Embedded predictive analysis of misrepresentation risk in GLM ratemaking models

Michelle Xia[1], Lei Hua*[2], and Gary Vadnais[3]

[1,2]Division of Statistics, Northern Illinois University, Dekalb, IL, 60115, U.S.A.

[3]Intact Insurance, Saint Hyacinthe, QC, J2S3BB, Canada

Abstract. Misrepresentation is a type of insurance fraud that happens frequently in policy applications. Due to the unavailability of data, such frauds are usually expensive or difficult to detect. Based on the distributional structure of regular ratemaking data, we propose a generalized linear model (GLM) framework that allows for an embedded predictive analysis on the misrepresentation risk. In particular, we treat binary misrepresentation indicators as latent variables under GLM ratemaking models, for rating factors that are subject to misrepresentation. Based on a latent logistic regression model on the prevalence of misrepresentation, the model identifies characteristics of policies that are subject to a high risk of misrepresentation. The method allows for multiple factors that are subject to misrepresentation, while accounting for other correctly measured risk factors. Based on the observed variables on the claim outcome and rating factors, we derive a mixture regression model structure that possesses identifiability. The identifiability ensures valid inference on the parameters of interest, including the rating relativities and the prevalence of misrepresentation. The usefulness of the method is demonstrated by simulation studies, as well as a case study using the Medical Expenditure Panel Survey data.

Key words. Misrepresentation, ratemaking, predictive analysis, generalized linear models, Bayesian inference, Markov chain Monte Carlo.

*Corresponding author; Email: Lhua@niu.edu

1 Introduction

For property and casualty insurance, ratemaking models are often determined using historical claim data based on rating factors that are predictive for loss severity and frequency. We refer to Klein et al. [2014], Bermúdez and Karlis [2015], David [2015], Hua [2015], Shi [2016] for references in this regard. For example, auto insurance rates are usually calculated using risk factors such as use of vehicle, annual mileage, claim and conviction history, age, gender and credit history of the insured or the applicant (Lemaire et al. [2016]). Due to the financial incentives, policy applicants may have a motivation to provide false statements on the risk factors. This type of fraud occurring on the policy application is referred to as insurance *misrepresentation* (Winsor [1995]). In auto insurance, information regarding risk factors such as use of vehicle and annual mileage is generally hard to obtain. Even for insurers that offer a voluntary discount on mileage tracking, the existence of anti-selection limits the capability of such programs in obtaining accurate information on the whole book of policies.

Misrepresentation is one particular cause of *misclassification*, the type of measurement error in binary or categorical variables, when the variable is recorded in a wrong category. Xia and Gustafson [2016], Sun et al. [2017], Xia and Gustafson [2017] used the term *unidirectional misclassification* for situations like misrepresentation where the error occurs only in the direction that is more favorable to the respondent. When there are misclassification errors in some of the variables, the general difficulty of modeling is the *unidentifiability* of the parameters. Model unidentifiability is a situation where the likelihood function possesses multiple global maximums. In health and accounting areas, Gustafson [2014] and Hahn et al. [2016] studied two cases of partially identified models arising from unidirectional misclassification, where some parameters of interest can not be estimated consistently. For these models, we may only be able to obtain a range of values for the partially identified parameters, even with an infinite amount of data.

Proposed by Brockman and Wright [1992], GLM ratemaking models are now very popular in property and casualty insurance areas (see, e.g., Haberman and Renshaw [1996]) . In GLM ratemaking models, possible misrepresentation in a rating factor is expected to cause an attenuation bias in the estimated risk effect (e.g., the relativity), resulting in an underestimation of the difference between the true positive and true negative groups. Despite discussions on operational changes in order to discourage such fraudulent behaviors, there seems to be little work concerning statistical models for predicting or evaluating the risks or expenses associated with such fraud. In Xia and Gustafson [2016], the authors studied the structure of GLM ratemaking models with a binary covariate (e.g., risk factor) that is subject to misrepresentation. The study revealed that all parameters are theoretically identifiable for GLM ratemaking models with common distributions assumed for claim severity and frequency. This suggests that we can estimate the rating relativities consistently, using regular ratemaking data. However, the study was based on a simplified situation where there is a single risk factor subject to misrepresentation, without including other risk factors.

In the current paper, we use regular ratemaking data to develop GLM ratemaking models that embed predictive analyses of the misrepresentation risk. The particular extensions include: (1) the adjustment for other correctly measured risk factors in the ratemaking model; (2) the incorporation of multiple rating factors that are subject to misrepresentation; (3) the embedding of a latent logistic model for how risk factors affect the prevalence of misrepresentation, which enables us to perform a predictive analysis (see, e.g., Frees et al. [2014], Shi and Valdez [2011]) on the misrepresentation risk. We derive mixture regression structures for the conditional distribution of the observed variables, confirming the identifiability of the models. We perform simulation studies based on finite samples, show that we can consistently estimate the model parameters including the rating relativities, and interpret how risk factors affect the prevalence of misrepresentation. The case study using the 2013 Medical Expenditure Panel Survey (MEPS) data illustrates the use of the proposed method in an insurance application.

The proposed model is expected to have an immediate impact on the actuarial practices of the insurance industry. Based on the theoretical identification of the model, traditional GLM ratemaking models can be extended to embed a predictive analysis of misrepresentation risk without requiring extra data on the misrepresentation itself. The analysis will provide information on the characteristics of the insured individuals or applicants who are more likely to misrepresent on self-reported rating factors. With predictive models that automatically update with new underwriting data, the prevalence of misrepresentation can be predicted at the policy level, during policy underwriting based on various risk characteristics. Therefore, underwriting interventions can be undertaken in order to minimize the occurrence of misrepresentation. In addition, the claims department may use the risk profiles to help identify fraudulent claims.

