Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 7, No. 1, Winter & Spring 2027
Research paper

Joint Partially Models for Dependent Frequency and Severity of Insurance Claims with Using Spline Functions

Ehsan Bahrami Samani[1]

[1] Department of Statistics, Shahid Beheshti University, Tehran, Iran.
e_bahrami@sbu.ac.ir

Abstract:
In this paper, the claim counts and claim amounts, both with and without missing values, are assumed to be correlated in the context of non-life insurance. The proposed methodology involves fitting joint random effects partially specified models based on the factorization of the joint distribution of claim counts and average claim amounts. Furthermore, the paper introduces joint partially aggregate claims models that account for correlated claim count and amount responses, addressing cases with missing values in both variables and incorporating nonignorable missing data mechanisms. To capture non-linear time effects on the joint model, various spline functions are employed. A full likelihood-based estimation procedure is adopted to obtain maximum likelihood estimates for the parameters of the joint partial models involving correlated claim frequencies and severities. For data scenarios where both variables contain nonignorable missing values, a corresponding pure premium formula is derived. The performance of the proposed model is compared to that of the independent aggregate claims model through a simulation study, focusing on differences in the resulting pure premiums. To assess model sensitivity, influence analysis is conducted by examining how small perturbations in the average percent difference (APD) affect likelihood displacement using influence graphs. Finally, the methodology is applied to a real Iranian automobile insurance dataset.

Keywords: Partially aggregate claims model, Claim frequency, Claim severity, Missing Data, Spline functions.
Classifications: 91G05; 62P05; 62J12; 62D10; 65D07.

1 Introduction

In the insurance research, one of the topics studied is the analysis of claim count and amounts variables and analyzing the effect of explanatory variables on them. For example, many researchers such as [26], [21], [11], [4], [25], and [12] can be mentioned. In these researches, generalized linear models (GLMs) have been used. The GLMs techniques are used to estimate the pure premium based on the observable characteristics of policyholders. In GLMs some monotone transformation of the means of the claim count and amounts variables and analyzing the effect of explanatory variables on them can then be expressed, where linear function of the explanatory variables such as age, sex, etc are the linear and multiplicative models

[1]Corresponding author
Received: 30/03/2026 Accepted: 29/06/2026
https://doi.org/10.22054/jmmf.2026.92058.1274

as special cases. All these researches propose an aggregate claims model to relax the independence assumption between the counts and the amounts of insurance claims without missing data. Also, separate analyses of the claim counts and amounts may give biased estimates of the parameters, and may result in misleading inferences. Consequently, we need to consider a method in which these variables can be modelled jointly. Also, [13] proposed the dependent aggregate claims model. They consider here a conditional approach in which the generalized linear model (GLM) for the component of the aggregate claims models is allowed to depend on the claim count. The simultaneous analysis of the explanatory variables on mixed responses creates accurate, unbiased and consistent results in contrast to that of separate analysis of responses. Models which considered responses simultaneously have been introduced by many researchers, for example [26], [3] and [19]. Different models can be used to take into account the correlation between responses raised due to the nature of the correlated data. A possibility is random effects modeling, which makes inferences about variability between respondents ([1, 17, 31, 32]). Various studies have been conducted using random effects in the analysis and modeling of insurance data. Paper [14] considers regression models for spatially indexed data, and allow for underlying spatial dependency patterns by including correlated spatial random effects. Again the dependence is defined through copulas here. Shi and Ahn define a bivariate (frequency/severity) random effects model for a Bonus-Malus system and use copulas to introduce the dependence [24]. Jeong et al. include random effects, including through generalized linear mixed models (GLMMs), a common way used by actuaries to add random effects to GLMs [20]. Cheung et al. use Bayesian credibility, another common way actuaries introduce random effects in aggregate claims models [2].

Missing values in data sets can arise for several reasons: 1) Human errors during data entry can lead to missing values. This could be due to incorrect filling of forms or miscommunication. 2) In surveys or questionnaires, respondents may skip certain questions or refuse to answer them, leading to missing values for those specific fields. 3) Technical issues or oversights during data collection can cause incomplete information. 4) Evolving definitions of variables over long periods may lead to some not being recorded. 5) Merging data from different sources can create mismatches and gaps. 6) Sensitive data may be omitted to protect individual privacy. 7) Some events, like minor injuries not requiring attention, may inherently lack data. 8) In large data sets, some entries may unintentionally be left blank. 9) Certain variables might not apply to all records, causing missing information. 10) Organizational shifts can temporarily disrupt data collection.
Addressing these missing values is vital for proper data cleaning and preprocessing, as they impact analysis and insights [34].

Missing data are a common challenge in insurance datasets. Some reasons for the missing data in insurance include the following:
1) Multiple sales channels, customers refusing to disclose data, and offices not

submitting the relevant data into their databases. 2) when the date of diagnosis is missing in critical illness insurance it is impossible to include these claims in diagnosis inception rate calculations. 3) Sometimes the amount of insurance data is limited and insurance companies have difficulty determining premium rates and related financial discussions. For example, in the model incidence rates for older ages, there are not very many insured people at older ages. In these cases population statistics can be used as another information source. The information about the insured persons in older ages can be considered as a kind of missing data, which according to the relationship of this statistic with the population statistics, this missing data can be surrogated. 4) In the insurance applicant's historical data, there will be the possibility of missing data. 5) Incomplete data in insurance is a common problem and prevalent in basically all data sets. This missing data are due to the non-recording of some information of policyholders. 6) Insurance data sets are often truncated for multiple reasons, such as deductibles and total loss limits. So, in this case, data is missing. 7) Some non-life insurance data include correlated claim count and amounts variables with missing data.

In such insurance studies, the claim count and amounts in non-life insurance, often the subjects do not respond in occasions which cause for missing responses. For example, the people claiming bit amounts might be more likely not to report their claims, or some claims may be more likely to be unreported depending on the cause of claim. 8) The personal legal insurance, also called group legal services insurance or prepaid legal services, is designed to make legal services more affordable. In this type of insurance, data observed a large number of claims under choice where claims adjusters did not record the legal option to which the claim was subject.

Missing data affect the quality of analyses, as well as the base and accuracy of estimates. The most common approach to deal with this problem in the literature is to exclude observations with missing information from analyses, typically referred to as a complete case analysis. There are often two kinds of missing pattern considered for insurance studies; one is the monotone pattern that is most common in medical surveys and is caused by subject's dropout before the study is completed, and the other is the non-monotone or general pattern that occurs when some of the subjects withdraw from the study at some occasion and return to the study again at some other occasion. Rubin, and Little and Schluchter made important distinctions between the various types of missing mechanisms for each of the above - mentioned patterns [23, 27]. They define the missing mechanism as missing completely at random (MCAR) if missingness is dependent neither on the observed responses nor on the missing responses, and missing at random (MAR) if, given the observed responses, it is not dependent on the missing responses (assuming that the parameters of missing mechanism and responses are distinct). Missingness is defined as not missing at random (NMAR) if it depends on the unobserved responses. From a likelihood point of view MCAR and MAR are ignorable, but NMAR is non-ignorable. For example, suppose there are two variables, the claim count and

the claim amounts in non-life insurance, where the claim amounts is subject to nonresponse. If the probability that the claim amounts is recorded (non-missing) is the same for all individuals, regardless of their claim count and amounts, then the data are MCAR. This is a very restrictive case of missingness, which requires both that missing data are MAR and that observed data are observed at random. If, however, the probability that the claim amounts is recorded varies according to the claim amounts of the respondent but does not vary according to the claim count of respondents within an age group, then the data are MAR, but not MCAR. In other words, the probability of missing data on the claim count depends on a person's age, but within each age group, the probability of missing the claim amounts is unrelated to the claim amounts. The last case of missingness calls for special attention in empirical work. That is, when the probability that the claim amounts is recorded varies according to the claim amounts within each age group, then the data are NMAR.

Derrig and Francis presented an early example of a comparison between tree-based methods and GLMs (for logistic regression) applied to insurance claims with missing values [6]. Also, Guelman and Guillén discuss a special case of missing value problem for auto insurance data [15]. Wang et al. define GLMs for claim counts and generalized linear mixed model (GLMMs) with missing claim count covariate for claim amounts, so in a way this is a form incorporating both, random effects and claim counts. In all the researches mentioned, the analyses are based on missing at random mechanism (MAR) [33].

The aim of this paper is to propose an approach for the joint partially aggregate claims model by using random effects with nonignorable missing data. Random effects are also introduced to take into account the correlation between claim count and amounts variables in different occasions. The joint partially model by using random effects with missing data is described in terms of a correlated the claim count and amounts variables. We consider the use of different spline functions for example The truncated power basis function (TBP), cubic splines and B-spline, to explore the non-linear effects of time on the joint partially aggregate claims model.

The Poisson distribution is a classical distribution to model count data. In real data, the relationship between the mean and variance implies a dispersion index, if the variance is greater than the mean, then the dispersion index is greater to 1. Over-dispersion relative to the Poisson distribution (i.e. where the variance is greater than the mean) is a common feature among real data. One distribution that allows for over dispersion is the negative binomial distribution. Although the negative binomial distribution is able to handle over dispersion, it is not in the case of under dispersion. A flexible distribution capable of modeling both under-dispersed and over-dispersed data. The COM-Poisson distribution is a flexible distribution for count data that allows for over- or under-dispersion, (via [28, 29]). In particular, it will be seen that when Conway-Maxwell Poisson counts are assumed and a log-link is used in the conditional random effects model, the pure premium resulting from

our general approach can be written as the product of three terms, i.e., the marginal mean frequency given random effect, a modified marginal mean severity given random effect, and a dependence correction term. Also, the effect of missing values on rate and premium calculations can also be investigated using appropriately the joint aggregate claims model by using random effects. This would give us a better understanding of the risk associated with relevant insurance products.

The rest of the paper is organized as follows. In Section 2, distributions for count data are presented. The models, likelihoods and estimation are presented by Section 3. In Section 4, the joint partially aggregate claims model with and without missing data are given. In Section 5, sensitivity analysis for the average percent difference (APD) is presented. Section 6, simulation studies are used to assure the performance of the model for a given data and in Section 7, the models are used for analyzing insurance data. Finally, in Section 8, the paper concludes with some remarks.

2 Distributions for Count Data

In this section, we shall review the distributions of count data that most widely used in the analysis of count data. The general form of the probability mass function (pmf) of a random variable Y of a standard power series (PS) distribution with the parameter $\theta > 0$ is given by

$$P(Y = y) = \frac{a(y)\,\theta^y}{g(\theta)}, \qquad y = 0, 1, \cdots \quad , \quad \theta > 0, \tag{1}$$

where $a(y) \geq 0$ and $g(\theta)$ is the normalizing constant. The mean and variance of Y are

$$E(Y) = \theta \frac{g^{(1)}(\theta)}{g(\theta)},$$
$$Var(Y) = \theta^2 [\frac{g^{(2)}(\theta)}{g(\theta)} - (\frac{g^{(1)}(\theta)}{g(\theta)})^2],$$

where $g^{(k)}$ is the k the derivative of g. The Poisson (P) and negative binomial (NB) distributions belong to this class. For the Poisson distribution with **the parameter** θ (Poisson(θ)), we have $g(\theta) = \exp(\theta)$ and $a(n) = \frac{1}{n!}$. Also, for the negative binomial distribution with parameters μ and σ (NB(μ,σ)), we have $\theta = \frac{\mu}{\mu+\sigma}$, $g(\theta) = (1-\theta)^\sigma$ and $a(n) = \frac{\Gamma(n+\sigma)}{\Gamma(\sigma)\Gamma(n+1)}$ (for known σ). The general form of the probability mass function (pmf) of a random variable Y of a Conway-Maxwell Poisson (CMP) distribution with the parameters θ and ν is defined as:

$$P(Y = y|\theta, \nu) = \frac{1}{h(\theta, \nu)} \frac{\theta^y}{y!^\nu}, \tag{2}$$

where $\lambda = E(Y^\nu) > 0$ is a shape parameter, $\nu > 0$ is a dispersion parameter and $h(\theta, \nu) = \sum_{y=0}^{\infty} \frac{\theta^y}{y!^\nu}$ normalizes the distribution. If $\nu = 1$ a CMP distribution is the exactly same as a Poisson distribution which indicates there is no additional dispersion. If $\nu > 1$ ($\nu < 1$) represents, then under dispersion (over dispersion) is obtained.

