Incorporating heterogeneity into the prediction of total claim amount

Aslıhan Şentürk Acar [*†], Uğur Karabey[‡] and Dario Gregori[§]

Abstract

This paper proposes an alternative predictor for the total claim amount of individuals that can be used for any type of non-life insurance products in which individuals may have multiple claims within one policy period. The impact of heterogeneity on expected total claim amount is investigated focusing on marginal predictions. Generalized linear mixed model (GLMM) is used for the amounts of loss per claim. Closed-form expression of the predictor is derived using marginal mean under GLMM and claim count distribution. Empirical studies are performed using a private health insurance data set of a Turkish insurance company. Proposed predictive model provides the lowest prediction errors among competing models according to the mean absolute error criterion.

Keywords: Generalized linear mixed model, aggregate loss, marginal mean, zero-inflation, insurance pricing.

2000 AMS Classification: 62P05, 62J12

1. Introduction

Insurance companies cover the unpredictable losses of a policyholder and policyholder pays premium against the insurance coverage in return. Actuaries construct homogeneous risk classes using explanatory variables (rating factors) related to policyholders and use statistical models to determine fair premiums. Generally two components are used for the premium; claim frequency and claim severity. Claim severity is the average cost per claim and the claim frequency is the average number of claims per time period [24]. Pure premium is calculated by the product of claim frequency and claim severity.

Aggregate loss is the total claim amount paid by the insurer for the claims that occur over a fixed period of time. For $j = 1, 2, \cdots, N_i$ let Y_{ij} be the individual

[*]Hacettepe University, Department of Actuarial Sciences, Beytepe, Ankara, Turkey Email: aslihans@hacettepe.edu.tr

[†]Corresponding Author.

[‡]Hacettepe University, Department of Actuarial Sciences, Beytepe, Ankara, Turkey Email: ukarabey@hacettepe.edu.tr

[§]University of Padova, Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic and Vascular Sciences, Padova, Italy Email: dario.gregori@unipd.it

claim amounts for insured i where N_i is the number of claims. Assume that the individual claim amounts follow an aggregate loss model that is defined as,

$$S_i = \sum_{j=1}^{N_i} Y_{ij}$$

where S_i is the aggregate loss of insured i. Conditional on N_i, Y_{ij} are assumed to be independent and identically distributed random variables. The distribution of N_i does not depend on the values of individual claim amounts. Letting Y_{ij} be i.i.d. with $Y_{ij} \stackrel{d}{=} Y_i$, expected aggregate loss is defined as,

$$\mathrm{E}(S_i) = E(N_i)E(Y_i) = E(N_i)E(\overline{Y_i})$$

where $\overline{Y_i}$ is the mean claim amount that is the ratio of total claim amount to the number of claims ([17],[28]). There exist an independence assumption between the claim severity and claim frequency in aggregate loss models. However in practice this assumption is not realistic since claim severity is expected to be affected by claim frequency. One way of modeling the dependence between these two components is to model claim severity conditional on the number of claims and to include the number of claims as an explanatory variable in severity model ([13], [16], [28]).

Generalized linear models (GLMs) are frequently used to model cross-sectional claim frequency and claim severity data that is observed in a fixed period of time. It is common to use gamma, lognormal or inverse Gaussian distributions for the right skewed claim severity data (see [14], [8] and [20]). For the frequency component, Poisson and negative binomial (NB) distributions are commonly used in the literature (see [6], [16] and [27]). Both deductible agreement and no claim discount (NCD) system cause presence of extra zeros in claim count data. Although NB model is preferred to overcome the overdispersion problem that exists in Poisson model, it is not sufficient enough to model excess zero counts [32]. In order to deal with excess zero counts, zero-inflated models ([18], [23]) and hurdle models [23] are frequently used in the literature (see [4], [12], [22] and [30]).

Observations are assumed to be independent in the GLMs. However, policyholders may experience multiple claim events in one policy period and the analyser might be interested in modeling correlation structure among the repeated claim amounts of the same individual in an insurance data such as health insurance. One of the sources of variability that has an impact on the correlation is between-individual heterogeneity [11]. There are two main approaches to model repeated measures; subject-specific models and marginal (population-averaged) models. Heterogeneity is explicitly modeled by using random effects in subject-specific models such as linear mixed model (LMM) and GLMM. In marginal models, dependent variable is modeled as a function of explanatory variables without taking into account heterogeneity [31]. The method of generalized estimating equation (GEE) is typically used to estimate the parameters of marginal models.

Frees et al. [13] used aggregate loss approach to predict annual inpatient and outpatient expenditures of individuals. They used LMM for the log-transformed expenditures per event and NB model for the number of events. In this study, an alternative predictor for the total claim amount of individuals is suggested for

non-life insurance pricing. Differently from Frees et al. [13], this study focuses on marginal predictions considering the difficulty of following the same policyholders over the years for yearly renewable insurance contracts. The objective is to obtain marginal predictions of total claim amounts by taking account of heterogeneity and to model claim amounts in original scale instead of using log transformation. For this purpose, GLMM is used for the amounts of loss per claim of policyholders. Marginal mean is obtained by averaging the conditional mean in GLMM over the distribution of random effects. Number of claims is used as an explanatory variable in GLMM in order to model the dependence between claim severity and claim frequency.