2 A GLM ratemaking model with misrepresentation

In Xia and Gustafson [2016], the authors studied the identification of a GLM when a covariate (e.g., a rating factor) is subject to misrepresentation. For misrepresentation in a binary rating factor (e.g., smoking status), we can formulate the model as follows. Denote by V and V^* the true and observed binary risk status, respectively. There is a chance for misrepresentation to occur if the individual has a positive risk status. In particular, we can write the conditional probabilities as

$$\begin{aligned} \mathrm{P}(V^* = 0 \,|\, V = 0) &= 1 \\ \mathrm{P}(V^* = 0 \,|\, V = 1) &= p, \end{aligned} \tag{2.1}$$

where we call p the *misrepresentation probability.*

Denote $\theta = \mathrm{P}(V = 1)$, the true probability of a positive risk status. We can derive the observed probability of a positive risk status as $\theta^* = \mathrm{P}(V^* = 1) = \mathrm{P}(V^* = 1 \,|\, V = 0)\mathrm{P}(V = 0) + \mathrm{P}(V^* = 1 \,|\, V = 1)\mathrm{P}(V = 1) = \theta(1 - p)$. In insurance applications, a quantity of interest is $q = \mathrm{P}(V = 1|V^* = 0)$, the percentage of reported negatives that corresponds to a misrepresented true positive risk status. We define q as the *prevalence of misrepresentation.* Using the Bayes's Theorem, the prevalence of misrepresentation can be obtained as

$$\begin{aligned} \mathrm{P}(V = 1|V^* = 0) &= \frac{\mathrm{P}(V^* = 0 \,|\, V = 1)\mathrm{P}(V = 1)}{\mathrm{P}(V^* = 0)} \\ &= \frac{\theta p}{1 - \theta(1 - p)} = q. \end{aligned}$$

Similarly, we can obtain $p = (1-\theta)q/[\theta(1-q)]$. Note, we can derive one conditional probability from the other, along with an estimate of the observed probability θ^* using samples of V^*. Note that the prevalence of misrepresentation q represents the percentage of misrepresented cases among applicants who reported a negative risk status. It quantifies the misrepresentation risk of a particular application, and thus determines the total number of misrepresented cases in the book of business.

In a GLM ratemaking model, the model considers the mean of a response variable Y from a loss or count distribution, conditioning on the true risk status V. In Xia and Gustafson [2016], the authors showed that the conditional distribution of the observed variables $(Y \,|\, V^* = 0)$ is a mixture of the two distributions of $(Y \,|\, V = 0)$ and $(Y \,|\, V = 1)$, and $(Y \,|\, V^* = 1)$ has the same distribution as $(Y \,|\, V = 1)$. Since one of the two mixture components can be informed from data with $V^* = 1$, the mixture model possesses the model identifiability in a general case when the response variable is non-binary. Furthermore, the authors showed the moment identifiability of the model by deriving the observable moments corresponding to Y and V^*. In an identified model, the likelihood function possesses one unique global maximum, allowing consistent estimation of all parameters including p and q. This means we can consistently estimate the true relativity and the probability p with regular ratemaking data. Here, we will extend the work under the GLM ratemaking framework in three directions: (a) to include additional rating factors that are correctly measured, a situation we commonly face in insurance ratemaking; (b) to simultaneously include multiple rating factors that are subject to misrepresentation; and (c) to relax the assumption that the prevalence of misrepresentation does not change with values of rating factors.

2.1 Model with correctly measured risk factors

In the GLM ratemaking model, we first assume that the loss outcome Y depends on the true status V. In order to formulate the problem for the first extension, we use $\mathbf{x} =(X_1, X_2, \cdots, X_K)$ to denote K correctly measured rating factors that are predictive of the loss outcome Y.

The GLM ratemaking model can be written as

$$g(\mu) = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_K X_K + \alpha_{K+1} V,$$

where $\mu = \mathrm{E}(Y)$ and $\mathrm{Var}(Y) = \varphi \mathrm{V}(\mu)$, with φ being the dispersion parameter and $\mathrm{V}(\cdot)$ being the variance function. Denote $\boldsymbol{\alpha}$ =(α_0, α_1, $\cdots$, α_{K+1}), and let $\boldsymbol{\varphi}$ denote all other parameters including the dispersion parameter. We may use $f_Y(y \,|\, \boldsymbol{\alpha}, \boldsymbol{\varphi}, V, \mathbf{x})$ to denote the conditional distribution function of $(Y \,|\, V, \mathbf{x})$ from the exponential family.

Assume that the misrepresentation is *non-differential* on Y (i.e., $Y \perp V^* \,|\, V, \mathbf{x}$). That is, the outcome Y does not depend on whether the applicant misrepresents on the risk factor, given the true status V and other risk factors $\mathbf{x}$. We can obtain the conditional distribution of the observed variables, $(Y \,|\, V^*, \mathbf{x})$, as

$$\begin{aligned} f_Y(y \,|\, V^* = 1, \mathbf{x}) &= f_Y(y \,|\, \boldsymbol{\alpha}, \boldsymbol{\varphi}, V = 1, \mathbf{x}) \\ f_Y(y \,|\, V^* = 0, \mathbf{x}) &= q\, f_Y(y \,|\, \boldsymbol{\alpha}, \boldsymbol{\varphi}, V = 1, \mathbf{x}) + (1-q) f_Y(y \,|\, \boldsymbol{\alpha}, \boldsymbol{\varphi}, V = 0, \mathbf{x}). \end{aligned} \tag{2.2}$$

Given fixed values of $\mathbf{x}$, the conditional distribution of $(Y \,|\, V^*, \mathbf{x})$ is a mixture of two distributions when $V^* = 0$, and it is a single distribution when $V^* = 1$. The mixture model in (2.2) with additional covariates $\mathbf{x}$ is called a *mixture regression model.* Such mixture models have been shown to possess identifiability, given that a mixture distribution is identifiable for the specific component distribution (see, e.g., Hennig [2000], and Grün and Leisch [2008b]). Owing to the identifiability for mixtures of distributions from the exponential family (Atienza et al. [2006, 2007]), the model in (2.2) will be fully identifiable for common loss severity and frequency distributions such as the gamma, Poisson and negative binomial distributions considered under the GLM ratemaking context. Hence, using regular ratemaking data containing $(Y, V^*, \mathbf{x})$, we can consistently estimate the mixture weight q (i.e., the prevalence of misrepresentation), and the regression coefficients in $\boldsymbol{\alpha}$ (and the corresponding relativities).

