16}$
Vehicle value	-0.0317	0.0259	-1.22	0.22
Gender Male	0.1899	0.0606	3.13	$1.8 \cdot 10^{-3}$
Area B	-0.0060	0.0903	-0.07	0.95
Area C	0.0236	0.0814	0.29	0.77
Area D	-0.0208	0.1100	-0.19	0.85
Area E	0.2001	0.1184	1.69	0.09
Area F	0.4781	0.1390	3.44	$6.0 \cdot 10^{-4}$
Age Band	-0.0599	0.0212	-2.83	$4.7 \cdot 10^{-3}$

Table 13: Regression Parameter Estimates - Dependent GLM Severity Model

Regression Parameter	**Estimate**	**Standard Error**
Intercept	7.2912	0.16291
Vehicle value	-0.0321	0.0262
Gender Male	0.1911	0.0612
Area B	0.0096	0.0913
Area C	0.0474	0.0823
Area D	-0.0086	0.1112
Area E	0.2102	0.1195
Area F	0.4872	0.1404
Age Band	-0.0591	0.0214
Number of Claims	0.3521	0.1131

Table 14: Regression Parameter Estimates - Independent GAM Severity Model

Regression Parameter	**Estimate**	**Standard Error**
Intercept	7.5951	0.0963
Gender Male	0.2217	0.0602
Area B	0.0118	0.0894
Area C	0.0261	0.0806
Area D	-0.0100	0.1090
Area E	0.2314	0.1174
Area F	0.5661	0.1387
Age Band	-0.0613	0.0210

model is not an adequate simplification of the dependent model (or *is* an oversimplification). The same tests were also conducted for the dependent severity GAM case. The extreme p-values obtained from both of the two tests indicate the coefficient for the frequency covariate $\tilde{\beta}_N$ is strongly significant in the dependent severity GAM model. Moreover, for both of the dependent severity GLM and GAM, $\hat{\beta}_N = 0.35213 > 0$ and $\hat{\hat{\beta}}_N = 0.369401 > 0$ imply that there is a positive correlation between the claim frequency and severity, which is consistent with what we expected, as outlined earleir. It is important to notice that as the claim frequency increases, the estimated mean of the average claim sizes also increases in the dependent severity model. Hence, using the framework of accounting for dependence in the dependent models can avoid ignoring the association effect between claim frequency and severity under both the GLM and GAM structures, by

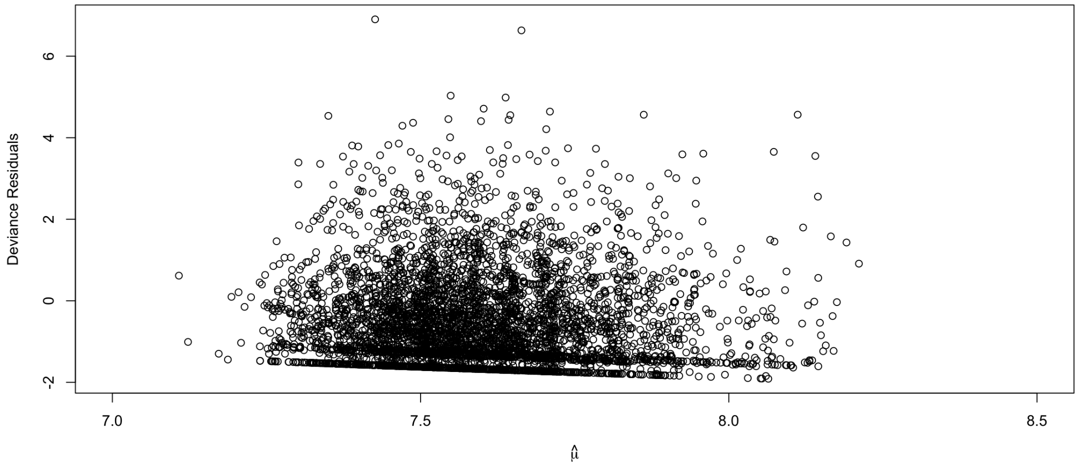
Table 15: Regression Parameter Estimates - Dependent GAM Severity Model