The organization of this paper is as follows. In Section 2 models for claim frequency and claim severity are defined. In Section 3 the proposed predictive model is introduced and the alternative models are summarized. Candidate models are fitted to Turkish private health insurance claims data and the predictive accuracy of the proposed model is compared with the alternative models in Section 4. In the last section, results are summarized and conclusions are drawn.

2. The Models

2.1. Generalized linear models. Generalized linear models are the generalizations of the ordinary linear models. The distribution of dependent variable is a member of exponential family such as normal, Poisson, binomial, gamma and inverse Gaussian. Probability (density) function of a random variable Y, that has a distribution in the exponential family, can be defined as,

$$f_Y(y;\theta,\phi) = \exp\{a(\phi)^{-1}[y\theta - \psi(\theta)] + c(y,\phi)\} \tag{2.1}$$

where θ is the canonical parameter and ϕ is the dispersion parameter. $a(.)$, $\psi(.)$ and $c(.)$ are specified functions. ϕ may be known or unknown. Mean and variance are given respectively by,

$$\mathrm{E}(Y) = \mu = \psi^{'}(\theta)$$

$$\mathrm{Var}(Y) = \sigma^2 = a(\phi)\psi^{''}(\theta)$$

where $\psi^{''}(\theta)$ is called the variance function and regarded as a function of μ, $\nu(\mu)$. $a(\phi)$ is generally in the form of $a(\phi) = \phi/w$ where w is a known prior weight [20].

On an individual basis, let Y_i be the dependent variable of the ith individual for $i = 1, \cdots, K$. Y_i are assumed to be independent. Linear predictor is a linear function of covariates:

$$\eta_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p$$

and the mean function is defined as,

$$\mu_i = h^{-1}(\eta_i) = h^{-1}(\boldsymbol{x_i^{'}\beta})$$

where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip})^{'}$ is the p-dimensional vector of covariates and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients [2]. $h(.)$ is the link function that specifies the relation between the linear predictor and the dependent variable. Regression parameters are usually estimated using maximum likelihood (ML) estimation method.

2.1.1. ***Models for claim frequency.*** Poisson and negative binomial models are the basic models for the claim frequency data. In recent years zero-inflated models and hurdle models are also preferable to accommodate excess zero counts. In zero-inflated models, zero counts may arise from both the point mass and the count component. Hurdle models assume that zeros come from the point mass and the hurdle component models zero versus positive counts [32]. These models are briefly defined as follows.

Poisson model. Let n_i be the observed number of claims of policyholder i. Given $\boldsymbol{x_i}$, probability function of Poisson random variable can be defined as,

$$P(N_i = n_i) = \frac{\exp(-\mu_i)\mu_i^{n_i}}{\Gamma(1+n_i)}.$$

For the logarithmic link function, the mean parameter is defined as,

$$E(N_i \mid \boldsymbol{x_i}) = \mu_i = exp(\boldsymbol{x_i^{'}\beta}).$$

Negative Binomial model. Negative binomial distribution is a mixture of Poisson and gamma distributions. Probability function with dispersion parameter α and mean parameter μ_i is given by,

$$P(N_i = n_i) = \frac{\Gamma(\alpha^{-1}+n_i)}{\Gamma(\alpha^{-1})\Gamma(n_i+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{n_i}.$$

Zero-inflated models. Probability function of zero-inflated distribution is defined as,

$$P(N_i = n_i) = \begin{cases} \pi_i + (1-\pi_i)f_2(0) & \text{if } n_i = 0 \\ (1-\pi_i)f_2(n_i) & \text{if } n_i > 0 \end{cases}$$

where $f_2(n_i)$ is the base count density of a distribution such as negative binomial and Poisson. π_i is the component that inflates the probability of a zero and can be estimated by probit or logit model. For the logistic regression, $\pi_i = \dfrac{exp(\boldsymbol{x_i^{'}\beta})}{1+exp(\boldsymbol{x_i^{'}\beta})}$.

Hurdle models. Probability function of hurdle distribution is defined as,

$$P(N_i = n_i) = \begin{cases} f_1(0) & \text{if } n_i = 0 \\ \dfrac{1-f_1(0)}{1-f_2(0)} f_2(n_i) & \text{if } n_i > 0 \end{cases}$$

where $f_2(n_i)$ is the count density function and $f_1(0)$ is the probability of observing a count of zero. $f_1(0)$ and $1-f_1(0)$ may be estimated by probit or logit model ([4],[7]).

2.1.2. ***Models for claim severity.*** Claim severity data is generally positive-valued and has a right skewed distribution. Generally there are two options to model these variables; using GLM with a distribution such as gamma, inverse Gaussian or transforming the data for normality and using normal linear model for the transformed data with lognormal distribution assumption [8]. The distribution function and the mean function for gamma distribution is given below [3],

$$f(y) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}y^{\alpha-1}e^{-\beta y}, \;\; E(Y) = \frac{\alpha}{\beta} = \exp(\boldsymbol{x^{'}\beta}).$$

2.2. Generalized linear mixed model. Generalized linear mixed model is an extension of GLMs to longitudinal, clustered or repeated measures data that includes subject-specific regression parameters ($\boldsymbol{b_i}$) in the linear predictor. Within-subject association between the repeated measures of the same subject is taken into account by the introduction of random effects. Regression coefficients in GLMM ($\boldsymbol{\beta}$) have subject-specific interpretation because they represent the impact of covariates on an individual's transformed mean response.