3 Models, Likelihoods and Estimation

In this Section, we proposed the joint partially linear models for the claim count and amounts variables and analyzing the effect of explanatory variables on them. Partially linear models(PLMs) are regression models in which the response depends on some explanatory variables linearly but on other covariates non-linearly and non-parametrically(via [7, 18, 30]).

3.1 The Joint Partially Linear Model

We structure the model hierarchically to clearly separate the observation-level distributions from the random-effects dependence structure.

Hierarchical Model Specification

Let Y_i denote the claim count and $\bar{Z}_i$ the average claim amount for policyholder i $(i = 1, \ldots, n)$. Let $X_i^{(q)}$, $W_i^{(q)}$, and $V_i^{(q)}$ be sub-vectors of covariates for $q = 1, 2$, where $q = 1$ corresponds to the count model and $q = 2$ to the severity model.

We assume that the distribution of Y_i given b_i is a power series distribution with the probability mass function (1) where a Generalized partially linear model is considered as follows:

$$[Y_i|b_i] \sim PS(\theta_i), \quad \log(E(Y_i|b_i)) = X_i^{(1)'}\beta + W_i^{(1)'}b_i + k_1(V_i^{(1)})$$

where β is the vectors of regression coefficients and b_i is random effect. Here, $k_1(\cdot)$ is an unknown smooth function capturing the non-linear effect of $V_i^{(1)}$ on the claim count. Full details on knot selection, smoothing, and asymptotics are provided in Section 3.1.

Also, let $\bar{Z}_i = \frac{\sum_{j=1}^{Y_i} Z_{ij}}{Y_i}$ denote the average claim amount. The aggregate claims can then be written as $S_i = Y_i\bar{Z}_i$ with the convention that $S_i = 0$ when $Y_i = 0$. We assume that the distribution Z_{ij} given u_i is in the case of the reproductive exponential dispersion family (EDF) where a EDF regression $(EDF(\mu_{ij}, \sigma^2))$ model as

$$[\bar{Z}_i|u_i] \sim EDF(\mu_i, \sigma^2/Y_i), \quad \log E(\bar{Z}_i|u_i) = X_i^{(2)'}\alpha + W_i^{(2)'}u_i + k_2(V_i^{(2)}),$$

where $\mu_i = E(\bar{Z}_i|b_i)$ and α is the vectors of regression coefficients and and u_i is random effect. Here, $k_2(\cdot)$ is an unknown smooth function capturing the non-linear effect of $V_i^{(2)}$ on the claim count. Full details on knot selection, smoothing, and asymptotics are provided in Section 3.1.

To model the claim count and amounts variables, one may use random effects to take into account the correlation between responses across time, which leads to the conditional independence of the vector of responses in different occasions given

subject-level effects b_i and u_i where

$$R_i = (b_i, u_i) \overset{iid}{\sim} MVNormal(0, \Sigma_{bu}),$$
$$\Sigma_{bu} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = Var(b_i)$, $\Sigma_{22} = Var(u_i)$ and $\Sigma_{12} = \Sigma'_{21} = Cov(b_i, u_i)$. Also, $X_i^{(q)}$ is a covariate vector for $q = 1, 2$, $W_i^{(q)}$ and $V_i^{(q)}$ be some sub-vectors of covariate of $X_i^{(q)}$. The parameter vector to be estimated is $\Theta = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \text{vech}(\Sigma_{bu})^\top, \sigma^2)^\top$. The magnitude of the off-diagonal block Σ_{12} captures the cross-dependence between the claim count and severity processes. If $\Sigma_{12} = \mathbf{0}$, the two random effects are uncorrelated, and consequently, the two responses are conditionally independent given the covariates. Note that this corrects the condition for conditional independence; the diagonal specification previously stated was a typographical error.

Model

The joint par The joint partially model for the two responses by using random effects is written as:

$$Y_i|b_i \sim PS(\theta_i), \qquad \log(E(Y_i|b_i)) = X_i^{(1)'}\beta + W_i^{(1)'}b_i + k_1(V_i^{(1)}) \quad (Model\ 1)$$
$$\bar{Z}_i|u_i \sim EDF(\mu_i, \sigma^2/Y_i), \quad logE(\bar{Z}_i|u_i) = X_i^{(2)'}\alpha + W_i^{(2)'}u_i + k_2(V_i^{(2)})$$

where

$$R_i = (b_i, u_i) \overset{iid}{\sim} MVNormal(0, \Sigma_{bu}),$$
$$\Sigma_{bu} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Large values of $\Sigma_{12} = \Sigma'_{21} = Cov(b_i, u_i)$ indicate a strong correlation between the two responses. The vector of parameters α and β, the parameters of Σ_{bu} should be estimated. If $\Sigma_{12} = diagonal\{1, ..., 1\}$, then the two responses are independent given the covariates.

Example 1. Suppose that $Y_i \sim Poisson(\theta_i)$ and that, the individual claim amounts $\bar{Z}_i$ have a Gamma distribution with mean μ_i and dispersion $\frac{\sigma}{Y_i}$. It then follows from the model (1) that

$$Y_i \sim Poisson(\theta_i), \qquad log(\theta_i) = \beta_0 + \beta_1 X_i^{(1)} + b_i + k_1(V_i^{(1)})$$
$$\bar{Z}_i \sim Gamma(\mu_i, \tfrac{\sigma}{Y_i}), \quad log(\mu_i) = \alpha_0 + \alpha_1 X_i^{(2)} + u_i + k_2(V_i^{(2)})$$

where

$$R_i = (b_i, u_i) \overset{iid}{\sim} MVNormal(0, \Sigma_{bu}),$$
$$\Sigma_{bu} = \begin{pmatrix} \sigma_b^2 & \sigma_b\sigma_u\rho_{bu} \\ \sigma_b\sigma_u\rho_{bu} & \sigma_u^2 \end{pmatrix},$$

and σ_b^2 (σ_u^2) is the variance of random effect b_i (u_i) and ρ_{bu} is the dependence parameters of the model(1).

Non-linear Components via Splines

In this model, $X_i^{(q)}$ has a linear relationship on the response but it is possible in many applications, a non-linear relationship has on the Y_i. This non-linear relationship is denoted by $k_q(t)$, $q = 1, 2$, so that $k_q(t)$ where $t = V_i^{(q)}$, is unknown and it should be estimated using some semi-parametric approaches such as spline functions. A spline function $k_q(t)$ with a set of knots sequence $\tau_{q1} < ... < \tau_{qL_q}$ and fixed the polynomial of d_q can be given by

$$k_q(t) = \sum_{l=1}^{L_q+d_q+1} \eta_{ql} B_{ql}(t),$$

where $B_{ql}(t)$ is basis function defining a vector space and η_{ql} is the associated spline coefficient. There are numerous options for the definition of the basis functions $B_{ql}(t)$ such as the truncated power series basis, the B-spline basis, cubic spline and the cardinal spline basis, (For more details, see [5]). The truncated power series basis is defined by the basis functions $B_{q1}(t) = 1$, $B_{q2}(t) = t$, ..., $B_{qd_{q+1}}(t) = t^{d_q}$, $B_{qd_{q+2}}(t) = (t - \tau_{q1})_+^{d_q}$,..., $B_{qd_{L+q+1}}(t) = (t - \tau_{qL})_+^{d_q}$, where $t_+ = tI_{(t>0)}$ and $I(.)$ is indicator function. For example, if $d_q = 3$ then the truncated power basis is a cubic spline function, with three knots (τ_1, τ_2, τ_3). Using representation given in

$$k(t) = \eta_0 + \eta_1 t + \eta_2 t^2 + \eta_3 t^3 + \eta_4 (t - \tau_1)^3 + \eta_5 (t - \tau_2)^3 + \eta_6 (t - \tau_3)^3.$$

The B-spline basis functions of degree d_q is given by

$$B_{ql}^0(t) = I_{(a_{ql} \le t < a_{q,l+1})},\ B_{ql}^{d_q}(t) = \frac{t - a_{ql}}{a_{q,l+d_q} - a_{ql}} B_{ql}^{d_q-1}(t) - \frac{a_{q,l+d_q+1} - t}{a_{q,l+d_q+1} - a_{q,l+1}} B_{q,l+1}^{d_q-1}(t),$$

where $l = 1, ..., L + d_q + 1$ and $a_1 \le ... \le a_{d_q} < ... < a_{d_q+L+1} < ... < a_{2d_q+L+2}$ is the knot sequence.

Likelihood Function

We proposed a model for a correlated the claim count and the average claim amount variables based on the factorization of the joint distribution of the two responses. The factorization of the joint distribution of Y and $\bar{Z}$ can be considered:

$$f(y, \bar{z}|b, u) = f(y|b) f(\bar{z}|u),$$

where $f(y, \bar{z}|b, u)$, $f(\bar{z}\,|u)$ and $f(y\,|u)$ denote the corresponding joint density function for Y and $\bar{Z}$, the corresponding density function for $\bar{z}$ given b and the density function for y given u, respectively.

Maximum likelihood estimates of the parameters can be obtained with commonly used algorithms for maximizing the likelihood. The log-likelihood function under the factorization model with random effects model (1) is

$$\log(L) = \sum_{i=1}^{n} \log(\int_{b_i} \int_{u_i} f(\bar{Z}_i\,|u_i)\, f(Y_i|b_i)\, \varphi_{12}(b_i, u_i; \Sigma_{bu})\, db_i du_i),$$

where $f(\bar{Z}_i\,|u_i)$, $f(Y_i\,|u_i)$ and $\varphi_{12}(.;.)$ denote the corresponding density function for $\bar{Z}_i$ given b_i, the density function for Y_i given u_i and the corresponding joint density function for b_i and u_i, respectively.

Estimation

A popular approach for this purpose is the use of penalized spline. The smoothing spline of degrees $2d_q-1$ is given by choosing $k_q(t)$ to minimize the following criterion

$$\log(L)+\delta_1\int(\frac{\partial^{d_1}}{\partial t^{d_1}}k_1(t))^2dt+\delta_2\int(\frac{\partial^{d_2}}{\partial t^{d_2}}k_2(t))^2dt,$$

where $\log(L)$ is the logarithm of the likelihood and $\frac{\partial^{d_q}}{\partial t^{d_q}}k_q(t)$ is the d_qth derivative of $k_q(t)$ and the value of d_q is often chosen to be 1 or 2 which lead to linear and cubic splines, respectively. Also, δ_1 and δ_2 are controlling the amount of smoothness. Some popular strategies to determine the optimal values of them are the use of generalized cross validation, AIC. The basic idea of AIC is to correct the log-likelihood of a fitted model for the effective number of parameters, (via [8]).

In this approach, we define the knots, the degree and adding of the basis functions as some new covariates to model (1).

The log-likelihood for model (2), which has been given in appendix **B** (Subsection B1). As the likelihood may be flat (the problem of using a selection model for missing mechanism), for finding the parameter estimates, our suggestion is to use the results of the parameter estimates of the independent univariate analysis as initial values for a more complicated model which includes the random effects and then using results of this model as initial values for a full model. This step by step modeling (where different initial values may be tried in each step to get sure about the results) may help to be certain about global optimization.

In our analysis, that the log-likelihood involves a double integral over the random effects b_i and u_i, which complicates the estimation process. In our current implementation, we approximate these integrals using numerical techniques instead of attempting to solve them analytically. Specifically, we utilize a combination of Monte Carlo integration and numerical quadrature methods to estimate the integrals.The general idea behind Monte Carlo methods is to use random sampling to obtain numerical results. Instead of calculating an integral directly, we estimate it by generating random points in the region of interest and evaluating the function at those points. We generate a large number of samples from the probability distributions of these random effects.