Regression Parameter	**Estimate**	**Standard Error**
Intercept	7.1773	0.1557
Gender Male	0.2244	0.0607
Area B	0.0293	0.0904
Area C	0.0526	0.0814
Area D	0.0054	0.1102
Area E	0.2434	0.1185
Area F	0.5818	0.1400
Age Band	-0.0607	0.0212
Number of Claims	0.3694	0.1121

accurately reflecting the dependence between N_i and $\bar{Y}_i$, thereby adjusting for the estimate of mean aggregate claims, $E[S_i]$, accordingly.

Figures 15, 16, 17 and 18 present the residuals plots with the deviance residuals against fitted values for assessing the goodness of fit of each severity model. The deviance residuals of severity GAMs are more compact and more evenly distributed around 0 as the linear predictor differs compared with their GLM counterparts under both the independent and dependent assumptions. There is clear evidence of GAMs improving the severity model fits since their residuals are closer to meet the model assumption of constant or homogeneous variance of the residuals compared with the parametric approach. In addition, compared with the independent severity models, the deviance residuals of the dependent models also distribute more compactly. Also, the deviance residuals are more evenly spaced around 0 and there is no obvious trend or pattern shown under both the GLM and GAM structures, which suggest the dependent severity models potentially provide a better model fit of the average claim sizes. Notice that for the frequency models, the largely above zero residuals corresponds to the non-zero claim counts. The straight line pattern shown in the bottom of each severity residual plot corresponds to the 'deductible' of $200 for the observed average claim amount. Recall that in Subsection 5.1, we found that the minimum value of average claim severity was $200, given a loss had occurred.

Figure 15: Residual Plot for the Independent Severity GLM



To test whether there really is a need to smooth the 'Vehicle value' predictor for both of the independent and dependent severity model, approximate F-tests were performed by using the `anova` function in R. For the independent case, the F-statistic is given by 2.3807 and the approximate p-value = 0.006263 is strongly significant. For the dependent case, the approximate F-statistic is 2.4348 with the corresponding p-value