For $i = 1, 2, \cdots, K$ and $j = 1, 2, \cdots, n_i$ let Y_{ij} be the jth outcome measured for cluster (individual) i. $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i})^{'}$ is the response vector for individual i. Probability (density) function of Y_{ij} from exponential family distribution is given by,

$$f_i(y_{ij} \mid \boldsymbol{b_i}, \boldsymbol{\beta}, \phi) = \exp\{\phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi)\}$$

and the mean function is defined as,

$$h(E(Y_{ij} \mid \boldsymbol{b_i})) = \eta_{ij} = \boldsymbol{x_{ij}^{'}\beta} + \boldsymbol{z_{ij}^{'}b_i} \tag{2.2}$$

where $h(.)$ is a known link function, η_{ij} is the linear predictor, $\boldsymbol{x_{ij}}$ be the $px1$ dimensional vector of fixed covariates at measurement occasion j and $\boldsymbol{z_{ij}}$ be the $qx1$ vector of covariates corresponding to $qx1$ vector of random effects $\boldsymbol{b_i}$. $\boldsymbol{b_i}$ is assumed to be independent Gaussian vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$ and Y_{ij}'s are assumed to be independent conditional on $\boldsymbol{b_i}$ ([21], [31]). LMM is a special case of GLMM where the link function is the identity link function and the distribution of responses is assumed to be normal [11].

Conditional mean and variance are defined as [3],

$$E[Y_{ij} \mid \boldsymbol{b_i}] = u_{ij} = \psi^{'}(\theta_{ij})$$

$$\mathrm{Var}[Y_{ij} \mid \boldsymbol{b_i}] = \phi\psi^{''}(\theta_{ij}) = \phi\nu(u_{ij}).$$

Marginal mean in GLMM can be obtained by averaging conditional mean over the distribution of random effects as follows,

$$\begin{aligned} E(Y_{ij}) = \mu_{ij} &= E\{E(Y_{ij} \mid \boldsymbol{b_i})\} \\ &= E\{h^{-1}(\boldsymbol{x_{ij}^{'}\beta} + \boldsymbol{z_{ij}^{'}b_i})\} \\ &= \int_{-\infty}^{+\infty} h^{-1}(\boldsymbol{x_{ij}^{'}\beta} + \boldsymbol{z_{ij}^{'}b_i}) f(\boldsymbol{b_i}) \mathrm{d}\boldsymbol{b_i} \end{aligned} \tag{2.3}$$

where $f(\boldsymbol{b_i})$ is the probability density function of $\boldsymbol{b_i}$ ([11], [31]). Generally this expression has no closed-form solution. Zeger et al. [31] approximated the expression for the marginal mean in Eq. (2.3) for standard link functions when the distribution of random effects is Gaussian with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$. For the logarithmic link function marginal mean is expressed as,

$$\mu_{ij} = \exp(\boldsymbol{x_{ij}^{'}\beta} + \boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}/2).$$

ML estimation of parameters in GLMMs is generally computationally infeasible because of the multiple integral that has no closed-form expression. To approximate the likelihood, various methods are suggested such as penalized quasi-likelihood (PQL) [5], Laplace approximation [26] and Gaussian quadrature methods [25].

2.3. Marginal models and generalized estimating equations. Another extension of GLMs to longitudinal data is the marginal model. Marginal models are also referred as population-averaged models since the inferences are for the population instead of subjects. Mean depends on covariates not on random effects and previous responses. The parameters of marginal model are estimated by the method of generalized estimating equations. Using the notations of previous subsection, a marginal model can be specified in three parts:

- Relation between the conditional expectation of Y_{ij} and the covariates is defined as,

  $h(E(Y_{ij} \mid X_{ij})) = h(\mu_{ij}) = \eta_{ij} = \boldsymbol{x_{ij}'}\boldsymbol{\beta}$.
- Given the covariates conditional variance of Y_{ij} is defined as,

  $\mathrm{Var}(Y_{ij}) = \phi\nu(\mu_{ij})$

  where $\nu(\mu_{ij})$ is the known variance function and ϕ is the scale parameter.
- It is assumed that the within–subject association is a function of association parameters in addition to mean function (μ_{ij}) [11]

Liang and Zeger [19] introduced generalized estimating equation as an extension of GLMs to analyse longitudinal data. Rather than specifying the multivariate distribution of an individual's observations, GEE requires only specifying the first two moments. Estimates of regression parameters can be obtained by solving the GEE defined as,

$$\mathrm{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}} \right)' \mathrm{Var}(\boldsymbol{Y_i})^{-1} (\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0 \tag{2.4}$$

where $\boldsymbol{\mu_i}$ is the mean response function and $\mathrm{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is similar to the quasi-score function proposed by Wedderburn [29] with additional association parameters ($\boldsymbol{\alpha}$) [9]. Here $Var(\boldsymbol{Y_i}) = \boldsymbol{V_i}(\boldsymbol{\alpha})$ depends not only on $\boldsymbol{\beta}$ but also on $\boldsymbol{\alpha}$ as defined below,

$$\boldsymbol{V_i} = \boldsymbol{A_i}^{1/2} \boldsymbol{R_i}(\boldsymbol{\alpha}) \boldsymbol{A_i}^{1/2}$$

where $\boldsymbol{R_i}(\boldsymbol{\alpha})$ is the working correlation matrix and $\boldsymbol{A_i}$ is a diagonal matrix with $\mathrm{Var}(Y_{ij}) = \phi\nu(\mu_{ij})$ along the diagonal [11]. Covariance structure across time is nuisance in this method and consistency of fixed regression parameters' estimates depends only on correct specification of the mean not on the correct choice of working correlation matrix [19].