In order to understand the model identification, we use a hypothetical example to visualize the conditional distribution of $(Y \,|\, V^*, \mathbf{x})$ under the mixture regression context. For better visualization of the mixture structure, we assume that the medical loss amount Y follows a lognormal distribution, with V^* being the smoking status. Given $\mathbf{x} = \boldsymbol{x}_0$ (i.e., assuming that the comparison is among individuals with the same other risk factors), $(\log(Y) \,|\, \mathbf{x} = \boldsymbol{x}_0)$ will have a mixture of two normal distributions for individuals who reported non-smoking, while it will have a normal distribution for those who reported smoking. Figure 1 gives an example on how the conditional distributions look like, with $\mathbf{x} = \boldsymbol{x}_0$ being fixed for both groups. For the dashed density, the two mixture components are the conditional distributions for the true non-smokers and smokers, and the prevalence of misrepresentation is the mixture weight for the true smokers that has a higher mean. Note that in the mixture regression model, the two components are regression models that cannot be visually presented, although the model possesses the identifiability for estimating all parameters. This mixture regression structure allows us to estimate the mixture weight (the prevalence of misrepresentation), the true risk effect of smoking (i.e., the difference in the mean of the two components), and the regression coefficients associated with each of the risk factors in $\mathbf{x}$, without observing V.

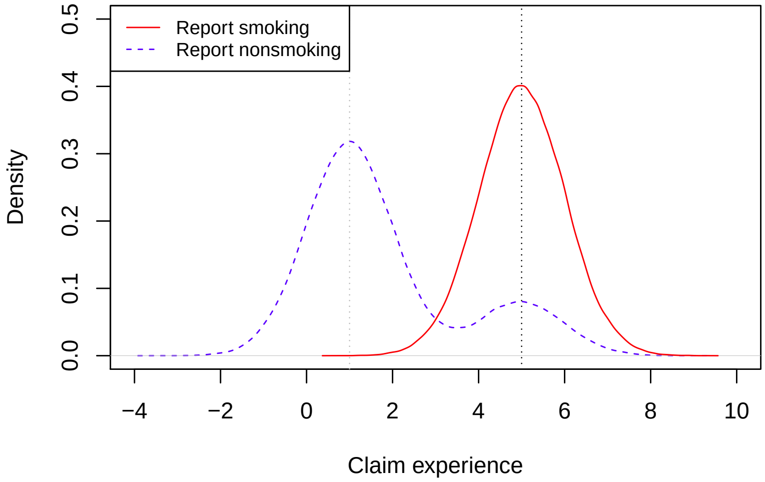


Figure 1: Logarithm of loss amount $\log(Y)$ by reported smoking status V^* under lognormal ratemaking models, when comparing individuals with same other risk characteristics $\mathbf{x}$.

Example 2.1 (Gamma model)**.** First, we give a simplified example of a gamma loss severity model commonly used for auto insurance. Denote by Y the amount of a liability loss for a given claim, by X the annual mileage traveled by the vehicle, and by V^* the observed risk status on vehicle use status (e.g., business or not) that is subject to misrepresentation. We can use a gamma GLM severity model given by

$$\begin{aligned}(Y \,|\, V,\, X) &\sim gamma\,(\varphi,\, \mu_{V,\,X})\\ \log(\mu_{V,\,X}) &= \alpha_0 + \alpha_1 V + \alpha_2 X\\ (V^* \,|\, V,\, X) &\sim Bernoulli((1-p)V),\end{aligned} \tag{2.3}$$

where φ is the scale parameter, and $\mu_{V,\,X}$ is the conditional mean of the gamma distribution given V and X. Note that the conditional distribution of the observed variables $(Y|V^*, \mathbf{x})$ have the same form as that in Equation (2.2). For this example, the conditional distribution $f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V,\, \mathbf{x})$ takes the form of the above gamma distribution, with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ and $\boldsymbol{\varphi} = \varphi$. Note that p and q are parameters of the model, for which we can perform inference using either frequentist or Bayesian approaches. For Bayesian methods, the posterior distributions of $(p|V^*, \mathbf{x})$ and $(q|V^*, \mathbf{x})$ do not have closed-forms. Hence, we will need to use Markov chain Monte Carlo (MCMC) techniques in order to make inference on the parameters.

2.2 Multiple risk factors with misrepresentation

For the second extension, we denote $\mathbf{v}$ =$(V_1,\, V_2,\, \cdots,\, V_J)$ as the true status of J rating factors that are subject to misrepresentation, and $\mathbf{v}^*$ =$(V_1^*,\, V_2^*,\, \cdots,\, V_J^*)$ as the corresponding observed values for these rating factors. We may assume that the response variable Y depends on the rating factors $\mathbf{v}$ through a parameter vector $\boldsymbol{\alpha}$ (e.g., one intercept and J regression coefficients). We further assume that $(Y \,|\, \mathbf{v})$ has the probability function $f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, \mathbf{v})$. Under the assumption of non-differential misclassification, the conditional distribution of $(Y \,|\, \mathbf{v}^*)$ will either be a single distribution when $\mathbf{v}^* = (1, 1, \cdots, 1)$, or a mixture distribution with the number of components and the mean of components determined by the values of the observed $\mathbf{v}^*$. For example, when there are two rating factors with misrepresentation (i.e., $\mathbf{v}$ =$(V_1,\, V_2)$), we can write the conditional distribution of observed variables, $(Y \,|\, \mathbf{v}^{*})$, as

$$\begin{aligned}
f_Y(y \,|\, V_1^* = 1,\, V_2^* = 1) =& f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 1,\, V_2 = 1)\\
f_Y(y \,|\, V_1^* = 0,\, V_2^* = 1) =& q_1\, f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 1,\, V_2 = 1) + (1 - q_1) f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 0,\, V_2 = 1)\\
f_Y(y \,|\, V_1^* = 1,\, V_2^* = 0) =& q_2\, f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 1,\, V_2 = 1) + (1 - q_2) f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 1,\, V_2 = 0)\\
f_Y(y \,|\, V_1^* = 0,\, V_2^* = 0) =& q_3\, f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 1,\, V_2 = 1) + q_4 f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 0,\, V_2 = 1)\\
&+ q_5 f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 1,\, V_2 = 0) + (1 - q_3 - q_4 - q_5) f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1 = 0,\, V_2 = 0),
\end{aligned} \tag{2.4}$$

where the corresponding prevalence of misrepresentation $q_1 = \mathrm{P}(V_1 = 1,\, V_2 = 1 \,|\, V_1^* = 0,\, V_2^* = 1)$, $q_2 = \mathrm{P}(V_1 = 1,\, V_2 = 1 \,|\, V_1^* = 1,\, V_2^* = 0)$, $q_3 = \mathrm{P}(V_1 = 1,\, V_2 = 1 \,|\, V_1^* = 0,\, V_2^* = 0)$, $q_4 = \mathrm{P}(V_1 = 0,\, V_2 = 1 \,|\, V_1^* = 0,\, V_2^* = 0)$ and $q_5 = \mathrm{P}(V_1 = 1,\, V_2 = 0 \,|\, V_1^* = 0,\, V_2^* = 0)$.