Figure 16: Residual Plot for the Dependent Severity GLM

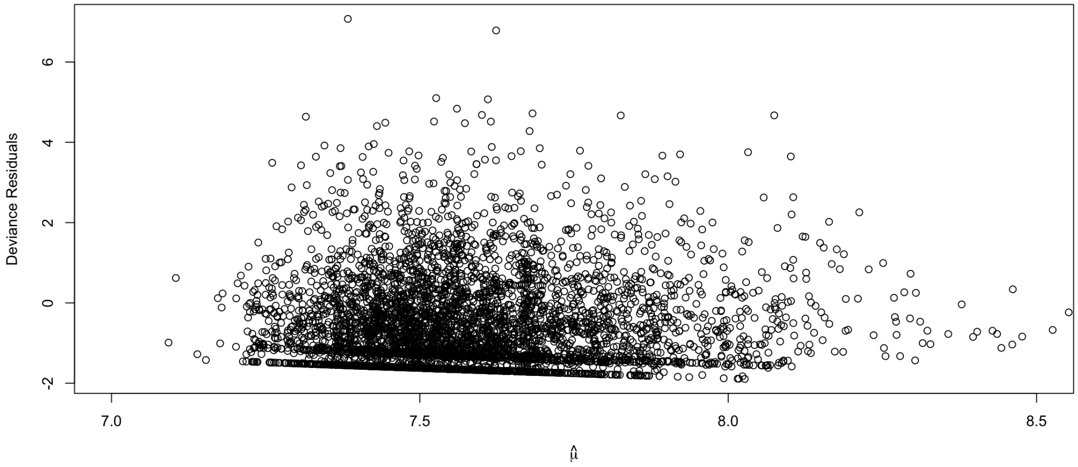


Figure 17: Residual Plot for the Independent Severity GAM

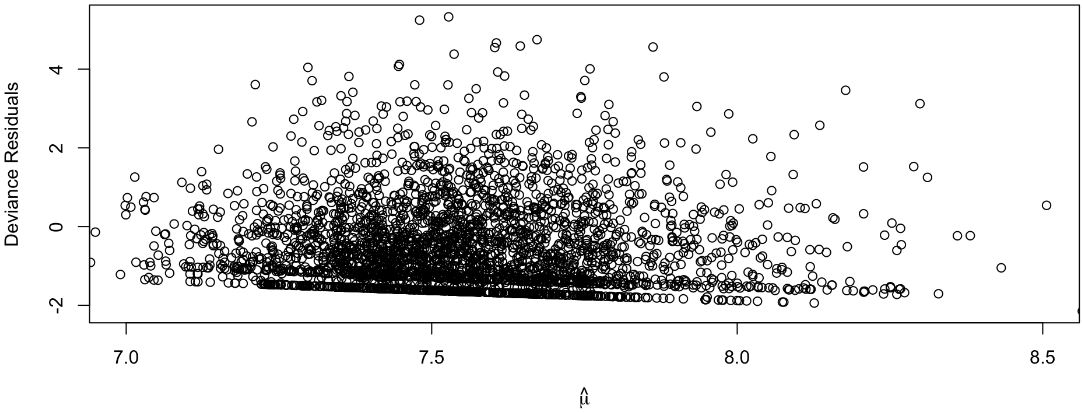
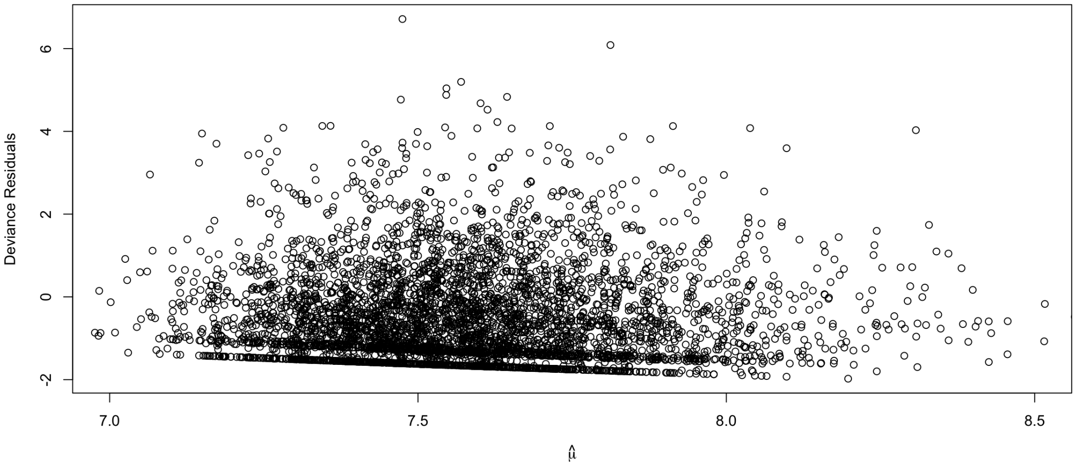


Figure 18: Residual Plot for the Dependent Severity GAM



equal to 0.005071, which is also significant at the 0.01 level. The parametric bootstrapping approach was used to obtain the actual distribution of the F-statistic. As shown in Table 16, compared with the bootstrap

p-values, the approximate p-values seem to be more liberal. This result is also confirmed by the unpublished work by C. Crainiceanu (Ruppert et al. 2003). Both of the bootstrap p-values are not significant at all. This might be because 'Vehicle value' is not a significant predictor in both the dependent and independent severity GLMs (but AIC selection criterion kept this predictor in both of the models).

For greater clarity, the "more liberal" approximate p-values from the GAM models implies that its p-values are smaller than the bootstrap p-values while all else is held fixed. That is to say, the predictors estimated using the GAM modeling method are more likely to be significant, so it's considered, "more liberal". From the opposite perspective, the bootstrapped p-values are "more conservative" since those values are less likely to indicate the predictor's significance while all else is held fixed. Ruppert et al. (2003) also reached that conclusion based on their study results.