3. Prediction of Total Claim Amount

In this section, proposed predictive models is described and the closed-form predictor of total claim amount is formulated. Alternative predictive models are also given in this section.

3.1. Proposed predictive model. The objective is to obtain marginal predictions of annual total claim amounts by taking into account heterogeneity among policyholders and to model claim amounts in original scale for the avoidance of retransformation issues. Accordingly, GLMM with logarithmic link function is used for the amounts of loss per claim of policyholders. Dependency between the

frequency and severity components is allowed by using claim count as an explanatory variable in GLMM. Marginal mean under GLMM is obtained by averaging conditional mean over the distribution of random effects. Closed–form predictor of the total claim amount is obtained by using the estimated distribution of the claim count variable.

Let Y_{ij} denotes the *jth* claim amount of policyholder i. Using GLMM description given in Eq. (2.2) and conditional on number of claims, random effects and fixed effects, mean of the response Y_{ij} is defined as,

$$E(Y_{ij} \mid N_i > 0, \boldsymbol{X_i}, \boldsymbol{Z_i}, \boldsymbol{b_i}) = \exp(\boldsymbol{x_{ij}^{'}}\boldsymbol{\beta} + N_i\beta_N + \boldsymbol{z_{ij}^{'}}\boldsymbol{b_i}) \tag{3.1}$$

where $\boldsymbol{b_i} \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{X_i}$ is the matrix of explanatory variables associated with the fixed effects, $\boldsymbol{Z_i}$ is the matrix of explanatory variables associated with the random effects and β_N is the regression coefficient related to the claim number. Conditional mean defined in Eq. (3.1) is marginalized by averaging over the distribution of random effects to express marginal mean ([15], [31]),

$$\begin{aligned} E_{\boldsymbol{b_i}}[E(Y_{ij} \mid N_i, \boldsymbol{X_i}, \boldsymbol{Z_i}, \boldsymbol{b_i})] &= E(Y_{ij} \mid N_i, \boldsymbol{X_i}, \boldsymbol{Z_i}) \\ &= \exp\left(\boldsymbol{x_{ij}^{'}}\boldsymbol{\beta} + N_i\beta_N + \frac{1}{2}\boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}\right). \end{aligned} \tag{3.2}$$

It follows that the expected total claim amount becomes,

$$\begin{aligned} E(S_i \mid \boldsymbol{X_i}) &= E[N_i E(Y_{ij} \mid N_i, \boldsymbol{X_i}, \boldsymbol{Z_i}) \mid \boldsymbol{X_i}] \\ &= E[N_i \exp(\boldsymbol{x_{ij}^{'}}\boldsymbol{\beta}) \exp(N_i\beta_N) \exp\left(\frac{1}{2}\boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}\right) \mid \boldsymbol{X_i}] \\ &= \exp(\eta_{ij}) \exp\left(\frac{1}{2}\boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}\right) E[N_i \exp(N_i\beta_N) \mid \boldsymbol{X_i}] \\ &= \exp(\eta_{ij}) \exp\left(\frac{1}{2}\boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}\right) E[\frac{\partial}{\partial\beta_N} \exp(N_i\beta_N) \mid \boldsymbol{X_i}] \\ &= \exp(\eta_{ij}) \exp\left(\frac{1}{2}\boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}\right) \frac{\partial}{\partial\beta_N} E[\exp(N_i\beta_N) \mid \boldsymbol{X_i}] \\ &= \exp(\eta_{ij}) \exp\left(\frac{1}{2}\boldsymbol{z_{ij}^{'}}\mathbf{D}\boldsymbol{z_{ij}}\right) \frac{\partial}{\partial\beta_N} M_{N_i|\boldsymbol{X_i}}(\beta_N) \end{aligned} \tag{3.3}$$

where $\eta_{ij} = \boldsymbol{x_{ij}^{'}}\boldsymbol{\beta}$ is the fixed effects part of GLMM and $M_{N_i|\boldsymbol{X_i}}(\beta_N)$ is the moment generating function of count variable evaluated at β_N. When the parameter estimates are replaced into Eq. (3.3) the predictor for total claim amount becomes,

$$\hat{S}_i = \exp(\hat{\eta}_{ij}) \exp\left(\frac{1}{2}\boldsymbol{z_{ij}^{'}}\hat{\mathbf{D}}\boldsymbol{z_{ij}}\right) \frac{\partial}{\partial\hat{\beta}_N} M_{N_i|\boldsymbol{X_i}}(\hat{\beta}_N). \tag{3.4}$$

In this study only random intercepts are used to take into account heterogeneity. Therefore, $\boldsymbol{z_{ij}} = 1$. Random intercepts (b_i) are assumed to be normally distributed with mean 0 and variance σ_b^2. Then the predictor in Eq. (3.4) reduces to,

$$\hat{S}_i = \exp(\hat{\eta}_{ij}) \exp\left(\frac{1}{2}\hat{\sigma}_b^2\right) \frac{\partial}{\partial\hat{\beta}_N} M_{N_i|\boldsymbol{X_i}}(\hat{\beta}_N) \tag{3.5}$$

where $\exp\left(\frac{1}{2}\hat{\sigma}_b^2\right)$ is the impact of heterogeneity on expected total claim amount.