In the likelihood function, a double integral involving random variables, the process estimation as follows:
1) We have (b_i, u_i) follows a bivariate normal distribution with zero mean and Σ_{bu}, we will generate random samples from that distribution.
2) For each set of randomly sampled values (a combination of (b_i, u_i), we evaluate a likelihood function. The likelihood function represents how well the sampled

random effects explain the observed data. This could involve substituting the sampled values into the function we are integrating and calculating how likely the observed outcomes are when using those sampled values.
3) The estimate of the double integral can be computed by averaging the likelihoods obtained from the sampled combinations. The average is then scaled by the total volume of the integration region. we denote the number of samples by M and let L be the likelihood for the sampled values (b_{ij}, u_{ij}), the estimate of the integral can be expressed as: This function may be used for minimization of a function of parameters. For maximization of a likelihood function one may minimize minus log likelihood function. The function **"nlminb"** uses optimization method of port routine which is given in [10]. The function **"nlminb"** uses a sequential quadratic programming (SQP) method to minimize the requested function. The details of this method can be found in [10]. The observed Hessian matrix may be obtained by **"nlminb"** function or may be provided by function **"fdHess"**. Standard deviations are obtained by the square-root of the inverse of the diagonal elements of the observed Hessian matrix.

3.2 The Joint Partially Linear Model with Missing Data

Missing data are highly prevalent in insurance and arise from various sources, including customers refusing to disclose information, offices failing to submit records, limited data at older ages, incomplete diagnosis dates in critical illness policies, and truncation due to deductibles or policy limits. In some cases, such as legal expense insurance, claims adjusters may not record all relevant details. These missing entries reduce the quality of analyses and the accuracy of estimates. The common practice of complete-case analysis (excluding incomplete observations) is often inefficient and inadequate.

Two main missing data patterns are identified in insurance studies: the monotone pattern, typically caused by subject dropout, and the non-monotone pattern, where subjects may leave and later return. Rubin distinguish three missing mechanisms ([23, 27]):

(i) **MCAR (Missing Completely at Random):** Missingness is independent of both observed and unobserved responses.

(ii) **MAR (Missing at Random):** Conditional on the observed data, missingness does not depend on the unobserved responses (assuming parameter distinctness).

(iii) **NMAR (Not Missing at Random):** Missingness depends on the unobserved responses themselves.

From a likelihood perspective, MCAR and MAR are ignorable, whereas NMAR is non-ignorable and requires explicit modeling.

Example: In non-life insurance data with claim counts and amounts, if the probability of recording the claim amount is constant across all individuals, the data are MCAR. If this probability varies with claim amount but is constant within age groups, the data are MAR. If it varies with claim amount even within age groups, the data are NMAR.

Most existing studies, including Derrig and Francis [6], Guelman and Guillén [15], and Wang et al. [33], have conducted their analyses under the MAR assumption.

Rubin proposed the various types of missing mechanisms, MCAR, MAR and NMAR [27]. To handle the missing mechanism issue, we assume R_{y_i} and $R_{\bar{z}_i}$ as the response indicators which take the value 1 if Y_i, respectively $\bar{Z}_i$ is observed and 0 is missing. We assume that the distribution of R_{y_i} and $R_{y_{\bar{z}_i}}$, conditional on $(X_i, Y_i, \bar{Z}_i)$ is multinomial distributions with probabilities $Pr(R_{y_i} = 1|M_{1i}, y_i)$ and $Pr(R_{\bar{z}_i} = 1|M_{2i}, \bar{z}_i)$ where probit models are considered to be

$$\begin{aligned} Pr(R_{y_i} = 1|M_{1i}, y_i) &= \Phi(M_{1i}'\eta_1 + \xi Y_i), \\ Pr(R_{\bar{z}_i} = 1|M_{2i}, \bar{z}_i) &= \Phi(M_{2i}'\eta_2 + \gamma \bar{Z}_i), \end{aligned}$$

where Φ is Standard normal cumulative distribution function, M_{1i} and M_{2i} are certain covariate matrices for the X_i dropout mechanisms, respectively. Also, η_1 and η_2 are the vector of parameters and the parameters ξ and γ, contain information about the missing mechanism for responses.

The factorization of the joint distribution of Y_i, $\bar{Z}_i$, R_{y_i} and $R_{\bar{z}_i}$ can also be considered:

$$\begin{aligned} f(y_i, \bar{z}_i, R_{y_i}, R_{\bar{z}_i}\|b_i, u_i) &= f(R_{y_i}, R_{\bar{z}_i}\|y_i, \bar{z}_i) f(\bar{Z}_i\|b_i) f(y_i\|u_i), \\ &= f(R_{y_i}\|M_{1i}, y_i) f(R_{\bar{z}_i}\|M_{2i}, \bar{z}_i) f(y_i\|b_i) f(\bar{Z}_i\|u_i). \end{aligned}$$

where $f(R_{y_i}|M_{1i}, y_i)$ and $f(R_{\bar{z}_i}|M_{2i}, \bar{z}_i)$ denote the corresponding density function for R_{y_i} given y_i and the density function for $R_{\bar{z}_i}$ given $\bar{z}_i$, respectively.

The joint model for the two responses is written as:

$$\begin{aligned} &Y_i|b_i \sim PS(\theta_i), \quad \log(E(Y_i|b_i)) = X_i^{(1)'}\beta + W_i^{(1)'}b_i + k_1(V_i^{(1)}), \qquad (Model\ 2) \\ &\bar{Z}_i|u_i, Y_i \sim EDF(\mu_i, \sigma^2/Y_i), \quad \log E(\bar{Z}_i|b_i) = X_i^{(2)'}\alpha + W_i^{(2)'}u_i + k_2(V_i^{(2)}), \\ &Pr(R_{y_i} = 1|M_{1i}, y_i) = \Phi(M_{1i}'\eta_1 + \xi Y_i), \\ &Pr(R_{\bar{z}_i} = 1|M_{2i}, \bar{z}_i) = \Phi(M_{2i}'\eta_2 + \gamma \bar{Z}_i). \end{aligned}$$

where

$$\begin{aligned} R_i = (b_i, u_i) &\overset{iid}{\sim} MVNormal(0, \Sigma_{bu}), \\ \Sigma_{bu} &= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \end{aligned}$$

For finding the condition for MCAR, the vector of parameters ξ and γ contain information about the missing mechanism of the responses. If $\xi = \gamma = 0$ then the missing mechanisms are completely at random (MCAR), otherwise, the missingness

is not at random. Note, that $\xi(\gamma)$ shows that the missingness of the claim count (amounts) responses is dependent on the claim count (amounts) responses.

The log-likelihood function under the model (2) is

$$\log(L) = \sum_{i=1}^{n} \log[\int_{b_i}\int_{u_i} f(R_{y_i}\,|M_{1i}, Y_i) f(R_{\bar{z}_i}\,\big|M_{2i}, \bar{Z}_i\,) \\ f(Y_i\,|\bar{z}_i, b_i\,, X_i) f(\bar{Z}_i|u_i, X_i)\varphi_{12}(b_i, u_i; \Sigma_{bu})\; db_i du_i].$$

Details on the logarithm of the Model (2) likelihood function are provided in the Appendix **A**.

4 The Joint Partially Aggregate Claims Model

In this Section, we propose the joint partially aggregate claims model, with and without missing data, for claim count reported by the policyholder and incurred claim amounts for the insured over a fixed time period, using random effects.

One common approach to the aggregate claims model is to assume independent between claim count and amounts and fit a separate model to each response variable. Let Y_i (count response) be the claim count reported by policyholder i over a fixed time period for $i = 1, 2, ..., n$. Also, let Z_{ij}, $j = 1, 2, ..., Y_i$, indicate is the jth claim amount for insured i incurred which it is assumed that conditionally on Y_i, the individual claim amounts, these Z_{ij} are independent.

For a given class of policyholders, the aggregate claims can be expressed as $S_i = \sum_{j=1}^{Y_i} Z_{ij}$, So, the mean and variance of S_i are given by $E(S_i) = E(Y_i)E(Z_{ij})$ and $Var(S_i) = E(Y_i)Var(Z_{ij}) + Var(Y_i)E(Z_{ij})^2$.

Garrido et al. (2016) proposed separate generalized linear models. In this model, this amounts to assuming that for given link functions $g_1(E(Y_i)) = X_i'\beta$ and $g_2(E(Z_{ij})) = X_i'\alpha$, where X_i is is the design matrix for the ith individual, β and α are vectors of regression coefficients. So, Y_i and Z_{ij} are assumed to be independent, one can write the mean aggregate claims in terms of the two marginal mean models as $E(S_i|X) = g_1^{-1}(X_i'\beta)g_2^{-1}(X_i'\alpha)$.

Theorm 1: The dependent partially aggregate claims model by using random effects for complete data under the model (1) is

$$E[S_i] = \mu_i^* M_{R_i}(t_i^*),$$

where $\mu_i^* = \exp[X_i^{(1)'}\beta + X_i^{(2)'}\alpha + k_1(V_i^{(1)}) + k_2(V_i^{(2)})]$ and $M_{R_i}(t_i^*) = \exp(\frac{1}{2}t_i'\Sigma_{bu}t_i)$ is the moment generating function of $R_i = (u_i, b_i)$ and $t_i^* = (W_i^{(1)}, W_i^{(2)})$ based on multivariate normal distribution.

Proof: The dependent partially aggregate claims model with using random effects

for complete data under model (1) is:

$$\begin{aligned}
&E[S_i] = E[E(S_i|b_i, u_i)],\\
&= E[E(Y_i\bar{Z}_i|b_i, u_i)],\\
&= E[E(Y_i|b_i)E(\bar{Z}_i|u_i)],\\
&= \int_{b_i}\int_{u_i} \exp[X_i^{(1)'}\beta + W_i^{(1)'}b_i + k_1(V_i^{(1)})]\\
&\quad \exp[X_i^{(2)'}\alpha + W_i^{(2)'}u_i + k_2(V_i^{(2)})]\varphi_{12}(b_i, u_i; \Sigma_{bu}))db_i du_i,\\
&= \mu_i^* M_{R_i}(t_i^*),
\end{aligned}$$

where $\mu_i^* = \exp[X_i^{(1)'}\beta + X_i^{(2)'}\alpha + k_1(V_i^{(1)}) + k_2(V_i^{(2)})]$ and $M_{R_i}(t_i^*) = \exp(\frac{1}{2}t_i'\Sigma_{bu}t_i)$ is the moment generating function of $R_i = (u_i, b_i)$ and $t_i^* = (W_i^{(1)}, W_i^{(2)})$ based on multivariate normal distribution.

Theorm 2: The variance of the partially aggregate claims for complete data under model (1)

$$Var(S_i) = Var(E(S_i|R_i)) + E[Var(S_i|R_i)],$$

where

$$\begin{aligned}
Var(E(S_i|R_i)) &= \mu_i^*[M_{R_i}(2t_i^*) - (M_{R_i}(t_i^*))^2],\\
E[Var(S_i|R_i)] &= \mu_i^* E[\exp(t'^*_i R_i)h(R_i)],
\end{aligned}$$

and

$$h(R_i) = \frac{Var(Y_i|b_i)}{E^2(Y_i|b_i)} + \frac{Var(\bar{Z}_i|u_i)}{E^2(\bar{Z}_i|u_i)} + \frac{2E(Y_i - E(Y_i|b_i))(\bar{Z}_i - E(\bar{Z}_i|u_i))}{E(Y_i|b_i)E(\bar{Z}_i|u_i)}.$$

where $\mu_i^* = \exp[X_i^{(1)'}\beta + X_i^{(2)'}\alpha + k_1(V_i^{(1)}) + k_2(V_i^{(2)})]$ and $M_{R_i}(t_i^*) = \exp(\frac{1}{2}t_i'\Sigma_{bu}t_i)$ is the moment generating function of $R_i = (u_i, b_i)$ and $t_i^* = (W_i^{(1)}, W_i^{(2)})$ based on multivariate normal distribution.

Proof: Appendix **B**.