Table 16: Approximating versus Bootstrapping p-values

Model	F**-value**	**Approximate** p**-value**	**Bootstrap** p**-value**
Independent Severity Models	2.3807	0.006263	0.282
Dependent Severity Models	2.4348	0.005071	0.247

Table 17: Results of Severity Model Comparisons

Model	**AIC**	**adjusted** R^2
Independent Severity GLM	52766.00	0.306%
Dependent Severity GLM	52744.00	0.274%
Independent Severity GAM	52732.22	1.570%
Dependent Severity GAM	52707.51	1.910%

Table 17 provides the AIC and adjusted R^2 for each severity model. GAMs have larger adjusted R^2 and AIC values under both the independent and dependent setting, which implies that the severity GAMs provide more reasonable and better fittings than their GLM counterparts. And the dependent severity GAM clearly has the lowest AIC and largest adjusted R^2 value, which indicate that accounting for the nonlinear relationship between the risk and the predictor 'Vehicle value', as well as the dependent effect between the claim frequency and severity, significantly improves the model fit compared with the parametric and/or the independent models. Notice that the adjusted R^2 values are very small. That is potentially because a large sample size was fitted by a limited number of rating variables. Also, we note that a low R-squared value indicates that the predictors in the models are not explaining much of the variation in the dependent variable. With more metadata and information of the source study dataset, further clean-up of the source dataset and/or adding more non-correlated predictors can be conducted to improve the models. Even if the R^2 values are small, the improvement of its values can still demonstrate GAMs superior performance over GLMs in this space, which is one of the primary aims of this paper.

5.9 Hold-out Dataset Validation

Notice that a hold-out dataset was also used to test the models derived from the training dataset, which was randomly sampled from and accounts for one third of the original dataset. Although there are some differences for the parameter estimations, very close estimates for the significant predictors were obtained and very similar conclusions were drawn as in the previous analysis of the training dataset.

To test whether there is a need to smooth the 'Vehicle value' predictor in the claim frequency model. The approximate F-value is given by 4.0811 with a significant p-value = 0.0005049 and the bootstrap p-value is given by 0.001, which both indicate there is a nonlinear trend for the 'Vehicle value' covariate. For the comparison of the parametric and nonparametric severity model under the independent and dependent setting, the bootstrap p-values are given by 0.420 and 0.467, respectively, which indicate there is no evidence

Table 18: Regression Parameter Estimates - Frequency GLM

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	-0.2895	0.4515	-0.64	0.52
Vehicle Value	0.0589	0.0224	2.63	$8.6 \cdot 10^{-3}$
Vehicle Body COUPE	-0.9137	0.4917	-1.86	0.06
Vehicle Body HBACK	-1.3128	0.4500	-2.92	$3.5 \cdot 10^{-3}$
Vehicle Body HDTOP	-1.3333	0.4732	-2.82	$4.8 \cdot 10^{-3}$
Vehicle Body MCARA	-0.6046	0.5868	-1.03	0.30
Vehicle Body MIBUS	-1.8646	0.5478	-3.40	$6.7 \cdot 10^{-4}$
Vehicle Body PANVN	-1.1877	0.4901	-2.42	0.02
Vehicle Body SEDAN	-1.2928	0.4497	-2.87	$4.0 \cdot 10^{-3}$
Vehicle Body STNWG	-1.3739	0.4505	-3.05	$2.3 \cdot 10^{-3}$
Vehicle Body TRUCK	-1.4576	0.4726	-3.08	$2.0 \cdot 10^{-3}$
Vehicle Body UTE	-1.4521	0.4579	-3.17	$1.5 \cdot 10^{-3}$
Age Band	-0.0979	0.0175	-5.59	$2.2 \cdot 10^{-8}$