In order to simplify the model, we may assume that there is no correlation in the two risk factors, and the occurrence of misrepresentation in one risk factor does not depend on the value or the misrepresentation status of the other. That is, we have $V_1 \perp V_2$, $\mathrm{P}(V_1^* = 0 \,|\, V_1 = 1,\, V_2) = \mathrm{P}(V_1^* = 0 \,|\, V_1 = 1) = p_1$, $\mathrm{P}(V_2^* = 0 \,|\, V_1,\, V_2 = 1) = \mathrm{P}(V_2^* = 0 \,|\, V_2 = 1) = p_2$, and $\mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = 1,\, V_2 = 1) = p_1 p_2$. Denote $\mathrm{P}(V_1 = 1) = \theta_1$ and $\mathrm{P}(V_2 = 1) = \theta_2$. Using Bayes' Theorem, we derive specific forms of the prevalence of misrepresentation, the q_j's, in the Appendix. In particular, q_1 and q_2 have the same form as q in Equation (2.2). That is, $q_j = \theta_j p_j / [1 - \theta_j (1 - p_j)]$, $j = 1, 2$. In addition, we have $q_3 = p_1 p_2 \theta_1 \theta_2 / [p_1 p_2 \theta_1 \theta_2 + p_1 \theta_1 (1 - \theta_2) + p_2 (1 - \theta_1) \theta_2 + (1 - \theta_1)(1 - \theta_2)]$, $q_4 = p_2 (1 - \theta_1) \theta_2 / [p_1 p_2 \theta_1 \theta_2 + p_1 \theta_1 (1 - \theta_2) + p_2 (1 - \theta_1) \theta_2 + (1 - \theta_1)(1 - \theta_2)]$, and $q_5 = p_1 \theta_1 (1 - \theta_2) / [p_1 p_2 \theta_1 \theta_2 + p_1 \theta_1 (1 - \theta_2) + p_2 (1 - \theta_1) \theta_2 + (1 - \theta_1)(1 - \theta_2)]$.

An alternative to the above conditional independence assumption is the assumption on the prevalence of misrepresentation, the q_j's. For example, we can assume that the applicant has the same misrepresentation probability $p_1 = p_2$ for the two risk factors. When the true status is positive for both risk factors, it is reasonable to assume that there are only two possibilities on the misrepresentation status: the applicant either does not misrepresent on any risk factor, or misrepresents on both of them. That is, we have the prevalence of misrepresentation being $q_4 = q_5 = 0$. In such a case, the last mixture distribution in (2.4) will only have two components.

This will dramatically simplify the model when there are more than two risk factors subject to misrepresentation.

From (2.4), the conditional distribution still has a mixture regression structure. Based on the observed status of $\mathbf{v}^*$, we know the number of components in the mixture regression structure, as well as which components (corresponding to the possible true values of $\mathbf{v}$) each mixture contains. Owing to the identifiability of mixture regression models (Hennig [2000], Grün and Leisch [2008b]), the model in (2.4) is identifiable for the common loss severity and frequency distributions assumed in a GLM ratemaking model. This means all the regression coefficients in $\boldsymbol{\alpha}$ (and the corresponding relativities), and the mixture weights (i.e., prevalence of misrepresentation q_j, $j = 1, 2, \cdots, 5$) can be consistently estimated from regular ratemaking data containing (Y, $\mathbf{v}^*$). The results can be extended straightforwardly to the case with more than two rating factors subject to misrepresentation. Note that the distributional assumptions we adopted in the two previous paragraphs are for the purpose of obtaining analytical forms of the prevalence of misrepresentation q_j's. The identifiability of the model is obtained from the mixture structure we have in (2.4) and does not depend on the specific forms of q_j's.

Example 2.2 (Negative binomial model)**.** We give an example of a negative binomial loss frequency model commonly used for auto insurance. Denote by Y the number of liability losses for a given policy year, and by (V_1^*, V_2^*) the observed risk status on the binary vehicle use and binary annual mileage (e.g., on whether it is over a certain threshold) that are subject to misrepresentation. In particular, we can write the negative binomial GLM model as

$$\begin{aligned} (Y \mid V_1, V_2) &\sim negbin\,(\varphi, \mu_{V_1, V_2}) \\ \log(\mu_{V_1, V_2}) &= \alpha_0 + \alpha_1 V_1 + \alpha_2 V_2, \end{aligned} \tag{2.5}$$

where φ is a dispersion parameter, and μ_{V_1, V_2} is the conditional mean of the negative binomial distribution given the true statuses (V_1, V_2). Note that the conditional distribution of the observed variables $(Y|V_1^*, V_2^*)$ have the mixture structure given in Equation (2.4), and the prevalence of misrepresentation q_j's are the mixture weights that can be estimated using data on (Y, V_1^*, V_2^*). For this example, the conditional distribution $f_Y(y \,|\, \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V_1,\, V_2)$ takes the form of the above negative binomial distribution, with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ and $\boldsymbol{\varphi} = \varphi$.

2.3 Embedded predictive analysis on misrepresentation risk

The last extension is very important for understanding the characteristics of the insureds or applicants who are more likely to misrepresent on certain self-reported rating factors. In addition to the regression relationship we are assuming between the mean of the loss outcome and the rating factors, we may further assume that the prevalence of misrepresentation q depends on certain risk factors. Without loss of generality, we assume the case in Equation (2.2) where there is one variable V subject to misrepresentation. Like in the case of zero-inflated regression models (see, e.g., Yip and Yau [2005]), we may use a latent binary regression model for the relationship between the prevalence of misrepresentation and the rating factors. That is,

$$g(q) = \beta_0 + \mathbf{z}\boldsymbol{\beta}, \tag{2.6}$$

where the link function $g(\cdot)$ can either take the logit or probit form, β_0 is an intercept and the vector $\boldsymbol{\beta}$ contains the effects of the rating factors on the prevalence of misrepresentation. For the latent model in (2.6), it usually requires a larger sample size to learn parameters with the same precision. In order to simplify the model in real practices, we recommend choosing a subset of meaningful risk factors from $\mathbf{x}$ as $\mathbf{z}$, in the case where there are many rating factors available.