Theorem 3: The dependent aggregate claims model by using random effects for missing data under the model (2) is:

$E[S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1] = \mu_i^* M_{R_i}(t_i^*) + E(h^*(R_i))$,

where

$h^*(R_i) = \frac{Cov(Y_i\bar{Z}_i, R_{\bar{z}_i}|R_i)}{Var(R_{\bar{z}_i}|R_i)}(1 - E(R_{y_i}|R_i)) + \frac{Cov(Y_i\bar{Z}_i, R_{y_i}|R_i)}{Var(R_{y_i}|R_i)}(1 - E(R_{\bar{z}_i}|R_i))$.

where $\mu_i^* = \exp[X_i^{(1)'}\beta + X_i^{(2)'}\alpha + k_1(V_i^{(1)}) + k_2(V_i^{(2)})]$ and $M_{R_i}(t_i^*) = \exp(\frac{1}{2}t_i'\Sigma_{bu}t_i)$ is the moment generating function of $R_i = (u_i, b_i)$ and $t_i^* = (W_i^{(1)}, W_i^{(2)})$ based on multivariate normal distribution.

Proof: Appendix **C**.

The Pearson residuals for S_i, assuming complete data, can take the form

$$r_{s_i} = \frac{S_i - E(S_i)}{\sqrt{Var(S_i)}},$$

where $E(S_i)$ and $Var(S_i)$ which have been given in Theorems 1 and 2.

The missing values of responses create problems for the usual residual diagnostics, (see [34]). These unconditional residuals will be misleading if missingness is MAR or NMAR as in these cases the values of S_i come from a conditional or truncated conditional distribution. We can examine the residuals of the S_i conditional on being observed. The Pearson's residuals for S_i can take the form

$$r_{s_i} = \frac{S_i - E(S_i|R_{y_i} = 1, R_{\tilde{z}_i} = 1)}{\sqrt{Var(S_i|R_{y_i} = 1, R_{\tilde{z}_i} = 1)}},$$

where $E(S_i|R_{y_i} = 1, R_{\tilde{z}_i} = 1)$ and $Var(S_i|R_{y_i} = 1, R_{\tilde{z}_i} = 1)$ which have been given in Theorem 3.

5 Sensitivity Analysis

We used sensitivity analysis for the average percent difference (APD). In sensitivity analysis, an assessment of the influence of minor perturbations of the model is important. In sensitivity analysis, a general method for assessing the local influence of minor perturbations on a the average percent difference (APD) is presented. Likelihood displacement and global influence will be used for measuring the influence of a perturbation into the average percent difference (APD) using models formulation in purposed models (model 1 and model 2), through $APD(\rho_{bu})$ to be perturbed.

Let $P_{i1}(\rho_{bu})$ be the pure premiums calculated according to (2) and $P_{i2}(0)$ be the pure premiums calculated the independent model $\rho_{bu} = 0$, respectively. Following similar steps in Garrido et al. (2016), we define the average percent difference (APD) in the expected aggregate claims as follows:

$$APD(\rho_{bu}) = \frac{1}{n}\sum_{i=1}^{n} APD_i(\rho_{bu}),$$

where $APD_i(\rho_{bu}) = \frac{P_{i1}(\rho_{bu})}{P_{i2}(0)} - 1$.

Let $L(\eta|APD(0) = 0)$ is the logarithm of likelihood using the purposed models (model (1) and model (2)) under $APD(0) = 0$ (the independent models) and $L(\eta|APD(\rho_{bu}))$ is the logarithm of likelihood using the purposed models (model (1) and model (2)) under different values of $APD(\rho_{bu})$ except zero. Also, let $\hat{\eta}$ and $\hat{\eta}_{\rho_{bu}}$ denote the maximum likelihood estimators under $L(\eta|APD(0))$ and $L(\eta|APD(\rho_{bu}))$. To assess the influence of varying $APD(\rho_{bu})$, we present the likelihood displacement defined as:

$$LD(APD(\rho_{bu})) = 2[L(\hat{\eta}) - L(\hat{\eta}_{\rho_{bu}})]. \qquad (3)$$

We can quantify the differences using likelihood displacement defined as (3). The $LD(APD(\rho_{bu}))$ will be large if $L(\eta\,|APD(\rho_{bu}))$ is strongly curved at $\hat{\eta}$ which means that η is estimated with high precision, and small otherwise. A graph of $LD(APD(\rho_{bu}))$ versus $APD(\rho_{bu})$ provides a global sensitivity analysis.

6 Simulations

In order to illustrate the effect of dependence in the aggregate claims model, simulation studies were carried out using a fictive portfolio involving $n = 10,000$ classes with and without missing data. Two simulation studies were considered. The simulation studies are performed to evaluate the performance of the proposed model (1). In this simulation study, we compared the biased and mean square error (MSE) of the estimates of parameters of the model (1) and independent model (when two responses of the model (1) are independent). Also, with using sensitivity analysis, we show how the estimates of parameters of the model (1) are affected by the level dependence parameters.

Parameter specifications: The true parameters for all scenarios are: $\beta_0 = 0$, $\beta_1 = 0.5$, $\alpha_0 = 0$, $\alpha_1 = 0.5$, $\sigma_b^2 = \sigma_u^2 = 1$, $\rho_{bu} = 0.5$, with overdispersion parameters $\sigma = 2$ (for NB) and $\nu = 2$ (for CMP).

Distributions: Claim counts are generated from Poisson (Scenario A), Negative Binomial (Scenario B), or Conway-Maxwell Poisson (Scenario C). Claim amounts follow a Gamma distribution with mean μ_i and dispersion σ^2/Y_i.

Evaluation metrics: We compute Bias, Mean Squared Error (MSE), and 95% coverage probabilities:

$$\text{Bias}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta),$$

$$\text{MSE}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2.$$

Sample sizes and replications: We use $n = 100, 1000, 10000$ and $R = 1000$ replications. All models are fitted using the `nlminb` routine in R with truncated power basis splines.

For each class, $i = 1, ..., m$, a covariate variable X_i was generated from a standard normal distribution and a vector of random effect (b_i, u_i) was generated from a bivariate normal distribution with zero means and matrix covariance $\Sigma_{bu} = \begin{pmatrix} \sigma_b^2 & \rho_{bu}\sigma_b\sigma_b \\ \rho_{bu}\sigma_b\sigma_b & \sigma_u^2 \end{pmatrix}$, where we let $\sigma_b^2 = \sigma_u^2 = 1$ and $\rho_{bu} = 0.5$.
Also, a claim count Y_i given the random effect b_i is generated by three distributions:
Scenario A: a Poisson distribution with true values of parameters $\beta_0 = 0$ and $\beta_1 = 0.5$.
$Y_i \sim Poisson(\theta_i)$, $\log(\theta_i) = \beta_0 + \beta_1 X_i + b_i + \sin(2\pi X_i)$. In 239 cases, it turned out that $Y_i = 0$, and hence S_i was set equal to 0.
Scenario B: a negative binomial distribution with true values of parameters $\beta_0 = 0$, $\beta_1 = 0.5$ and $\sigma = 2$. $Y_i \sim NB(\mu_i, \sigma)$, $\log(\mu_i) = \beta_0 + \beta_1 X_i + b_i + \sin(2\pi X_i)$. In 239 cases, it turned out that $Y_i = 0$, and hence S_i was set equal to 0.
Scenario C: a Conway-Maxwell Poisson distribution with true values of parameters $\beta_0 = 0$, $\beta_1 = 0.5$ and $\nu = 2$. $Y_i \sim CMP(\theta_i, \nu)$, $\log(\theta_i) = \beta_0 + \beta_1 X_i + b_i + \sin(2\pi X_i)$.

In 239 cases, it turned out that $Y_i = 0$, and hence S_i was set equal to 0.
In the remaining 9761 cases for scenario **A**, 9761 cases for scenario **B** and 9761 cases for scenario **C** an average claim size $\bar{Z}_i$ was generated by simulating Y_i individual claim amounts from a Gamma distribution with true values of parameters $\alpha = 0$, $\alpha_1 = 0.5$ and $\sigma = 2$.

$$\bar{Z}_i \sim Gamma(\mu_i, \frac{\sigma}{Y_i}), \qquad \log(\mu_i) = \alpha_0 + \alpha_1 X_i + u_i - \cos(2\pi X_i).$$

In order to ensure the results of the numerical algorithm, the model is fitted by **nlminb** package from software **R**. We use the proposed joint partially linear model by spline functions with the truncated power basis. The results of the truncated power basis show that all the parameters are estimated well.

Table 1: Results of the simulation study for scenarios A, B and C.

Scenario A		n=100		n=1000		n=10000	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.000	0.036	0.211	0.020	0.143	0.003	0.015
β_1	0.500	0.491	0.213	0.495	0.039	0.498	0.012
α_0	0.000	0.021	0.151	0.013	0.100	0.003	0.013
α_1	0.500	0.512	0.142	0.508	0.089	0.501	0.012
ρ_{bu}	0.500	0.490	0.123	0.495	0.051	0.499	0.018
σ_b^2	1.000	1.187	0.183	1.019	0.064	1.003	0.019
σ_u^2	1.000	1.132	0.139	1.076	0.059	1.005	0.017
Scenario B		n=100		n=1000		n=10000	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.000	0.026	0.138	0.010	0.012	0.002	0.003
β_1	0.500	0.521	0.103	0.512	0.071	0.501	0.002
α_0	0.000	0.034	0.101	0.011	0.053	0.003	0.007
α_1	0.500	0.511	0.124	0.509	0.083	0.502	0.009
ρ_{bu}	0.500	0.489	0.122	0.492	0.091	0.505	0.008
σ_b^2	1.000	1.190	0.197	1.019	0.078	1.003	0.002
σ_u^2	1.000	1.122	0.133	1.073	0.050	1.005	0.007
σ	2.000	2.119	0.216	2.063	0.148	2.005	0.008
Scenario C		n=100		n=1000		n=10000	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.000	0.011	0.119	0.010	0.010	0.001	0.001
β_1	0.500	0.507	0.101	0.502	0.031	0.501	0.002
α_0	0.000	0.022	0.100	0.010	0.044	0.002	0.005
α_1	0.500	0.509	0.113	0.503	0.069	0.501	0.008
ρ_{bu}	0.500	0.490	0.102	0.495	0.082	0.499	0.005
σ_b^2	1.000	1.183	0.182	1.019	0.072	1.001	0.001
σ_u^2	1.000	1.111	0.123	1.055	0.048	1.001	0.006
ν	2.000	2.019	0.206	2.013	0.138	1.005	0.008

The aim of this simulation study is the accuracy and efficiency of the estimations. For this we consider $n = 100,\ 1000,\ 100000$ for scenarios A, B and C. In this analysis we use 1000 sets of simulation. Table 1 contains the average estimated values of the parameters for scenarios A, B and C. The parameter estimates by the model(1)for scenarios A, B and C are close to the true values of the parameters. Of course, the more the value of n the better the estimates. So, the accuracy and efficiency of the estimations are increased. Figure 1 shows box plots for the mean square error of $E(S)$ $(MSE(E(S)))$ for the model (1) and the independent model under scenarios **A**, **B** and **C** with $n = 10,000$. The mean square error of estimates of the parameters under the independent model for scenarios **A**, **B** and **C** are generally larger than those for the model (1). So, comparing parameter-specific estimates show that, $MSE(E(S))$ for the independent model under scenarios **A**, **B** and **C** estimates were mostly larger than those for the model (1).

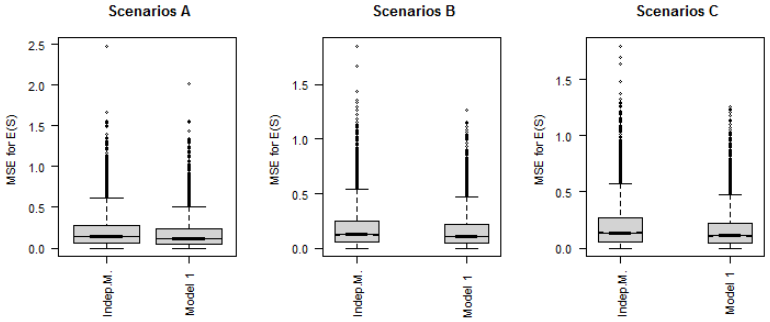

Figure 1: Box plots of $MSE(E(S))$ for the model (1) and the independent model under Scenarios **A**, **B** and **C**.