As indicated in [15], relation between the regression coefficients of marginalized GLMM and GLMM depends on the components of covariance matrix ($\boldsymbol{D}$). When there is just random intercept term as random effects in the model, intercept parameter of marginal model increase by $\left(\frac{1}{2}\hat{\sigma}_b^2\right)$ while all other regression coefficients remain unchanged.

3.2. Alternative models. When the marginal model (GEE) with logarithmic link function is used for the amounts of loss per claim and observed number of claims is included as an explanatory variable in the model, expected total claim amount becomes,

$$\begin{aligned} E(S_i \mid \boldsymbol{X_i}) &= E[N_i E(Y_{ij} \mid N_i, \boldsymbol{X_i}) \mid \boldsymbol{X_i}] \\ &= E[N_i \exp(\boldsymbol{x_{ij}'}\boldsymbol{\beta} + N_i\beta_N) \mid \boldsymbol{X_i}] \\ &= \exp(\eta_{ij}) E[N_i \exp(N_i\beta_N) \mid \boldsymbol{X_i}] \\ &= \exp(\eta_{ij}) E\left[\frac{\partial}{\partial\beta_N}\exp(N_i\beta_N) \mid \boldsymbol{X_i}\right] \\ &= \exp(\eta_{ij})\frac{\partial}{\partial\beta_N} M_{N_i|\boldsymbol{X_i}}(\beta_N). \end{aligned} \tag{3.6}$$

Eq. (3.6) is similar to mean aggregate loss cost equation derived by Schulz [28] that models both mean claim amount and claim count by GLM. Difference is that GLM is for cross-sectional data and GEE is for repeated measures data. According to Eq. (3.6) predictor for total claim amount becomes,

$$\hat{S}_i = \exp(\hat{\eta}_{ij})\frac{\partial}{\partial\hat{\beta}_N} M_{N_i|\boldsymbol{X_i}}(\hat{\beta}_N). \tag{3.7}$$

Lastly, the predictor of Frees et al. [13] that used LMM for $ln(y_{ij})$ is given. y_{ij} is defined as the amount of inpatient or outpatient expenditure in jth event of individual i. Frees et al. [13] introduced the predictor of annual expenditures as,

$$\hat{S}_i = \exp(\hat{b}_i + \boldsymbol{x_{2i}'}\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) M_{N_i}'(\hat{\beta}_N) \tag{3.8}$$

where N_i is the number of events for either inpatient stays or outpatient visits, $\boldsymbol{x}_{2i}$ is a vector of individual level explanatory variables, $b_i \sim N(0, \sigma_b^2)$ are the random effects and $M_{N_i}'(\hat{\beta}_N)$ is the derivative of moment generating function of N_i at $\hat{\beta}_N$ [13].

For the prediction of total claim amount, moment generating function of the claim count variable is necessary according to Eq. (3.5), Eq. (3.7) and Eq. (3.8). Moment generating function of the negative binomial distribution, Poisson distribution and zero-inflated distribution can be found in [13], [28] and [10], respectively. Moment generating function of hurdle distribution can be defined as in the following,

$$
\begin{aligned}
M_N(t) = E(e^{tN}) &= \sum_{k=0}^{\infty} e^{tk} P(N=k) \\
&= e^0 P(N=0) + \sum_{k=1}^{\infty} e^{tk} P(N=k) \\
&= f_1(0) + \sum_{k=1}^{\infty} e^{tk} \frac{1-f_1(0)}{1-f_2(0)} f_2(k) \\
&= f_1(0) + \frac{1-f_1(0)}{1-f_2(0)} \left[\sum_{k=1}^{\infty} e^{tk} f_2(k) + e^0 f_2(0) - e^0 f_2(0) \right] \\
&= f_1(0) + \frac{1-f_1(0)}{1-f_2(0)} \left[\sum_{k=0}^{\infty} e^{tk} f_2(k) - e^0 f_2(0) \right] \\
&= f_1(0) + \frac{1-f_1(0)}{1-f_2(0)} \left[M_2(t) - e^0 f_2(0) \right] \\
&= f_1(0) + \frac{1-f_1(0)}{1-f_2(0)} \left[M_2(t) - f_2(0) \right]
\end{aligned} \tag{3.9}
$$

where $M_2(t)$ is the moment generating function of the count variable [1].

4. Application

4.1. About the data. Empirical studies have been done with a private health insurance (yearly renewable) data set that belongs to a major insurance company in Turkey. The models are fitted using claims data of year 2010 and predictive accuracy of the models is assessed using year 2011 data. There are 21496 and 22057 policies that have started or are renewed in 2010 and 2011, respectively. All policy numbers belong to year 2010 are different from year 2011 policies. The term 'claim amount' stands for the amount paid to the policyholder under the terms of policy in return for the reported claim. Note that the deductibles and policy limits are ignored in this study.

Sample consists of individuals whose ages are between 18–70. Each policyholder has one year of exposure. Explanatory variables are age, gender, marital status, policy package of policyholder and the province in which the policyholder lives. All explanatory variables are categorical except age. Provinces are categorized according to the policy numbers. Policy packages are categorized according to the economical properties and benefit types. For the claim severity models, number of claims are also used as an explanatory variable. Table 1 shows the description of categorical explanatory variables.