Example 2.3 (Poisson model)**.** We give an example of a Poisson loss frequency model commonly used for auto insurance. Denote by Y the number of liability losses for a given policy year, by V^* the observed risk status on the binary vehicle use status with possible misrepresentation, and by X the annual mileage. In particular, we can write the Poisson GLM model as

$$
\begin{aligned}
(Y \mid V,\, X) &\sim \mathit{Poisson}\,(\mu_{V,\,X}) \\
\log(\mu_{V,\,X}) &= \alpha_0 + \alpha_1 V + \alpha_2 X \\
\operatorname{logit}(q) &= \log\left(\frac{q}{1-q}\right) = \beta_0 + \beta_1 X,
\end{aligned}
\tag{2.7}
$$

where $\mu_{V,\,X}$ is the conditional mean of the Poisson distribution given the true status $(V,\, X)$. Note that the conditional distribution of the observed variables $(Y|V^{*}, \mathbf{x})$ have the same form as that in Equation (2.2), except for the fact that q varies with the covariate X. Here, the logit model on q is a latent model that requires no additional information other than observed data on X. For this example, the conditional distribution $f_Y(y \mid \boldsymbol{\alpha},\, \boldsymbol{\varphi},\, V,\, \mathbf{x})$ takes the form of the above Poisson distribution, with $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)$ and there is no parameter $\boldsymbol{\varphi}$ needed for the Poisson model.

When there is a latent binomial or multinomial regression model (2.6) on the mixture weight q, the mixture in (2.2) is called *a mixture regression model with concomitant variables* (see, e.g., Grün and Leisch [2008a]). According to Hennig [2000] and Grün and Leisch [2008b], such a mixture regression model is identifiable, provided that a simple mixture of the component distributions is identifiable. The identifiability will ensure that we would be able to make inference on how different characteristics (e.g., demographics and other risk factors) affect the prevalence of misrepresentation q. Based on the conditional distribution of the observed loss outcomes and risk factors, the model will require no extra data on the misrepresentation itself.

Owing to the identifiability, we will be able to estimate the regression coefficients $\boldsymbol{\beta}$ for the latent model on the prevalence of misrepresentation using regular ratemaking data. The statistical significance of the variables indicates what characteristics of the applicants or insureds affect the probability that a reported negative risk status corresponds to a misrepresented true positive status. Using such estimated predictive models that automatically update with new loss experience data, the prevalence of misrepresentation can be predicted at the policy level, during policy underwriting based on various risk characteristics. The applications with a high predicted prevalence of misrepresentation can then be selected for an underwriting investigation on self-reported risk factors. For the claims department, the predicted prevalence of misrepresentation can be used as a flag for identifying potential fraudulent claims. Such a risk profile based on the predicted prevalence will be helpful for identifying misrepresentation on policy applications, as well as fraud in insurance claims.

3 Simulation studies

We perform simulation studies on the performance of the model under finite sample scenarios, in order to illustrate the model identifiability. In particular, we study the model performance under three scenarios where we have: (1) other correctly measured rating factors; (2) multiple rating factors subject to misrepresentation; and (3) a misrepresented rating factor with the prevalence of misrepresentation varying with other factors.

3.1 Model implementation

The proposed models given in Equations (2.2) to (2.7) all have tractable analytical forms. Hence, we can write out the full likelihood functions. For the model implementation, it is more convenient to work with the complete-data likelihood with a latent variable denoting the misrepresentation status.

Here, we use the gamma loss severity model in Example 2.1 to illustrate the implementation of the proposed models. Denote by $(y_1,\, v_1^*,\, x_1)$, $(y_2,\, v_2^*,\, x_2)$, $\cdots$, $(y_n,\, v_n^*,\, x_n)$ a random sample of observed variables of size n. Maximum likelihood estimation (e.g., based on the expectation maximization algorithm, EM, McLachlan and Krishnan [2007]) and Bayesian inference (based on MCMC) uses the complete-data likelihood that includes the *unobserved* (i.e., latent) status on the misrepresentation. For observations where $v_i^* = 0$, denote by z_i $(i = 1, 2, \cdots, n)$ the latent binary indicator on whether Observation i is misrepresented (whether the true status $v_i = 1$, i.e., whether Sample i is from the component distribution for true smokers). The introduction of the latent misrepresentation indicators leads to a multiplicative likelihood function convenient for obtaining the log-likelihood function for the EM or MCMC algorithm. In particular, the log-likelihood function can be written as

$$l(\alpha_0, \alpha_1, \alpha_2, \varphi) = \sum_{i=1}^{n} v_i^* \log \phi_2(y_i) + \sum_{i=1}^{n} (1 - v_i^*) \left\{ (1 - z_i) \log \left[(1 - q)\phi_1(y_i) \right] + z_i \log \left[q\phi_2(y_i) \right] \right\},$$

where $\phi_1(y_i) = f(y_i|v_i = 0,\, x_i)$ is the conditional distribution for the regression model of the true non-smokers, and $\phi_2(y_i) = f(y_i|v_i = 1,\, x_i)$ is that of the true smokers.

Note that regular GLM can be implemented either using the frequentist approach based on maximum likelihood estimation (MLE) or Bayesian inference based on MCMC. This is true for the proposed model as well. For the current paper, we use Bayesian inference that is convenient to conduct, as well allowing prior information to be incorporated on the parameters of the interest, when external information is available. Information regarding likelihood-based inference based on the EM algorithm as well as that regarding the complete-data likelihood function can be found in standard references such as McLachlan and Krishnan [2007].