To investigate the performance of the model(2) and to compare it with that of an independent model (when two responses of the model (2) are independent), the following simulations will be studied. Also, with using sensitivity analysis, we show how the estimates of parameters of the model (2) are affected by the level dependence parameters. Response variables (Y_i with three scenarios (**A**, **B** and **C**) and $\bar{Z}_i$) and covariate variables (X_i and M_{2i}) are generated as Section 5.1.1. In simulations for incomplete data, missing mechanism was added to responses. The coefficient associated with the covariate in the missing probability is chosen to be $\eta_{01} = \eta_{02} = 0$ and $\eta_{11} = \eta_{12} = 1$ and true parameter are as $\xi = 1$ and $\gamma = 1$, which implies that this missingness mechanism is NMAR. The responses R_{y_i} and $R_{\bar{z}_i}$ are obtained by generating a Bernoulli random variable with probability of success $Pr(R_{y_i} = 1|M_{i1}, y_i) = \Phi(\eta_{01} + M_{1i}\eta_{11} + \xi Y_i)$ and $Pr(R_{\bar{z}_i} = 1|M_{i2}, \bar{z}_i) = \Phi(\eta_{02} + M_{2i}\eta_{12} + \gamma \bar{Z}_i)$ then assigning Y_i and $\bar{Z}_i$ as missing if $R_{y_i} = 1$ and $R_{\bar{z}_i} = 1$ and otherwise, $R_{y_i} = 0$ and $R_{\bar{z}_i} = 0$ are observed. Incomplete data are generated

from following models:

$$
\begin{aligned}
&Y_i|b_i \sim PS(\theta_i), \quad \log(\theta_i) = \beta_0 + X_i\beta_1 + b_i + \sin(2\pi X_i),\\
&\bar{Z}_i|u_i, Y_i \sim Gamma(\mu_i, \sigma^2/Y_i), \qquad \log\mu_i = \alpha_0 + X_i\alpha_1 + u_i - \cos(2\pi X_i),\\
&Pr(R_{y_i} = 1|M_{1i}, Y_i) = \Phi(\eta_{01} + M_{1i}\eta_{11} + \xi y_i),\\
&Pr(R_{\bar{z}_i} = 1|M_{2i}, \bar{z}_i) = \Phi(\eta_{02} + M_{2i}\eta_{12} + \gamma\bar{Z}_i).
\end{aligned}
$$

where

$$
\begin{aligned}
&(b_i, u_i) \overset{iid}{\sim} MVNormal(0, \Sigma_{bu}),\\
&\Sigma_{bu} = \begin{pmatrix} \sigma_b^2 & \sigma_u\sigma_b\rho_{bu} \\ \sigma_u\sigma_b\rho_{bu} & \sigma_u^2 \end{pmatrix},
\end{aligned}
$$

In order to ensure the results of the numerical algorithm, the model is fitted by **nlminb** package from software **R**. For the integral approximation in the real data application, we set the number of Monte Carlo samples to $M = 500$. To assess the stability of this choice, we conducted a sensitivity analysis with $M = 100$, $M = 500$, and $M = 1000$. The resulting parameter estimates and their standard errors remained virtually unchanged for $M \geq 500$, indicating that the numerical integration is stable and that $M = 500$ provides sufficiently accurate approximations.

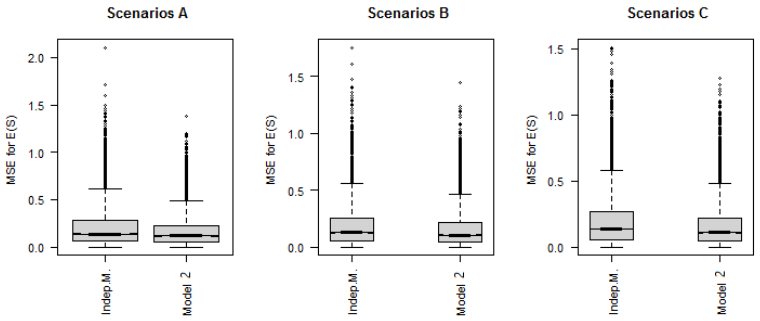

Figure 2: Box plots of $MSE(E(S))$ for the model(2) and independent model under Scenarios **A**, **B** and **C**.

Figure 2 shows box plots for mean square error of $E(S)$ $(MSE(E(S)))$ for the model in equation (3) and independent model under NMAR mechanism for scenarios **A**, **B** and **C**. The mean square error of estimates of the parameters under the independent model for scenarios **A**, **B** and **C** are generally larger than those for the model (2).

Comparing parameter-specific estimates show that, $MSE(E(S))$ for the independent model under NMAR mechanism for scenarios **A**, **B** and **C** estimates were generally larger than those for the model (2).

7 Application

7.1 Data

The data were collected to measure the number of automobile claims (Y) as a count response (we did not consider its values to be zero), average claim amount ($\bar{Z}$) as a continuous response. We describe data contain $m = 46952$ observations, Positive claim counts were observed for 5714 of these dataset. Also, the 2890 individuals claiming bit amounts might be more likely to don't report their claims, or some claims may be more likely to be unreported, depending on the cause of claim. There are several reasons for this including multiple sales channels, customers refusing to disclose data, and offices not submitting the relevant data into their databases. So the claim count and claim amounts are missing (missing rate 6.155%). Table 2, shows Distribution of claim count and average amounts reported by policyholder. It also gives the frequency and percentage of different claim count. As it can be seen more than 87.8 percent of individuals haven't claim count. The total of average count amount (Rials) is 107363957.

Table 2: Distribution of claim count and average amounts reported by the policyholder.

Claim count	Frequency	Percentage	Average amount
Missing	2890	6.155	-
0	38348	81.910	-
1	4778	10.176	24096861
2	771	1.642	29769483
3	55	0.117	53497613
Total	46952		107363957

Also, the vehicle age (X_1), the vehicle insurance discount (X_2), the vehicle value (X_3) and gender (X_4, male=1 and female=0) as explanatory variables from the company insurance of Iran in the year 2018. Table 3 shows the descriptive statistics for the explanatory variables.

Table 3: The descriptive statistics for the explanatory variables.

The explanatory variables	Notation	Mean	S.E.
The vehicle age	X_1	4.96	4.007
The vehicle insurance discount(The unit is Iranian Rial (IRR))	X_2	2174959	4335527
The vehicle value(IRR)	X_3	554350792	760444320
Gender	X_4	73%(Male)	

Figure 3 shows 2D-Dot plot of the average amounts claim for three levels of claim count (left graphs) and 2D-Dot plot of the amounts claim for four levels of claim count (right graphs). Figure 4 shows 2D-Dot plot of the vehicle value for four levels

of claim count (left graphs) and 2D-Dot plot of the vehicle insurance discount for four levels of claim count (right graphs).

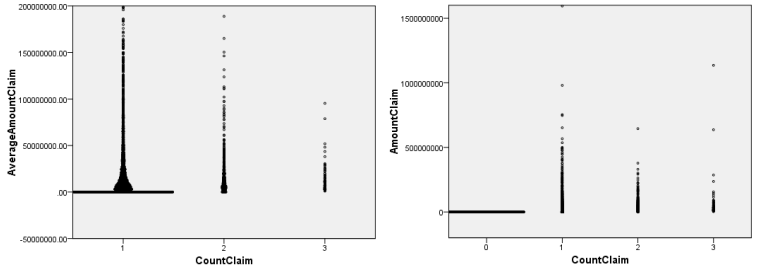

Figure 3: 2D-Dot plot of average amounts value for three levels of claim count (left graph) and 2D-Dot plot of the amounts claim for four levels of claim count (right graph).

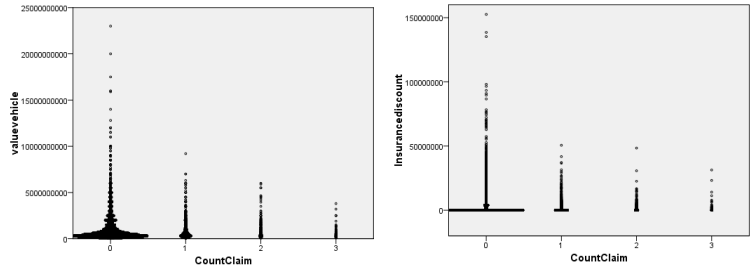

Figure 4: 2D-Dot plot of the amounts claim for three levels of claim count (left graph) and 2D-Dot plot of amounts claim for four levels of claim count (right graph).

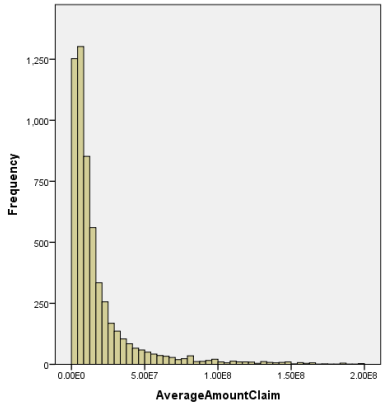

Figure 5: Average claim amount histogram without Missing data

Figure 5 shows average claim amount histogram without missing values. Also, a one-sample Kolmogorov-Smirnov test of the assumption of gamma distribution for average claim amount is not rejected (Kolmogorov-Smirnov Z= 588.859 P-value=0.00001). So, the distribution of the average claim amount was taken to be gamma.

Next, let's investigate the assumptions related to the claim count. One of the investigated hypotheses is the statistical distribution of the claim count. Distributions such as Poisson distribution, negative binomial and Conway Maxwell Poisson distribution are proposed as candidate distributions for the claim count. The choice of these distributions depends on the conditions related to the claim count. One of these conditions is the problem of overdispersion in the claim count. This problem occurs when the average value of the claim count is less than the variance of the claim count. In this case, we cannot use the Poisson distribution for the claim count (because in the Poisson distribution, the mean and the variance are equal). Therefore, distributions such as negative binomial distribution and Conway Maxwell Poisson distribution are used due to overdispersion. Using the likelihood ratio test, we investigate the problem of overdispersion and, as a result, obtain a suitable distribution for the number of claims. We can test for overdispersion with a likelihood ratio test based on Poisson, negative binomial and Conway-Maxwell Poisson distributions.

Let $LL(P)$, $LL(NB)$ and $LL(CMP)$ be logarithm of likelihood under Poisson, negative binomial and Conway-Maxwell Poisson distributions. Also, Let $LRT(P - NB)$ is the likelihood ratio test statistic with an approximate chi-square distribution with $d = 1$ degrees of freedom based on the null hypothesis is stated by saying that the count claim is Poisson distribution and the alternative hypothesis is negative binomial distributions and $LRT(P - CMP)$ is the likelihood ratio test statistic with an approximate chi-square distribution with $d = 1$ degrees of freedom based on the null hypothesis is stated by saying that the count claim is Poisson distribution and the alternative hypothesis is Conway-Maxwell Poisson distribution for count claim.

In this data, We calculated $LL(P) = 20631.209$, $LL(NB) = 20326.411$ and $LL(CMP) = 20217.013$. Also, $LRT(P - NB) = 409.596$ (with 1 d.f and P-value=0.0004) and $LRT(P - CMP) = 828.329$ (with 1 d.f and P-value=0.003). Therefore, the null hypothesis, which indicates the Poisson distribution for the distribution of the number of claims, is rejected, and two other distributions (negative binomial distribution and Conway Maxwell Poisson distribution) are considered as candidates for the distribution of this variable. Now, using the AIC criterion, the appropriate distribution is selected from the negative binomial distribution and the Conway Maxwell Poisson distribution. The Akaike Information Criterion (AIC) for claim count under negative binomial and Conway-Maxwell Poisson distributions are 40654.822 and 40430.026. So, claim count has overdispersion and claim count is Conway-Maxwell Poisson.