Table 1. Categorical explanatory variables

Factor	Variable	Description
Policy package	Eco, only inpatient	3
	Eco, inpatient and outpatient	2
	Not eco, only inpatient	1
	Not eco, inpatient and outpatient	0
Province	Other	4
	Kocaeli,Bursa,Muğla,Antalya,Adana,Tekirdağ	3
	İzmir	2
	Ankara	1
	İstanbul	0
Gender	Female	1
	Male	0
Marital status	Married	1
	Single / widowed	0

4.2. Modeling claim frequency. Claim numbers of 2010 policyholders range between 0-113 with a mean value of 3.803. 44% of 2010 policyholders have zero claims. In order to accommodate excess zero counts, zero-inflated and hurdle models are used for the annual number of claims in addition to Poisson and negative binomial models. According to the Pearson dispersion statistic (7.22>1), Poisson model is overdispersed. Due to the overdispersion problem, negative binomial model is used for the count part of zero-inflated model and hurdle models. Logarithmic link function is used for Poisson GLM, NB GLM, count parts of zero-inflated negative binomial (ZINB) and hurdle NB models. Logit model is preferred for the zero components of ZINB and hurdle NB models.

Table 2 shows the results of count data models that are fitted to 2010 number of claims. According to the Akaike information criterion (AIC) values, Poisson model provides worst fit whereas zero-inflated negative binomial model provides best fit to the number of claims. Zero-augmented models (ZINB, hurdle NB) provide better fit than the traditional Poisson and NB models.

Based on the count parts of zero-augmented models, Poisson and NB model, parameter estimates have close values in general. Province, age and gender are statistically significant determinants of claim counts in all models. Although mean functions of four models are similar for count parts, zero-augmented models expand the mean function by altering the likelihood of zero counts.

Parameter estimates for zero parts of ZINB and hurdle NB model are quite different. The reason is that zero hurdle component estimates the probability of observing positive claim while zero-inflation component estimates the probability of observing a zero claim from the point mass [32]. According to the zero part results of ZINB, coefficients for Ankara, İzmir provinces and marriage are statistically significant. For the zero part of hurdle NB model, coefficients for the economical packages that provide both inpatient and outpatient benefits and the category 4 of provinces are not statistically significant (see Table 2).

Considering the results, Poisson model is not used for the prediction of year 2011 total claim costs due to the overdispersion problem and poor goodness-of-fit.

Table 2. Parameter estimates of regression models for claim frequency data

Parameter	Estimate (Std. Error)					
	Poisson	NB	ZINB		Hurdle NB	
			Count part	Zero part	Count part	Zero part
Intercept	1.217 (0.016)*	1.038 (0.049)*	1.356 (0.041)*	-1.056 (0.098)*	1.460 (0.041)*	0.278 (0.065)*
Not eco, inpatient	-2.334 (0.023)*	-2.379 (0.035)*	-2.400 (0.048)*	-0.092 (0.146)	-2.365 (0.047)*	-1.510 (0.039)*
Eco, inpatient&outpatient	-0.033 (0.012)*	-0.021 (0.037)	-0.025 (0.030)	-0.021 (0.064)	-0.040 (0.030)	0.035 (0.050)
Eco, inpatient	-2.785 (0.103)*	-2.803 (0.131)*	-2.993 (0.237)*	-0.849 (1.325)	-2.890 (0.232)*	-1.869 (0.136)*
Ankara	-0.225 (0.013)*	-0.236 (0.039)*	-0.149 (0.032)*	0.279 (0.066)*	-0.137 (0.033)*	-0.295 (0.051)*
İzmir	-0.305 (0.018)*	-0.334 (0.050)*	-0.280 (0.042)*	0.184 (0.090)*	-0.254 (0.043)*	-0.298 (0.064)*
Prov. category 3	-0.125 (0.014)*	-0.139 (0.042)*	-0.127 (0.034)*	0.043 (0.077)	-0.107 (0.035)*	-0.126 (0.055)*
Prov. category 4	-0.319 (0.024)*	-0.305 (0.066)*	-0.320 (0.056)*	-0.011 (0.129)	-0.317 (0.057)*	-0.156 (0.083)
Age	0.005 (0.000)*	0.009 (0.001)*	0.009 (0.001)*	-0.001 (0.002)	0.006 (0.001)*	0.009 (0.001)*
Female	0.336 (0.007)*	0.364 (0.022)*	0.351 (0.018)*	-0.040 (0.041)	0.328 (0.019)*	0.210 (0.030)*
Married	0.024 (0.007)*	0.005 (0.023)	0.109 (0.018)*	0.360 (0.042)*	0.126 (0.019)*	-0.250 (0.030)*
AIC	168824	94331	**92928**		**92965**	

*Statistically significant at 5% level

4.3. Modeling claim severity. Claim amounts per claim of 2010 policyholders range between 1-108268 Turkish lira. Mean, median and standard deviation is 330, 101 and 1620 respectively. Because of right skewed distribution of claim amounts, GLM for mean claim amounts; GLMM and GEE for the amounts of loss per claim are used with gamma distribution assumption. Number of claims is included as a weight in GLM for mean claim amounts. Logarithmic transformation is also applied to the amounts of loss per claim and LMM is used with normality assumption. Exchangeable correlation structure is assumed for GEE and the parameters of GLMM is estimated by PQL method. Results of GLM, GLMM and GEE fits to 2010 claim severity data are given in Table 3.