Treating z_i $(i = 1, 2, \cdots, n)$ as latent variables, the Bayesian models can be implemented in the software package R using MCMC methods such as the Metropolis-Hastings algorithm (as was done in Xia and Gustafson [2016]). In addition, we may use Bayesian software packages such as WinBUGS and OpenBUGS to implement the models, by introducing a latent status on misrepresentation. In particular, the BUGS code for implementing the gamma model in Example 2.1 is as follows.

```
model {
  for (i in 1:n){
    V_star[i] ~ dbin(theta_star,1)
    Y[i] ~ dgamma(alpha, beta[i])
    beta[i] <- alpha/exp(aa0 + aa1*V[i] + aa2*X[i])
    V[i] <- V_star[i] + (1-V_star[i])*Z[i]
    Z[i]  ~ dbin(q,1)
  }
  theta_star <- theta*(1-p)
  q <- theta*p/(1-theta*(1-p))

  # Prior distributions
  p ~ dunif(0, 1)
  theta ~ dunif(0, 1)
  aa0 ~ dnorm(0, 0.1)
  aa1 ~ dnorm(0, 0.1)
  aa2 ~ dnorm(0, 0.1)
  alpha ~ dgamma(0.5, 0.5)
}
```

Using the observed values of (y_1, v_1^*, x_1), (y_2, v_2^*, x_2), $\cdots$, (y_n, v_n^*, x_n), the above BUGS program will output posterior samples of the parameters α_0, α_1, α_2, p, q, θ and φ. Note that the above normal and gamma prior distributions are chosen as vague priors that will work well for cases where the true parameters have values near 1, as in the simulation study. For real applications depending on the scales of the response and covariates, we may need to re-set the super-parameters or standardize continuous covariates in order for the normal and gamma priors to cover a range reasonably larger than the scale of the parameters. The use of WinBUGS and OpenBUGS for actuarial modeling was illustrated in earlier papers such as Scollnik [2001, 2002]. With R packages such as R2WinBUGS and BRUGS, we will be able to perform repeated computations using R.

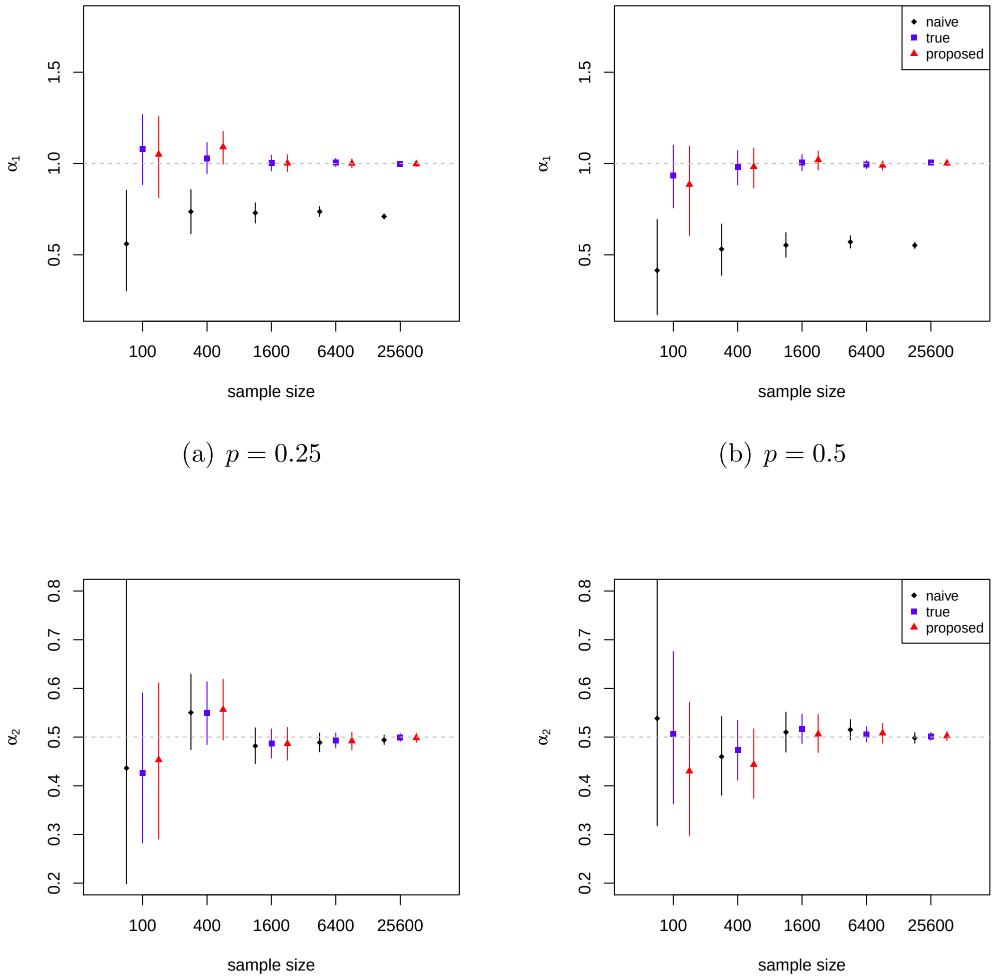
3.2 Impact from a correctly measured rating factor

Here, we include an additional risk factor X that is correctly measured. In the simulation study, we first generate the true risk status V, the additional factor X and the claim outcome Y from the true distributional structure for the gamma severity model in Example 2.1. The samples of V are then modified based on the true values of p in order to obtain the corresponding observed samples of V^*. The proposed model uses simulated samples of (Y, V^*, X) to estimate the parameters in the gamma distributional structure given in Example 2.1.

For the true risk factor V, we generate a single sample of size n, using a Bernoulli trial with the probability $\theta = 0.5$. Two different values of p, 0.25 and 0.5, are used as misrepresentation probabilities for obtaining the corresponding samples of V^*. The samples of the additional factor X are generated from a gamma distribution with the shape and scale parameters being (2, 0.5). The corresponding samples of Y are then generated from those of V and X, with regression coefficients (α_0, α_1, α_2) being $(1.2, 1, 0.5)$, and a gamma scale parameter $\varphi = 5$.

For the simulation study, we consider the five sample sizes of 100, 400, 1,600, 5,400, and 25,600. We compare the results from the proposed model in (2.2) with naive estimates from gamma regression using the observed values of V^*, pretending there to be no misrepresentation. We denote the true model as gamma regression using the "unobserved" values of V that we used earlier for obtaining the samples of V^*. For all the models, independent normal priors with mean 0 and variance 10 are used for the regression coefficients. For the probability parameters θ and p, uniform priors on (0, 1) are used. We run three chains with randomly generated initial values. The first 15,000 samples are dropped to ensure that the Markov chain has converged. In order to reduce the autocorrelation in the posterior samples, we take every 10th sample for our model acknowledging misrepresentation. For all the other models with faster convergence, we take every 10th sample after dropping the first 1,500 samples.