Table 19: Regression Parameter Estimates - Frequency GAM

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	-0.2123	0.4558	-0.47	0.64
Vehicle Body COUPE	-0.8326	0.4977	-1.67	0.09
Vehicle Body HBACK	-1.2818	0.4563	-2.81	$5.0 \cdot 10^{-3}$
Vehicle Body HDTOP	-1.3823	0.4793	-2.88	$4.0 \cdot 10^{-3}$
Vehicle Body MCARA	-0.6921	0.5930	-1.17	0.24
Vehicle Body MIBUS	-1.9032	0.5542	-3.43	$6.0 \cdot 10^{-4}$
Vehicle Body PANVN	-1.1863	0.4953	-2.40	0.02
Vehicle Body SEDAN	-1.2803	0.4556	-2.81	$5.0 \cdot 10^{-3}$
Vehicle Body STNWG	-1.4266	0.4556	-3.13	$1.7 \cdot 10^{-3}$
Vehicle Body TRUCK	-1.5070	0.4781	-3.15	$1.6 \cdot 10^{-3}$
Vehicle Body UTE	-1.4924	0.4204	-3.55	$3.9 \cdot 10^{-4}$
Age Band	-0.0949	0.0177	-5.36	$8.3 \cdot 10^{-8}$

Table 20: Adjusted R^2 and AIC for the Frequency Models

	Adjusted R^2	**AIC**
GLM	1.77 %	11867.00
GAM	1.85 %	11854.96

to reject that smoothing of the ‘Vehicle value’ predictor is not necessary under both the GLM and GAM structure; however, the ‘Vehicle value’ predictor is not significant in the independent severity models, and the AIC criterion as well as residual plots are in favor of the independent and dependent severity GAM over the parametric approach. The number of claims effect are estimated to be 0.2451 and 0.2836 for the dependent GAM and GLM severity model, respectively, which are both larger than 0 and close to what we found in the training dataset. Table 25 shows the AIC values for the frequency and conditional severity models, respectively. Similar to the results obtained from the training dataset, the GAM AIC is lower than the corresponding GLM AIC and the dependent model’s AIC is lower than the independent model’s AIC.

The residual plots for the test set were also very similar to those obtained from the training set (but we omit the plots here for the sake of space).

Table 21: Regression Parameter Estimates - Independent GLM Severity Model

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	7.7127	0.1536	50.21	$< 2 \cdot 10^{-16}$
Vehicle value	-0.0559	0.0373	-1.50	0.13
Gender Male	0.2283	0.0872	-2.62	0.01
Area B	0.0363	0.1277	0.28	0.78
Area C	0.1137	0.1167	0.97	0.33
Area D	-0.1360	0.1573	-0.86	0.39
Area E	0.1065	0.1779	0.60	0.55
Area F	0.4131	0.2006	2.06	0.04
Age Band	-0.0576	0.0296	-1.95	0.05

Table 22: Regression Parameter Estimates - Dependent GLM Severity Model

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	7.4428	0.2380	31.27	$< 2 \cdot 10^{-16}$
Vehicle value	-0.0563	0.0376	-1.50	0.13
Gender Male	0.2250	0.0881	2.55	0.01
Area B	0.0487	0.1292	0.38	0.71
Area C	0.1187	0.1180	1.01	0.31
Area D	-0.1232	0.1590	-0.77	0.44
Area E	0.1075	0.1798	0.60	0.55
Area F	0.4054	0.2027	2.00	0.05
Age Band	-0.0564	0.0299	-1.89	0.06
Number of Claims	0.2451	0.1651	1.48	0.14

Table 23: Regression Parameter Estimates - Independent GAM Severity Model

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	7.5493	0.1363	55.39	$< 2 \cdot 10^{-16}$
Gender Male	0.2250	0.0868	2.59	0.01
Area B	0.0892	0.1273	0.70	0.48
Area C	0.1503	0.1164	1.29	0.20
Area D	-0.1000	0.1574	-0.64	0.53
Area E	0.1641	0.1778	0.92	0.35
Area F	0.5617	0.2020	2.78	0.01
Age Band	-0.0562	0.0296	-1.90	0.06

In sum, compared with GLMs, GAMs can provide a more flexible approach, capable of dealing with the more complicated trend structure of the data. The new modelling approach of GAMs for aggregate claims under dependence improved the model fit when the data had nonparametric trends of the continuous rating predictors with an attendant association or dependence between severity and frequency components.