Table 3 shows that the parameter estimates of three models have close values in general. Number of claims is statistically significant in all models and it has a positive effect on claim severity. While package, age and marital status are statistically significant determinants of claim severity, province factors except category 4 are not statistically significant in all models. Estimate of correlation parameter is 0.0619 and 0.0793 according to GEE and GLMM, respectively. The correlation among the claim amounts of policyholders is very low for this data set. Estimate of variance component for random effects in GLMM ($\hat{\sigma}_b^2$) is 0.2943.

4.4. Prediction results. Point predictions for the total claim amounts of 2011 policyholders are obtained in this part. 36% of 22057 policyholders had zero claims in 2011. External model validation method is used to compare the predictive accuracy of proposed predictive model with the alternative models. Parameter estimates obtained from the models fit to 2010 data and the information of 2011

Table 3. Parameter estimates of regression models for claim severity data

Parameter	Estimate (Std. Error)		
	GLM	**GEE-Exc.**	**GLMM**
Intercept	4.866 (0.073)*	4.831 (0.080)*	4.622 (0.041)*
Claim number	0.014 (0.001)*	0.013 (0.002)*	0.013 (0.001)*
Not eco, inpatient	1.964 (0.106)*	1.944 (0.067)*	1.975 (0.049)*
Eco, inpatient&outpatient	-0.288 (0.052)*	-0.281 (0.045)*	-0.234 (0.029)*
Eco, inpatient	2.093 (0.463)*	2.072 (0.260)*	2.077 (0.203)*
Ankara	-0.113 (0.059)	-0.091 (0.049)	-0.001 (0.033)
İzmir	-0.072 (0.080)	-0.054 (0.094)	-0.046 (0.043)
Prov. category 3	-0.001 (0.063)	0.049 (0.079)	-0.039 (0.035)
Prov. category 4	0.253 (0.109)*	0.296 (0.112)*	0.213 (0.059)*
Age	0.015 (0.002)*	0.016 (0.002)*	0.011 (0.001)*
Female	-0.124 (0.033)*	-0.087 (0.036)*	-0.002 (0.018)
Married	0.070 (0.032)*	0.074 (0.035)*	0.041 (0.018)*

*Statistically significant at 5% level

policyholders are used to obtain point predictions. Two criteria are used to compare the models, namely root mean squared error (RMSE) and mean absolute error (MAE) that are defined respectively,

$$RMSE = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(S_i - \hat{S}_i)^2}$$

$$MAE = \frac{1}{K}\sum_{i=1}^{K} \mid S_i - \hat{S}_i \mid .$$

For the predictive model in which LMM is used for the log-transformed claim amounts, only negative binomial model is used for the number of claims. The predictors of random intercepts in LMM are assumed zero as all policy numbers of 2010 are different from policy numbers of 2011. Results are given in Table 4.

Table 4 shows that the two predictive models using GLM for mean claim amounts and zero-augmented models for the number of claims have the lowest RMSE. Two GEE models follow these models. In terms of MAE, proposed predictive models that take into account between-individual heterogeneity perform the best. Especially in terms of MAE, predictive models that use zero-augmented models for the number of claims perform better than the models that use traditional NB model. Predictive model that uses LMM for log-transformed claim amounts per claim and NB model for the number of claims performs worst in terms of both RMSE and MAE. The predictive models that use GEE and GLM for claim severity have close RMSE and MAE values in accordance with the model fit results.

Table 4. Comparison of predictive models

Predictive Model	**RMSE**	**MAE**
$\overline{Y}_i$ GLM+N_i NB	5728.490	1735.818
$\overline{Y}_i$ GLM+N_i ZINB	**5726.671**	1704.296
$\overline{Y}_i$ GLM+N_i Hurdle NB	**5726.702**	1704.875
ln (Y_{ij}) LMM+N_i NB	5797.197	2197.447
Y_{ij} GEE+N_i NB	5728.624	1732.412
Y_{ij} GEE+N_i ZINB	**5727.021**	1705.120
Y_{ij} GEE+N_i Hurdle NB	**5727.039**	1705.644
Y_{ij} GLMM+N_i NB	5738.657	**1607.140**
Y_{ij} GLMM+N_i ZINB	5739.528	**1587.891**
Y_{ij} GLMM+N_i Hurlde NB	5739.819	**1588.590**

5. Conclusion

This study suggests an alternative predictor for the total claim amount by using marginalized GLMM for the amount of loss per claim of policyholders. Suggested predictive model is useful where insureds have a tendency to have multiple claim events within one policy period and when the interest is on obtaining the marginal predictions by taking into account heterogeneity among individuals. With the proposed model, heterogeneity is introduced as a nuisance parameter that changes the interpretation and predictions.

Number of claims is used as an explanatory variable in claim severity models in order to model the dependence between two components of aggregate loss. According to the analysis results of the application study, number of claims has a significant positive effect on claim severity.

Zero-inflated models and hurdle models accommodate the excess zeros for claim frequency data. Analysis and prediction results show that zero-augmented models improve model fit for the annual number of claims and predictions of total claim amounts.

Using MAE and RMSE criteria, predictive accuracy of proposed model is compared with alternative models. In terms of MAE, suggested predictive model performs best among the competing models. As a result, by taking into account both between-individual heterogeneity and excess zero counts, pricing process will be fulfilled efficiently without ignoring important characteristics of insurance claims data.