The Kendall correlations of the claim count and average amounts reported by

policyholder is 0.124 (P-value= 0.001). Based on the results of the simulation study, we can expect to find a higher value for correlation by our model. This emphasizes that two responses should be modelled simultaneously. Taking into account the correlation, leads us to obtain a more precise estimation of standard errors of estimates and so a better inference. Also, According to economic indicators in Iran, the value of the vehicle has grown a lot. Therefore, it can have a non-linear relationship with count claim and the average amounts claim. Also the joint partially linear model is considered by using the truncated power basis (TPB), B-spline for $k_q(X_{i3})$ for $q = 1, 2$ and linear.

7.2 Models for Analyzing Data

For comparative purposes, three models for model (1) are considered. These models consider with using model (1) by linear and non-linear effect of X_3 with the B-spline and truncated power basis (TPB). These models are

$$\begin{cases} Y_i \sim Conway - MaxwellPoisson(\theta_i, \nu), \\ \log \theta_{it} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 X_{4i} + \sigma_b b_i + k_1(X_{3i}), \\ \bar{Z}_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_4 X_{4i} + \sigma_u u_i + k_2(X_{3i}), \end{cases}$$

where u_i and b_i are independent.

Also, for comparative purposes, three models for model (2) are considered. These models consider with using model(1) by linear and non-linear effect of X_3 with the B-spline and truncated power basis (TPB). These models are with the following form

$$\begin{cases} Y_i \sim Conway - MaxwellPoisson(\theta_i, \nu), \\ \log \theta_{it} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 X_{4i} + \sigma_b b_i + k_1(X_{3i}), \\ \bar{Z}_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_4 X_{4i} + \sigma_u u_i + k_2(X_{3i}), \\ Pr(R_{y_i} = 1|X, y_i) = \Phi(\eta_{01} + \eta_{11} X_{1i} + \eta_{21} X_{2i} + \eta_{31} X_{3i} + \eta_{41} X_{4i} + \xi Y_i), \\ Pr(R_{\bar{z}_i} = 1|X, \bar{z}_i) = \Phi(\eta_{02} + \eta_{12} X_{1i} + \eta_{22} X_{2i} + \eta_{32} X_{3i} + \eta_{42} X_{4i} + \gamma \bar{Z}_i). \end{cases}$$

where

$$(b_i, u_i) \overset{iid}{\sim} MVNormal(0, \Sigma_{bu}),$$
$$\Sigma_{bu} = \begin{pmatrix} \sigma_b^2 & \sigma_u \sigma_b \rho_{bu} \\ \sigma_u \sigma_b \rho_{bu} & \sigma_u^2 \end{pmatrix},$$

With respect to the spline implementation, we have added a new paragraph in the Application section (Section 7) that now explicitly specifies all technical details. For the vehicle value variable X_3, we used cubic splines with polynomial degree $d_q = 3$, three interior knots placed at the 25th, 50th, and 75th percentiles of the covariate distribution. The smoothing parameters δ_1 and δ_2 were selected by minimizing the AIC over a grid of candidate values, yielding $\delta_1 = 0.002$ and $\delta_2 = 0.005$ for the final model. These specifications were kept fixed across all missing-data mechanisms to

ensure comparability, and sensitivity checks with alternative knot numbers confirmed the robustness of our conclusions. All these details are now clearly presented in the revised manuscript.

7.3 Results

Results of using model (1) by linear and non-linear with the B-spline and truncated power basis (TPB), are given in Table 4.

Table 4: Results of using model (1) by linear and non-linear with the B-spline and truncated power basis (TPB). Parameters significant at 5% level are highlighted in bold.

	B-Spline		Linear		TPB	
Count Claim Model (Equation for Y_i)						
	Est.	S.E.	Est.	S.E.	Est.	S.E.
Constant	**-1.670**	**0.024**	**-1.532**	**0.019**	**-1.673**	**0.023**
X_1	**-0.038**	**0.014**	**-0.045**	**0.012**	**-0.039**	**0.013**
X_2	**-0.073**	**0.004**	**-0.063**	**0.006**	**-0.074**	**0.004**
$k_1(X_3)$: linear term	**0.015**	**0.008**	**0.016**	**0.002**	**0.013**	**0.005**
$k_1(X_3)$: quadratic term	**0.016**	**0.005**	-	-	**0.014**	**0.004**
$k_1(X_3)$: cubic term	-	-	-	-	**0.012**	**0.003**
X_4	**1.875**	**0.027**	**1.735**	**0.025**	**1.675**	**0.021**
Average Claim Amount Model (Equation for $\bar{Z}_i$)						
Constant	**16.703**	**0.031**	**14.318**	**0.021**	**16.705**	**0.033**
X_1	**-0.011**	**0.006**	**-0.015**	**0.003**	**-0.013**	**0.007**
X_2	0.012	0.013	0.025	0.021	**0.010**	**0.021**
$k_2(X_3)$: linear term	**0.011**	**0.003**	**0.016**	**0.002**	**0.010**	**0.002**
$k_2(X_3)$: quadratic term	**0.012**	**0.006**	-	-	**0.014**	**0.003**
$k_2(X_3)$: cubic term	-	-	-	-	**0.016**	**0.007**
X_4	0.021	0.032	0.020	0.025	0.023	0.026
ρ_{bu}	**0.578**	**0.083**	**0.580**	**0.081**	**0.572**	**0.081**
σ_u^2	**0.182**	**0.016**	**0.186**	**0.020**	**0.181**	**0.021**
σ_b^2	**0.213**	**0.024**	**0.216**	**0.021**	**0.213**	**0.026**
ν	**2.102**	**0.316**	**2.033**	**0.315**	**2.013**	**0.305**
-loglike	111403.794		111410.673		111400.502	

Notes: (1) For the TPB model with cubic splines and three interior knots, the spline function is $k_q(t) = \eta_{q1}t + \eta_{q2}t^2 + \eta_{q3}t^3 + \eta_{q4}(t-\tau_1)_+^3 + \eta_{q5}(t-\tau_2)_+^3 + \eta_{q6}(t-\tau_3)_+^3$. The reported coefficients correspond to $\eta_{q1}, \eta_{q2}, \eta_{q3}$; the truncated basis coefficients are omitted. (2) $k_1(X_3)$ refers to the claim count model, and $k_2(X_3)$ refers to the average claim amount model. (3) "–" indicates not applicable.

Model (1) with truncated power basis (TBP) is the best fitting model. Table 4 shows significant effect of the vehicle age (X_1), the vehicle insurance discount ($X2$), the vehicle value ($X3$) and gender (male, $X4$) on count claim and a significant effect of the vehicle age (X_1) and the vehicle value ($X3$) on average claim amount.

The variance of random effects (σ_u^2 and σ_b^2) has effect on two responses. The claim count of the male is more than that of females of the claim count. The model (1) with truncated power basis (TBP) also shows if the vehicle insurance discount is increased then the count claim and count amount are decreased. Also, if the vehicle value is increased then the count claim and average claim amount are increased. The model (1) with truncated power basis (TBP) shows a positive correlation between count claim and amount claim ($\hat{\rho}_{bu}$=0.572, P-value=0.001). This gives the conclusion that high values of count claim are associated with high values of amount claim. In this model, due to the consideration of the correlation between the responses, gender and the vehicle insurance discount affect the average claim amount. So, all the explanatory variables have an effect on the count claim and average claim amount. In Table 5, three different forms of model (2) with truncated power basis (TBP) are considered. The first form, assumes a missing completely at random mechanism for the average claim amount which means no affect covariate on $\Phi^{-1}[Pr(R_{\bar{z}_i} = 1|X, \bar{z}_i)]$. In this case, model (2) gives the same results as model (1). The second form of model (2) considered the missing mechanism for the average claim amount as missing at random (MAR). In this model shows a significant effect of the vehicle value ($X3$) and gender (male, $X4$) on $\Phi^{-1}[Pr(R_{y_i} = 1|X, Y_i)]$ and $\Phi^{-1}[Pr(R_{\bar{z}_i} = 1|X, \bar{z}_i)]$. In the third form of the model, there is no constrain for the missing mechanism of the count claim and the average claim amount (NMAR). We use the likelihood ratio to test for NMAR. By these results we can conclude that the two responses are correlated and also the missing indicators for the count claim and the average claim amount ($R_{y_i} = 1|X, Y_i$) and $R_{\bar{z}_i} = 1|X, \bar{z}_i)$) is related to the average claim amount. This leads to have not at random missing mechanism (Deviance= 116.726 with 1 d.f., P-value=0.000). All the covariates are also effective in responding to the responses. To select the best-fitting model, we computed the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all competing models. For the base models (Table 4), the TPB specification yields the lowest AIC (222837.004), indicating superior fit compared to the B-spline and linear alternatives. Although the BIC for TPB is slightly higher than that of the linear model, the additional flexibility in capturing non-linear effects justifies its selection. For the missing-data models (Table 5), the NMAR assumption clearly outperforms both MCAR and MAR, with substantially lower AIC (222760.622) and BIC (222907.856). This is consistent with the likelihood ratio test (Deviance = 116.726, df = 1, p-value < 0.001), confirming that the missingness is non-ignorable. Consequently, the final selected model is Model (2) with the TPB spline basis and the NMAR missing mechanism, which provides the best overall fit with the lowest information criteria.

Table 5: Results using model (2) by truncated power basis (TPB) with missing mechanism (model with assumption of missing completely at random, MCAR, model with missing at random, MAR, assumption and model with not imposing any assumption on the missing mechanisms, NMAR), (Parameters significant at 5% level are highlighted in bold))

	MCAR		*MAR*		*NMAR*	
Parameter	Est.	S.E.	Est.	S.E.	Est.	S.E.
Constant (β_0)	**-1.532**	**0.019**	**-1.641**	**0.017**	**-1.571**	**0.013**
The vehicle age (β_1)	**-0.045**	**0.012**	**-0.047**	**0.011**	**-0.043**	**0.010**
The vehicle insurance discount(β_2)	**-0.063**	**0.006**	**-0.060**	**0.006**	**-0.059**	**0.004**
The vehicle value	**0.016**	**0.002**	**0.014**	**0.002**	**0.015**	**0.003**
	0.015	**0.003**	**0.014**	**0.003**	**0.013**	**0.002**
	0.016	**0.004**	**0.011**	**0.004**	**0.014**	**0.003**
Gender (β_4)	**1.735**	**0.025**	**1.639**	**0.024**	**1.695**	**0.020**
Constant (α_0)	**14.318**	**0.021**	**14.110**	**0.021**	**13.208**	**0.018**
The vehicle age (α_1)	**-0.015**	**0.003**	**-0.014**	**0.003**	**-0.016**	**0.002**
The vehicle insurance discount (α_2)	**0.025**	**0.011**	**0.026**	**0.012**	**0.021**	**0.009**
The vehicle value	**0.015**	**0.003**	**0.014**	**0.004**	**0.012**	**0.001**
	0.011	**0.002**	**0.013**	**0.002**	**0.012**	**0.003**
	0.012	**0.006**	**0.015**	**0.001**	**0.012**	**0.006**
Gender (α_4)	**0.020**	**0.005**	**0.018**	**0.006**	**0.014**	**0.002**
ρ_{bu}	**0.580**	**0.081**	**0.593**	**0.073**	**0.593**	**0.073**
σ_u^2	**0.186**	**0.020**	**0.185**	**0.019**	**0.180**	**0.018**
σ_b^2	**0.216**	**0.021**	**0.215**	**0.017**	**0.213**	**0.017**
ν	**2.033**	**0.315**	**2.036**	**0.312**	**2.037**	**0.310**
Constant (η_{01})	-	-	**1.931**	**0.211**	**1.157**	**0.113**
The vehicle age (η_{11})	-	-	1.284	1.471	1.403	1.269
The vehicle insurance discount (η_{21})	-	-	1.501	1.755	1.649	1.310
The vehicle value	-	-	**1.246**	**0.025**	**1.031**	**0.011**
Gender (η_{41})	-	-	**1.187**	**0.079**	**1.319**	**0.048**
Count Claim (ξ)	-	-	-	-	**1.187**	**0.056**
Constant (η_{02})	-	-	**1.960**	**0.235**	**1.160**	**0.142**
The vehicle age (η_{12})	-	-	1.293	1.574	1.375	1.370
The vehicle insurance discount (η_{22})	-	-	1.543	1.760	1.650	1.307
The vehicle value (η_{32})	-	-	**1.247**	**0.028**	**1.046**	**0.010**
Gender (η_{42})	-	-	**1.192**	**0.085**	**1.321**	**0.053**
Average Claim Amount (γ)	-	-	-	-	**1.290**	**0.503**
-loglike	111410.671		111408.670		111350.311	

Using Pearson residuals, introduced for model (2) with NMAR mechanism. The Pearson's residuals for S_i under model (III) can take the form

$$r_{s_i} = \frac{S_i - E(S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1)}{\sqrt{Var(S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1)}},$$

We find five outliers with residuals larger than 3. Plots of the residuals (bean plot and box plot) for S_i under model (III) given Figure 6.