Figure 2 presents the 95% equal-tailed credible intervals for the regression effects α_1 and α_2 of the true risk status V and X, for each of the five sample sizes. The credible intervals are based on 5,000 posterior samples, with an effective size over 4,500. For MCMC, an effect size close to the nominal size indicates there is very little autocorrelation in the posterior samples that may jeopardize the efficiency in the estimation. We observe that the naive estimates are biased downward compared to those from the true models using the corresponding values of V. That is, misrepresentation in the risk factor causes an attenuation effect in the naive estimates for the risk effect α_1 for V, an effect commonly seen in measurement error modeling. The center of the posterior distribution for the regression effect α_1 from the proposed model is very



(a) $p = 0.25$

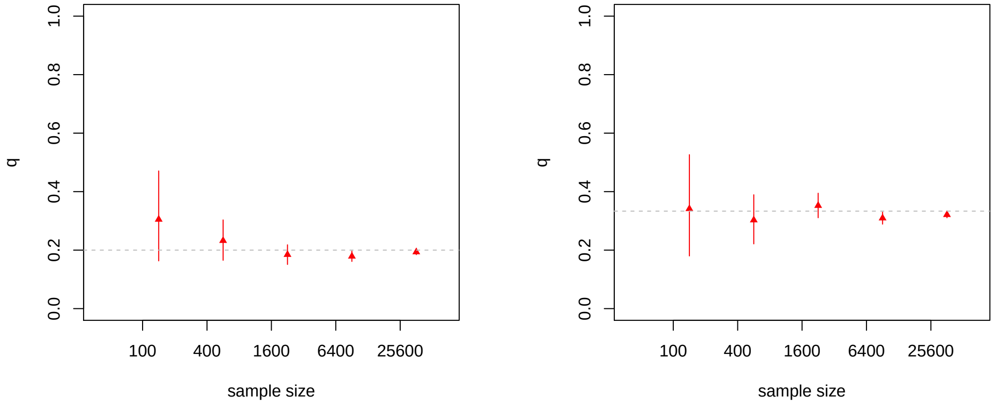
(b) $p = 0.5$

(c) $p = 0.25$

(d) $p = 0.5$

Figure 2: Credible intervals for the risk effects of V (top) and X (bottom) for the gamma loss severity model. The dashed line marks the true value.

close to that from the true model. There seems to be no noticeable difference concerning the estimation of α_2. The proposed model seems to give wider credible intervals, acknowledging the additional uncertainty due to the existence of the misrepresentation.



(a) $p = 0.25$ ($q = 0.2$)

(b) $p = 0.5$ ($q = 0.33$)

Figure 3: 95% credible intervals for the prevalence of misrepresentation q for the gamma loss severity model.

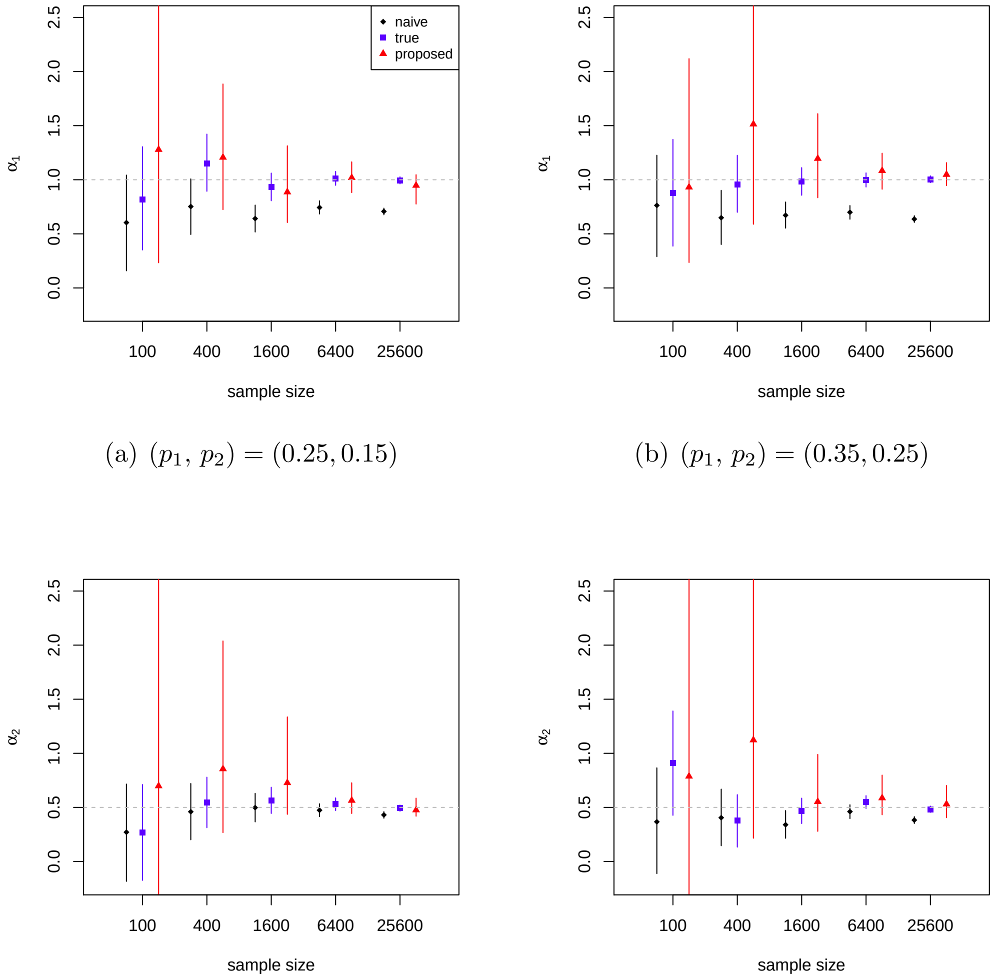
Figure 3 presents the credible intervals of the prevalence of misrepresentation q from the proposed model. For insurance applications, the prevalence q is a more meaningful measure that will allow us to obtain the number of misrepresented cases directly from data on the reported status. The credible intervals for the misrepresentation probability p have very similar patterns, and thus will not be presented here. The credible interval becomes narrower as the sample size increases, with all the intervals covering the true value of the probability. In both figures, there is larger variability in the estimation for the case with $p = 0.50$, where the issue of misrepresentation is more severe.