Acknowledgment

The first author has been supported by Scientific Research Projects Coordination Unit of Hacettepe University during her studies at the University of Padova. The authors would like to thank the anonymous referees for their valuable suggestions.

References

[1] Acar Şentürk, A. *Impact of heterogeneity on aggregate loss models in health insurance*, Ph.D. Thesis, Department of Actuarial Sciences, Hacettepe University, Turkey, 2016.

[2] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., Tahndi, N. *A Practitioner's guide to generalized linear models*, Casualty Actuarial Society 2004 Discussion Papers 1-116, Arlington, Virginia, 2004.

[3] Antonio, K. and Valdez, E. A. *Statistical concepts of a priori and a posteriori risk classification in insurance*, AStA Advances in Statistical Analysis, **96**(2), 187-224, 2012.

[4] Boucher, J. P., Denuit, M., Guillén, M. *Risk classification for claim counts:A comparative analysis of various zero-inflated mixed poisson and hurdle models*, North American Actuarial Journal, **11**(4), 110-131, 2007.

[5] Breslow, N.E. and Clayton, D.G. *Approximate inference in generalized linear mixed models*, Journal of the American Statistical Association, **88**(421), 9-25,1993.

[6] Brockman, M. J. and Wright, T. S. *Statistical motor rating: Making effective use of your data*, Journal of the Institute of Actuaries, **119**(03), 457-543, 1992.

[7] Cameron, C. A. and Trivedi, P. *Regression analysis of count data*, Econometric Society Monographs, 2nd edn, (New York: Cambridge University Press, 2013).

[8] De Jong, P. and Heller, G. Z. *Generalized linear models for insurance data*, (Cambridge University Press, 2008).

[9] Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S. L. *Analysis of longitudinal data*, 2nd edn., (New York: Oxford University Press, 2002).

[10] Edwin, T. K. *Power series distributions and zero-inflated models*, Doctoral Dissertation, University of Nairobi, 2014.

[11] Fitzmaurice, G. M., Laird, N. M., Ware, J. H. *Applied longitudinal analysis*, (New Jersey: John Wiley & Sons, 2004).

[12] Flynn, M. and Francis, L. A. *More flexible glms zero-inflated models and hybrid models*, Casualty Actuarial Society, 148-224, 2009.

[13] Frees, E. W., Gao, J., Rosenberg, M. A. *Predicting the frequency and amount of health care expenditures*, North American Actuarial Journal, **15**(3), 377-392, 2011.

[14] Fu, L. and Moncher, R. B. *Severity distributions for GLMs: Gamma or lognormal? Evidence from Monte Carlo simulations*, Casualty Actuarial Society Discussion Paper Program, 149-230, 2004.

[15] Grömping, U. *A note on fitting a marginal model to mixed effects log-linear regression data via GEE*, Biometrics, **52**(1), 280-285, 1996.

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

[17] Klugman, S. A., Panjer, H. H., Willmot, G. E. *Loss models: From data to decisions*, 2nd edn., (New York: John Wiley & Sons, 2004).

[18] Lambert, D. *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics, **34**(1), 1-14, 1992.

[19] Liang, K. Y. and Zeger S. L. *Longitudinal data analysis using generalized linear models*, Biometrika, **73**(1), 13-22, 1986.

[20] McCullagh, P. and Nelder, J. A. *Generalized linear models*, 2nd edn., (London: Chapman and Hall, 1989).

[21] Molenberghs, G. and Verbeke G. *Models for discrete longitudinal data*, (New York: Springer, 2005).

[22] Mouatassim, Y., Ezzahid, E. H. *Poisson regression and zero-inflated Poisson regression: Application to private health insurance data*, European Actuarial Journal, **2**(2), 187-204, 2012.

[23] Mullahy, J. *Specification and testing of some modified count data models*, Journal of Econometrics, **33**(3), 341-365, 1986.

[24] Ohlsson, E. and Johansson, B. *Non-life insurance pricing with generalized linear models*, (Berlin: Springer, 2010).

[25] Pinheiro, J.C. and Bates, D.M. *Approximations to the loglikelihood function in the nonlinear mixed-effects model*, Journal of Computational and Graphical Statistics, **4**(1), 12-35, 1995.

[26] Raudenbush, S. W., Yang, M. L., Yosef, M. *Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation*, Journal of Computational and Graphical Statistics, **9**(1), 141-157, 2000.

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

[28] Schulz, J. *Generalized linear models for a dependent aggregate claims model*, Master's Thesis, Department of Mathematics and Statistics, Concordia University, Montreal, Canada, 2013.

[29] Wedderburn, R. W. M., *Quasi-likelihood function, generalized linear models and the Gauss-Newton method*, Biometrika, **61**(3), 439-447, 1974.

[30] Yip, K. C. H. and Yau, K. K. W. *On modeling claim frequency data in general insurance with extra zeros*, Insurance: Mathematics and Economics, **36**(2), 153-163, 2005.

[31] Zeger, S. L., Liang, K. Y., Albert, P. S. *Models for longitudinal data: A generalized estimating equation approach*, Biometrics, **44**(4), 1049-1060, 1988.

[32] Zeileis, A., Kleiber, C., Jackman, S. *Regression models for count data in R*, Journal of Statistical Software, **27**(8), 1-25, 2008.