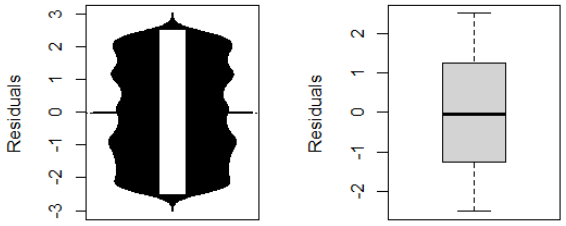

Figure 6: Bean plot of the residuals for S_i (left graph) and box plot of the residuals for S_i (right graph).

To search for sensitivity analysis, we find likelihood displacement for different values of $ADP(\rho_{bu})$ in the interval $(-1, 1)$ is plotted in Figure 7. This curve shows a high curvature around $ADP(0) = 0$, so that the differing values of $LD(ADP(\rho_{bu}))$ affects the model results, hence final results of the model is highly sensitive to the dependent parameter $ADP(\rho_{bu})$ for the purposed model (2) with NMAR mechanism..

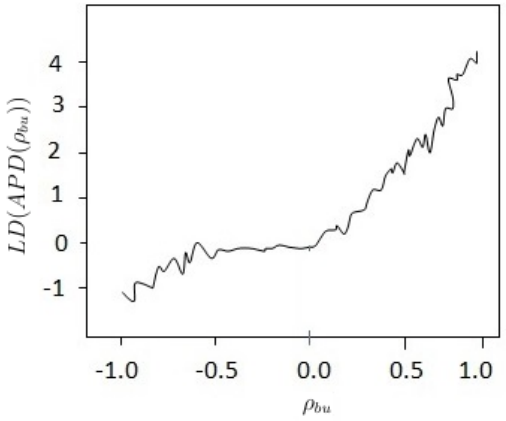

Figure 7: Likelihood displacement against values of $ADP(\rho_{bu})$.

This reflects the influence of the condition show that the influence of $APD(\rho_{bu})$ on the pure premium turns out to be significant. This approach measures the sensitivity in the neighborhood of the proposed model (model (2)) to see whether admitting the correlation between the claim count and the amount of the claim of the proposed models (model (2)) would affect the likelihood function and the expected aggregate loss of the joint model. The sensitivity of the aggregate pure premium to the dependence parameter ρ_{bu} is assessed using the likelihood displacement measure

LD(APD). As shown in Figure 7, plotting LD(APD) against ρ_{bu} is appropriate since ρ_{bu} is the independent variable. The computed value of LD(APD) at the estimated correlation $\hat{\rho}_{bu} = 0.572$ is 0.172, indicating that the model results are sensitive to the dependence parameter. A clarifying note has been added to the figure caption to avoid any ambiguity regarding the axis labels.

8 Conclusion

In this paper, we have presented joint partially linear models for aggregate claims, comprising claim counts and average claim amounts, under both complete and missing data settings. The proposed framework incorporates correlated random effects to capture dependence between frequency and severity components, and uses spline-based non-linear structures for continuous covariates. A selection model approach is adopted to handle non-ignorable missingness (NMAR), with MCAR and MAR as special cases. Estimation is carried out via maximum likelihood with numerical integration over the random-effects distribution. Residual diagnostics are also introduced for outlier detection. Simulation studies indicate that, under the assumed distributional settings (Poisson, Negative Binomial, or CMP for claim counts, and Gamma for claim amounts) and with the chosen spline specifications, the proposed joint models yield estimates with small bias and MSE, and outperform independent models that ignore cross-dependence. For the missing-data case, the NMAR model shows improved fit when missingness is indeed non-ignorable; however, this result is conditional on correct specification of the missing-data mechanism, and sensitivity analyses are recommended in practice. The proposed models are primarily developed for two responses (frequency and severity). Extensions to more than two variables and cases with missing covariates are not addressed here and remain topics for future research. Furthermore, the performance of the spline approximation depends on knot selection and smoothing parameters, which require careful choice in applications. Overall, the proposed methodology provides a flexible framework for joint modeling of insurance claims, but its conclusions should be interpreted with due regard to the underlying assumptions.

Appendix A. The Log-Likelihood Function Under the Model (2)

The log-likelihood function under the model (2) is

$$
\begin{aligned}
L &= \sum_{i=1}^{n} log[\int\int f(Y_i, \bar{Z}_i, R_{y_i}, R_{\bar{z}_i}|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i], \\
&= \sum_{i=1}^{n} log[\int\int f(R_{y_i}, R_{\bar{z}_i}|X_i, M_{1i}, M_{2i}, , Y_i, \bar{Z}_i, b_i, u_i)f(Y_i|X_i, b_i)f(\bar{Z}_i, |X_i, u_i)\varphi_{12}(b, u_i)db_i du_i], \\
&= \sum_{i=1}^{n} log[\int\int f(R_{y_i}|M_{i1}, y_i)f(R_{\bar{z}_i}|M_{i2}, \bar{Z}_i)f(Y_i|X_i, b_i)f(\bar{Z}_i|X_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i].
\end{aligned}
$$

1) For individuals with both Y_i and $\bar{Z}_i$ observed, the likelihood is:

$$
\begin{aligned}
L_{y_i, \bar{z}_i, obs} &= \int\int P(Y_i = y_i, \bar{Z}_i = \bar{z}_i, R_{y_i} = 1, R_{\bar{z}_i} = 1|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b, u_i)db_i du_i, \\
&= \int\int P(R_{y_i} = 1|y_i, M_{i1})P(R_{\bar{z}_i} = 1|\bar{Z}_i, M_{2i})f(y_i|X_i, b_i)f(\bar{z}_i|X_i, u_i)\varphi_{12}(b, u_i)db_i du_i,
\end{aligned}
$$

where by subscript "obs" we mean the observed value.
2) For individual who observe only Y_i and the likelihood is:

$$\begin{aligned}
&L_{y_i,obs} = \int\int P(Y_i = y_i, R_{y_i} = 1, R_{\bar{z}_i} = 0|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i,\\
&= \int_{\bar{z}_i}[\int\int f(y_i, \bar{z}_i, R_{y_i} = 1, R_{\bar{z}_i} = 0|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i]d\bar{z}_i,\\
&= \int_{\bar{z}_i}[\int\int P(R_{\bar{z}_i} = 0|M_{i1}, \bar{z}_i)P(R_{y_i} = 1|M_{i1}, y_i)f(y_i|X_i, b_i)f(\bar{z}_i|X_i, b_i)\varphi_{12}(b_i, u_i)db_i du_i]d\bar{z}_i.
\end{aligned}$$

3) For individual who observe only $\bar{Z}_i$ and the likelihood is:

$$\begin{aligned}
&L_{\bar{z}_i,obs} = \int\int P(\bar{Z}_i = \bar{z}_i, R_{\bar{z}_i} = 1, R_{y_i} = 0|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i,\\
&= \sum_{y_i}[\int\int f(\bar{z}_i, y_i, R_{\bar{z}_i} = 1, R_{y_i} = 0|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i],\\
&= \sum_{y_i}[\int\int P(R_{\bar{z}_i} = 0|M_{i1}, \bar{z}_i)P(R_{y_i} = 1|M_{i2}, y_i)f(y_i|X_i, b_i)f(\bar{z}_i|X_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i].
\end{aligned}$$

4) For individuals with both Y_i and $\bar{Z}_i$ missing the likelihood is:

$$\begin{aligned}
&L_{y_i,\bar{z}_i,mis} = \int\int P(R_{\bar{z}_i} = 0, R_{y_i} = 0|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i,\\
&= \sum_{y_i}[\int_{\bar{z}_i}[\int\int f(\bar{z}_i, y_i, R_{\bar{z}_i} = 0, R_{y_i} = 0|X_i, M_{i1}, M_{i2}, b_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i]d\bar{z}_i],\\
&= \sum_{y_i}[\int_{\bar{z}_i}[\int\int P(R_{\bar{z}_i} = 0|M_{i1}, \bar{z}_i)P(R_{y_i} = 0|M_{i2}, y_i)f(y_i|X_i, b_i)f(\bar{z}_i|X_i, u_i)\varphi_{12}(b_i, u_i)db_i du_i]d\bar{z}_i],
\end{aligned}$$

where by subscript "miss" we mean the missing value.

Appendix B. Proof for Theorm 2

we have variance of the aggregate claims

$$Var(S_i) = Var(E(S_i|R_i)) + E[Var(S_i|R_i)],$$

where

$$\begin{aligned}
&Var(E(S_i|R_i)) = Var(E(Y_i\bar{Z}_i|R_i)),\\
&= Var(E(Y_i|b_i)E(\bar{Z}_i|u_i)),\\
&= (\mu^*)^2 Var(exp(t_i'^* R_i)),\\
&= (\mu^*)^2[E(exp(2t_i^{*'} R_i)) - E^2(exp(t_i^{*'} R_i))],\\
&= \mu_i^*[M_{R_i}(2t_i^*) - (M_{R_i}(t_i^*))^2].
\end{aligned}$$

where $\mu_i^* = \exp[X_i^{(1)'}\beta + X_i^{(2)'}\alpha + k_1(V_i^{(1)}) + k_2(V_i^{(2)})]$ and $M_{R_i}(t_i^*) = exp(\frac{1}{2}t_i'\Sigma_{bu}t_i)$ is the moment generating function of $R_i = (u_i, b_i)$ and $t_i^* = (W_i^{(1)}, W_i^{(2)})$ based on multivariate normal distribution.

To obtain $E[Var(S_i|R_i)]$, we need to obtain the variance of the product of two random variables. [16] proposed the Variance of the product of K Random Variables. For the random variable $F_1, ..., F_K$, Goodman proposed $Var(\prod_{i=1}^K F_i)$ is

$$Var(\prod_{i=1}^K F_i) = \prod_{i=1}^K E^2(F_i)[\sum_{i=1}^K G_i + 2\sum_{i<j} E(\delta_i\delta_j)],$$

where $G_i = \frac{Var(F_i)}{E^2(F_i)}$ and $\delta_i = \frac{F_i - E(F_i)}{E(F_i)}$ [16]. So, we have

$$\begin{aligned}
&E[Var(S_i|R_i)] = E[Var(Y_i\bar{Z}_i|R_i)],\\
&= E[E^2(Y_i|b_i)E^2(Y_i|b_i)\ [h(R_i, X_i)]],\\
&= \mu_i^* E[exp(t_i'R_i)h(R_i, X_i)],
\end{aligned}$$

where

$$h(R_i, X_i) = \frac{Var(Y_i|b_i)}{E^2(Y_i|b_i)} + \frac{Var(\bar{Z}_i|u_i)}{E^2(\bar{Z}_i|u_i)} + \frac{2E(Y_i - E(Y_i|b_i))(\bar{Z}_i - E(\bar{Z}_i|u_i))}{E(Y_i|b_i)E(\bar{Z}_i|u_i)}.$$

Appendix C: The Proof for Theorem 3

The dependent aggregate claims model by using random effects for missing data under the model (2) is:

$$E[S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1] = \int\int E(Y_i\bar{Z}_i|R_{y_i} = 1, R_{\bar{z}_i} = 1, b_i, u_i)\phi_{12}(b_i, u_i)db_i du_i,$$

When $R_i = (b_i, u_i)$ in the likelihood function is multivariate the integral in this function is a multivariate integral which may be calculated by a Monte Carlo method (see, Ch. 8 of [9]).