Table 24: Regression Parameter Estimates - Dependent GAM Severity Model

Regression Parameter	**Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	7.2323	0.2272	31.83	$< 2 \cdot 10^{-16}$
Gender Male	0.2212	0.0876	2.53	0.01
Area B	0.1075	0.1286	0.84	0.40
Area C	0.1597	0.1175	1.36	0.17
Area D	-0.0809	0.1589	-0.51	0.61
Area E	0.1725	0.1795	0.96	0.34
Area F	0.5621	0.2040	2.76	0.01
Age Band	-0.0549	0.0298	-1.84	0.07
Number of Claims	0.2836	0.1644	1.73	0.08

Table 25: Results of Severity Model Comparisons

Model	**AIC**
Independent Severity GLM	26593.00
Dependent Severity GLM	26589.00
Independent Severity GAM	26576.37
Dependent Severity GAM	26570.16

Table 26: Approximating versus Bootstrapping p-values

Model	F**-value**	**Approximate** p**-value**	**Bootstrap** p**-value**
Independent Severity Models	1.8196	0.04409	0.420
Dependent Severity Models	1.7678	0.05138	0.467

6 Conclusion

This paper extends the framework of GLMs to GAMs for aggregate claims under both the independent and dependent settings, which relaxes the restriction of GLMs on the transformed mean of the response variable, thereby providing for the possibility of a much better model fit. This new modelling approach, namely, *GAMs for Aggregate Claims Under Dependence*, combines the advantages of allowing for dependence between claim frequency and severity with the nonparametric trend terms in the model fits, which can help actuaries better quantify the insurance risk.

In this research, frequentist GAMs based on cubic penalized regression splines were first introduced to predict pure premiums of property and casualty insurance data. Compared with ordinary regression splines, the penalized version is easier to implement, as the wiggliness penalty term will take care of the model selection process and control the amount of smoothing of the model fittings. Consequently, the selection of location and number of knots do not play an overly disproportionate role in determining the model fittings anymore, as long as the maximum basis dimension is chosen to be sufficiently large. In addition, it combines the advantages inherent in both the parametric and nonparametric approaches. Not only can the inference of parametric models be applied, but it also has an explicit local nature allowing for more flexible model fits.

The simulation study validated the new modelling techniques of GAM fittings with cubic penalized regression splines for aggregate claims. By introducing pre-specified dependent and nonlinear terms in the aggregate claims model, the effects were illustrated and compared by different model fits. The case study further applied this new approach to a one-year vehicle claims dataset and a hold-out dataset was used to test and validate the model fittings derived from the training set. Both the simulation and case study results suggest that the GAMs performed better than their GLM counterparts on estimating the pure premiums, and the dependent approach provided a better fit than the corresponding independent models under the same GLM or GAM structure. In summary, the new proposed approach, when compared with the parametric and/or independent approach, provided a more accurate representation of the insurance data considered and ultimately lead to a more precise estimate of the pure premium.

7 Future Work

The GAMs in this paper were only fit by cubic penalized regression splines in a univariate smoothing form. Other nonparametric smoothing methods and bivariate smoothing techniques might also be considered and compared with this approach for some more complicated cases. A more thorough and comprehensive simulation study that explores more nonlinear functions and more of the model parameter space would be worth pursuing. Further study of performance properties, such as coverage, length, etc. of the Bayesian credible intervals would also be opportune.