To obtain $E(Y_i\bar{Z}_i|R_{y_i} = 1, R_{\bar{z}_i} = 1, b_i, u_i)$, we use an approximation based on conditional expectations and regression with binary variables.

Joe proposed new approximations for multivariate normal probabilities for rectangular regions, based on conditional expectations and regression with binary variables [22]. The main ideas are quite simple and depend on conditional expectations and regression for binary variables; the form of the multivariate normal distribution is not used, so that the approximations are also usable for other multivariate normal distributions. Our idea of approximation is similar to Joe's idea of approximation [22]. The first step is an approximation of $E(Y_i\bar{Z}_i|R_{y_i} = 1, R_{\bar{z}_i} = 1, b_i, u_i)$ by

$$E(Y_i\bar{Z}_i|b_i, u_i) + \Sigma_{12}\Sigma_{11}^{-1}(1 - E(R_{y_i}), 1 - E(R_{\bar{z}_i}))',$$

where Σ_{12} is a row vector consisting of the entries covariance $Y_i\bar{Z}_i$ and $(R_{y_i}, R_{\bar{z}_i}))'$ and Σ_{11} is a matrix covariance with element $Cov(R_{y_i}, R_{\bar{z}_i})$.

So, we have

$$\begin{aligned} E(Y_i\bar{Z}_i|R_{y_i} = 1, R_{\bar{z}_i} = 1, b_i, u_i) &= E(Y_i|R_i)E(\bar{Z}_i|R_i) \\ &+ \frac{Cov(Y_i\bar{Z}_i, R_{\bar{z}_i}|R_i)}{Var(R_{\bar{z}_i}|R_i)}(1 - E(R_{y_i}|R_i)) + \frac{Cov(Y_i\bar{Z}_i, R_{y_i}|R_i)}{Var(R_{y_i}|R_i)}(1 - E(R_{\bar{z}_i}|R_i)), \end{aligned}$$

where $R_i = (b_i, u_i)$ and

$$\begin{aligned} Var(R_{\bar{z}_i}|R_i) &= E(R_{\bar{z}_i}^2|R_i) - [E(R_{\bar{z}_i}|R_i)]^2, \\ &= E(R_{\bar{z}_i}|R_i) - [E(R_{\bar{z}_i}|R_i)]^2, \\ &= E(E(R_{\bar{z}_i}|Z_iR_i)) - [E(E(R_{\bar{z}_i}|Z_iR_i))]^2, \\ &= E(Y_i\Phi(M_{1i}'\eta_1 + \xi Y_i))(1 - E(Y_i\Phi(M_{1i}'\eta_1 + \xi Y_i))). \end{aligned}$$

Also, $Var(R_{y_i}|R_i)$ is also obtained similarly.

We obtain $Cov(Y_i\bar{Z}_i, R_{y_i}|R_i)$ as follows:

$$E(Y_i\bar{Z}_i, R_{y_i}|R_i) - E(Y_i\bar{Z}_i|R_i)\ E(R_{y_i}|R_i) = E(Y_iR_{y_i}|R_i)E(\bar{Z}_i|R_i) - E(Y_i|R_i)E(\bar{Z}_i|R_i)E(R_{y_i}|R_i),$$

where

$$E[E(Y_iR_{y_i}|R_i, Y_i)] = E[Y_iE(R_{y_i}|R_i, Y_i)] = E[Y_iP(R_{y_i} = 1|R_i, Y_i)] = E(Y_i\Phi(M_1'\eta_1 + \xi Y_i))$$

and $E(\bar{Z}_i|R_i) = \exp(X_i'\alpha + V_i'u_i),\ E(Y_i|R_i) = \exp(X_i'\beta + W_i'b_i),\ E(R_{y_i}|R_i) = \Phi(M_{1i}'\eta_1 + \xi Y_i).$

Also, $Cov(Y_i\bar{Z}_i, R_{\bar{z}_i}|R_i)$ is also obtained similarly. Finally,

$$E[S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1] = \mu_i^* M_{R_i}(t_i) + E(h^*(R_i, X_i)),$$

where

$$h^*(R_i, X_i) = \frac{Cov(Y_i\bar{Z}_i, R_{\bar{z}_i}|R_i)}{Var(R_{\bar{z}_i}|R_i)}(1-E(R_{y_i}|R_i))+\frac{Cov(Y_i\bar{Z}_i, R_{y_i}|R_i)}{Var(R_{y_i}|R_i)}(1-E(R_{\bar{z}_i}|R_i)).$$

Finally, to obtain $Var(S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1)$, we have

$$E[S_i^2|R_{y_i} = 1, R_{\bar{z}_i} = 1] - (E[S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1])^2,$$

where $E[S_i^2|R_{y_i} = 1, R_{\bar{z}_i} = 1]$ is also obtained as $E[S_i|R_{y_i} = 1, R_{\bar{z}_i} = 1]$.

Abbreviations

MCAR:Missing completely at random, MAR: Missing at random, MNAR: Missing not at random, obs: the observed value, miss: the missing value, loglike: logarithm of likelihood, TPB: truncated power basis, P: Poisson, NB: negative binomial, CMP: Conway-Maxwell Poisson, pmf: the probability mass function, PS: power series,EDF:exponential dispersion family, APD: the average percent difference.

Acknowledgements

We would like to thank the editor and referees for reading and giving many improving comments.

Authors' Contributions

EBS was supervisor of the paper and proposed the idea of considering nonresponse according to its reasons in modeling the survey results in the form of a multivariate selection model. He also proposed sensitivity analysis for the proposed model.

Funding

Not applicable.

Availability of Data and Materials

The data that support the findings of this study have been provided by the Central Insurance of Iran. Such data are not publicly available.

Declarations

Competing Interests

The authors declare that they have no competing interests.

Bibliography

[1] D. M. Berridge, and D. M. Dos Santos, *Fitting a random effects model to ordinal recurrent events using existing software*, J. Statist. Comput. Simul., (1996) 55, 73-86.

[2] E. C. K. Cheung, W. Ni, R. Oh, and J. K. Woo, *Bayesian credibility under a bivariate prior on the frequency and the severity of claims* Insurance: Mathematics and Economics, (2021) 100, 274-295.

[3] D. R. Cox, and N. Wermuth, *Response models for mixed binary and quantitative variables*, Biometrika, (1992) 79(3): 441-461.

[4] C. Czado, R. Kastenmeier, E. C. Brechmann, and A. Min, *A mixed copula model for insurance claims and claim sizes*, Scand. Actuar. J., (2021) 4, 278-305.

[5] C. De Boor, *A practical guide to splines*, Springer-Verlag, (1987) New York.

[6] R. Derrig, and L. Francis, *Distinguishing the Forest from the TREES: A Comparison of Tree Based Data Mining Methods*, Casualty Actuarial Society Forum, (2006) 1-49.

[7] R. F. Engle, C. W. J. Granger, J. Rice, and A. Weiss, *Semiparametric estimates of the relation between weather and electricity sales*, *Journal of the American statistical Association*, (1986) 81(394), 310–320.

[8] P. H. Eilers, and B. D. Marx, *Flexible smoothing with B-splines and penalties*, Stat Sci, (1996) 11(2), 89–102.

[9] L. Fahrmeir, and G. Tutz, *Multivariate statistical modelling based on generalized linear models*, Springer, New york (2001).

[10] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, New York (2000).

[11] E. W. Frees, J. Gao, and M. A. Rosenberg, *Predicting the frequency and amount of health care expenditures*, N. Am. Actuar. J. (2011) 15 (3), 377-392.

[12] E. W. Frees, G. Lee, and L. Yang, *Multivariate frequency-severity regression models in insurance* Risks, (2016) 4(1), 1-36.

[13] J. Garrido, C. Genest, and J. Schulz, *Generalized linear models for dependent frequency and severity of insurance claims*, Insurance: Mathematics and Economic, (2016) 70, 205-215.

[14] S. Gschlössl, and C. Czado, *Spatial modelling of claim frequency and claim size in non-life insurance*, Scandinavian Actuarial Journal, (2007)3, 202-225.

[15] L. Guelman, and M. Guillén, *A causal inference approach to measure price elasticity in Automobile Insurance*, Expert Systems with Applications, (2014) 41, 387-396.

[16] L. Goodman, *The Variance of the Product of K Random Variables*, Journal of the American Statistical Association, (1962) 57(297): 54-60.

[17] D. A. Harvile, and R. W. Mee, *A mixed model procedure for analyzing ordered categorical data*, Biometrics, (1984) 40, 393-408.

[18] T. J. Hastie, and R. J. Tibshirani, *Generalized additive models*, CRC press, (1990) 43.

[19] J. J. D. Heckman, *Dummy Endogenous variable in a simultaneous Equations system*, Econometrica, (1978) 46 (6), 931-59.

[20] H. JEONG, E. VALDEZ, J. AHN, AND A. PARK, *Generalized linear mixed models for dependent compound risk models*, SSRN ELECTRONIC JOURNAL (2019) https://dx.doi.org/10.2139/ssrn.3045360.

[21] B. JØRGENSEN, M. C. P. DE SOUZA, *Fitting Tweedie's compound Poisson model to insurance claims data*, SCAND. ACTUAR. J., (1994) 1, 69-93.

[22] H. JOE, *Approximations multivariate normal rectangle probabilities based on conditional expectation*, *Journal of the American Statistical Association*, (1995) 90: 957-967.

[23] R. J. LITTLE, AND M. SCHLUCHTER, *Maximum likelihood estimation for mixed continuous and categorical data with missing values*, BIOMETRIKA, (1987) 72, 497-512.

[24] R. OH, P. SHI, AND J. AHN, *Bonus-Malus premiums under the dependent frequency-severity modeling*, SCANDINAVIAN ACTUARIAL JOURNAL, (2020) 3, 172-195.

[25] O. A. QUIJANO-XACUR, AND J. GARRIDO, *Generalised linear models for aggregate claims: To Tweedie or not?*, EUR. ACTUAR. J., (2015) 5 (1), 181-202.

[26] A. E. RENSHAW, *Modelling the claims process in the presence of covariates*, ASTIN BULL, (1994) 24 (2), 265-285.

[27] D. B. RUBIN, *Inference and missing data*, BIOMETRICA, (1976) 82, 669-710.

[28] G. SHMUELI, T. P. MINKA, AND J. B. KADANE, S. BORLE, AND P. BOATWRIGHT, *A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution*, APPL. STAT. (2005) 54, 127-142.

[29] K. F. SELLERS, S. BORLE, AND G. SHMUELI, (2012). *The COM-Poisson model for count data: A survey of methods and applications*, APPL. STOCH. MODEL. BUS. (2012) 28, 104-116.

[30] J. C. STONE, *Additive regression and other nonparametric models*, THE ANNALS OF STATISTICS. (1985) 13(2). 689–705.

[31] G. TUTZ, *Regression for Categorical Data*, CAMBRIDGE UNIVERSITY PRESS, Cambridge (2012).

[32] G. VERBEKE, AND G. MOLENBERGHS, *Linear mixed models in practice: A SAS Oriented Approach*, SPRINGER (1997).

[33] N. WANG, L. QIAN, N. ZHANG, AND Z. LIU, *Modelling the aggregate loss for insurance claims with dependence*, COMMUNICATIONS IN STATISTICS - THEORY AND METHODS, (2021) 50(9), https://doi.org/10.1080/03610926.2019.1659368.

[34] T. H. BOUKADOUM, AND K. BOUKHETALA, *A Stochastic Process Perspective on Hybrid Log-Normal and Machine Learning Models for Financial Risk under Left-Censored Data*, JOURNAL OF MATHEMATICS AND MODELING IN FINANCE, Allameh Tabataba'i University Press, (2026) 6(1).

How to Cite: Ehsan Bahrami Samani[1], *Joint Partially Models for Dependent Frequency and Severity of Insurance Claims with Using Spline Functions*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 7, No. 1, Pages:1–32, (2027).

The Journal of Mathematics and Modeling in Finance (JMMF) is licensed under a Creative Commons Attribution NonCommercial 4.0 International License.