8 Appendix A

If the individual claim size, conditional on the number of claims $Y_{ij}|N_i$, has a gamma distribution with mean μ and dispersion parameter ϕ, then the average claim amount $\bar{Y}|N_i$ also follows a gamma distribution with mean μ and dispersion ϕ/N_i (Song 2007). So, in this dependent setting, $\bar{Y}_i|N_i$ follows a gamma distribution with mean μ_{i2}^D and dispersion $\phi_{2i} = \phi'/n_i$, where $Y_{ij}|N_i \sim (\mu_{i2}^D, \phi')$. Then the variance of the aggregate claims under the auspices of dependence is as follows:

$$\begin{aligned}
Var(S_i|\boldsymbol{X}_i) &= Var(E[S_i|\boldsymbol{X}_i, N_i]|\boldsymbol{X}_i) + E[Var(S_i|\boldsymbol{X}_i, N_i)|\boldsymbol{X}_i] \\
&= Var(E[N_i\bar{Y}_i|\boldsymbol{X}_i, N_i]|\boldsymbol{X}_i) + E[Var(N_i\bar{Y}_i|\boldsymbol{X}_i, N_i)|\boldsymbol{X}_i] \\
&= Var(N_iE[\bar{Y}_i|\boldsymbol{X}_i, N_i]|\boldsymbol{X}_i) + E[N_i^2 Var(\bar{Y}_i|\boldsymbol{X}_i, N_i)|\boldsymbol{X}_i] \\
&= Var(N_i\mu_{i2}^D|\boldsymbol{X}_i) + E[N_i^2\phi_{2i}(\mu_{i2}^D)^2|\boldsymbol{X}_i] \\
&= Var(N_i\exp(\tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2 + N_i\beta_N)|\boldsymbol{X}_i) + E[N_i\phi'\exp(2\tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2 + 2N_i\beta_N)|\boldsymbol{X}_i] \\
&= \exp(2\tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2)Var(N_i\exp(N_i\beta_N)|\boldsymbol{X}_i) + \phi'\exp(2\tilde{\boldsymbol{X}}_{i2}^\top\tilde{\boldsymbol{\beta}}_2)E[N_i\exp(2N_i\beta_N)|\boldsymbol{X}_i] \qquad (45) \\
&= (\tilde{\mu}_{i2})^2\Big\{\mu_{i1}^2\exp\{\mu_{i1}(e^{2\beta_N}-1)+4\beta_N\} + \mu_{i1}\exp\{\mu_{i1}(e^{2\beta_N}-1)+2\beta_N\} \\
&\quad -\mu_{i1}^2\exp\{2\mu_{i1}(e^{\beta_N}-1)+2\beta_N\}\Big\} + \phi'(\tilde{\mu}_{i2})^2\mu_{i1}\exp\{\mu_{i1}(e^{2\beta_N}-1)+2\beta_N\} \\
&= \mu_{i1}(\tilde{\mu}_{i2})^2\Big\{\mu_{i1}\exp\{\mu_{i1}(e^{2\beta_N}-1)+4\beta_N\}+ \\
&\quad (\phi'+1)\exp\{\mu_{i1}(e^{2\beta_N}-1)+2\beta_N\} - \mu_{i1}\exp\{2\mu_{i1}(e^{\beta_N}-1)+2\beta_N\}\Big\},
\end{aligned}$$

which was originally derived by Schulz (2013).

References

Antonio, K. & Beirlant, J. (2008), 'Issues in claims reserving and credibility: A semiparametric approach with mixed models', *Journal of Risk and Insurance* **75**(3), 643–676.

Boucher, J., Côté, S. & Guillen, M. (2017), 'Exposure as duration and distance in telematics motor insurance using generalized additive models', *Risks* **5**(4), 54.

Czado, C., Kastenmeier, R., Brechmann, E. & Min, A. (2012), 'A mixed copula model for insurance claims and claim sizes', *Scandinavian Actuarial Journal* (4), 278–305.
URL: *https://doi.org/10.1080/03461238.2010.546147*

Denuit, M. & Lang, S. (2004), 'Nonlife ratemaking with Bayesian GAMs', *Insurance: Mathematics and Economics* **35**(3), 627–647.

Dionne, G., Gouriéroux, C. & Vanasse, C. (2001), 'Testing for evidence of adverse selection in the automobile insurance market: A comment', *Journal of Political Economy* **109**(2), 444–453.

Eilers, P. & Marx, B. (1996), 'Flexible smoothing with B-splines and penalties', *Statistical Science.* **11**(2), 89–121.
URL: *https://doi.org/10.1214/ss/1038425655*

Faraway, J. (2016), *Extending the linear model with R*, Chapman & Hall/CRC Texts in Statistical Science Series, CRC Press, Boca Raton, FL.

Garrido, J., Genest, C. & Schulz, J. (2016), 'Generalized linear models for dependent frequency and severity of insurance claims', *Insurance: Mathematics and Economics* **70**, 205–215.
URL: *https://doi.org/10.1016/j.insmatheco.2016.06.006*

Goldburd, M., Khare, A. & Tevet, D. (2016), *Generalized linear models for insurance rating*, Vol. 5 of *CAS Monograph Series*, Casualty Actuarial Society.

Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, Ltd., London.

Jong, P. & Heller, G. Z. (2008), *Generalized linear models for insurance data*, Vol. 5 of *International Series on Actuarial Science*, Cambridge University Press.

Jørgensen, B. & Paes de Souza, M. (1994), 'Fitting Tweedie's compound Poisson model to insurance claims data', *Scandinavian Actuarial Journal* (1), 69–93.
URL: *https://doi.org/10.1080/03461238.1994.10413930*

Klein, N., Denuit, M., Lang, S. & Kneib, T. (2014), 'Non-life rate-making and risk management with Bayesian generalized additive models for location, scale, and shape', *Insurance: Mathematics and Economics* **55**, 225–249.

Klein, N., Kneib, T. & Lang, S. (2015), 'Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data', *Journal of the American Statistical Association* **110**(509), 405–419.

Klugman, S., Panjer, H. & Willmot, G. (2012), *Loss models*, Wiley Series in Probability and Statistics, fourth edn, John Wiley & Sons, Inc., Hoboken, NJ; Society of Actuaries, Schaumburg, IL.

McCullagh, P. & Nelder, J. (1989), *Generalized linear models*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.
URL: *https://doi.org/10.1007/978-1-4899-3242-6*

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

Quijano, X., Oscar, A. & Garrido, J. (2015), 'Generalised linear models for aggregate claims: to Tweedie or not?', *European Actuarial Journal* **5**(1), 181–202.
URL: *https://doi.org/10.1007/s13385-015-0108-5*

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003), *Semiparametric regression*, Vol. 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge.
URL: *https://doi.org/10.1017/CBO9780511755453*

Schilling, E. G. & Neubauer, D. V. (2009), *Acceptance sampling in quality control*, Statistics: Textbooks and Monographs, second edn, CRC Press, Boca Raton, FL.
URL: *https://doi.org/10.1201/9781584889533*

Schulz, J. (2013), 'Generalized linear models for a dependent aggregate claims model', *Master's thesis, Concordia University, Montreal, Canada* .

Shi, P., Feng, X. & Ivantsova, A. (2015), 'Dependent frequency-severity modeling of insurance claims', *Insurance Math. Econom.* **64**, 417–428.

Smyth, G. & Jørgensen, B. (2002), 'Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling', *Astin Bulletin* **32**(1), 143–157.
URL: *https://doi.org/10.2143/AST.32.1.1020*

Song, P. (2007), *Correlated data analysis: modeling, analytics, and applications*, Springer Series in Statistics, Springer, New York.

Stasinopoulos, D. & Rigby, R. (2007), 'Generalized additive models for location scale and shape (GAMLSS) in R', *Journal of Statistical Software* **23**(7), 1–46.

Verrall, R. (1996), 'Claims reserving and generalised additive models', *Insurance: Mathematics and Economics* (19), 31–43.

Werner, G. & Modlin, C. (2016), *Basic ratemaking*, fifth edn, Casualty Actuarial Society.

Wood, S. (2017), *Generalized additive models*, Texts in Statistical Science Series, second edn, CRC Press, Boca Raton, FL.

Yang, Y., Qian, W. & Zou, H. (2018), 'Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models', *Journal of Business and Economic Statistics* **36**(3), 456–470.