3.3 Multiple rating factors with misrepresentation

The second case we study is when we have two rating factors that are subject to misrepresentation. For the negative binomial loss frequency model in Example 2.2, we generate a single sample of size $n = 1,000$ for the true risk statuses (V_1, V_2), using two Bernoulli trials with the binomial probabilities $\theta_1 = 0.5$ and $\theta_2 = 0.4$, respectively. Two different sets of values of (p_1, p_2), $(0.25, 0.15)$ and $(0.35, 0.25)$, are used for the two risk factors for obtaining the corresponding samples of (V_1^*, V_2^*). The corresponding samples of negative binomial counts Y are then generated from those of V_1 and V_2, with regression coefficients $(\alpha_0, \alpha_1, \alpha_2)$ being $(-1, 1, 0.5)$, and a dispersion parameter $\varphi = 5$. The proposed model uses simulated samples of (Y, V_1^*, V_2^*) to estimate the parameters in the negative binomial distributional structure given in Example 2.2.

We compare the results from the proposed model in (2.4) with naive estimates from negative binomial regression using the observed values of (V_1^*, V_2^*), pretending there to be no misrepresentation. We denote the true model as negative binomial regression using the corresponding values of (V_1, V_2) that we used earlier for obtaining the samples of (V_1^*, V_2^*). For all the models, independent normal priors with mean 0 and variance 10 are used for the regression coefficients. For the probability parameters θ_1, θ_2, p_1 and p_2, uniform priors on (0, 1) are used. Other MCMC details are the same as those for the gamma model.

Figure 4 presents the 95% equal-tailed credible intervals for the regression coefficients α_1 and α_2 of the true risk statuses V_1 and V_2, for each of the five sample sizes. As expected, we observe that the naive estimates are biased downward to a certain extent, compared to those from the true models using the corresponding values of V_1 and V_2. That is, misrepresentation in the risk factor causes an attenuation effect in the naive estimates. The centers of the posterior distributions for the regression effects α_1 and α_2 from the proposed model are very close to those from the true model, for sufficiently large sample sizes. For the negative binomial model, the proposed method gives much wider credible intervals, when compared with those from a Poisson model we tried with two misrepresented risk factors. The existence of misrepresentation seems to cause a larger efficiency loss in the negative binomial model than its Poisson counterpart,



(a) $(p_1,\, p_2) = (0.25, 0.15)$

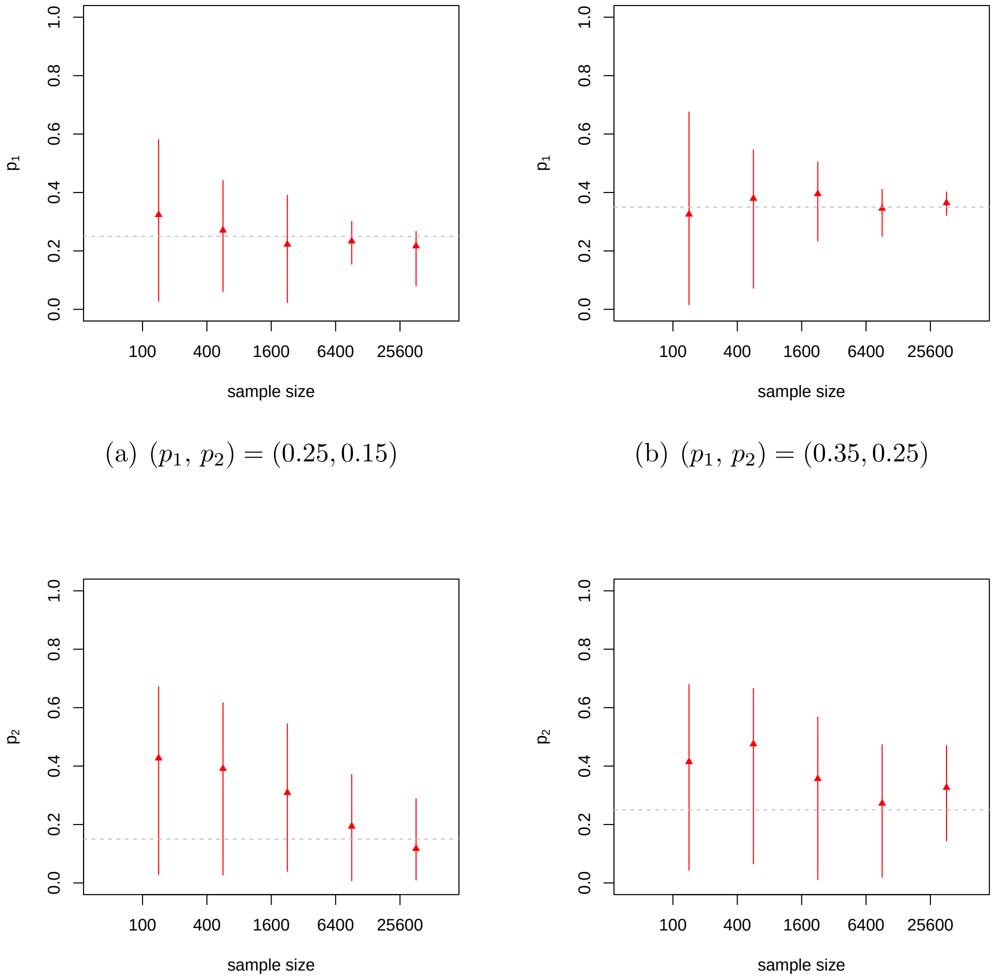
(b) $(p_1,\, p_2) = (0.35, 0.25)$

(c) $(p_1,\, p_2) = (0.25, 0.15)$

(d) $(p_1,\, p_2) = (0.35, 0.25)$

Figure 4: Credible intervals for the risk effects of V_1 (top) and V_2 (bottom) for the negative binomial loss frequency model. The dashed line marks the true value.

owing to the weak identification (Xia and Gustafson [2016]). The results for the Poisson model have similar patterns as those in Figures 2 and 3, and are not presented here. We further observe that the model works better when the sample size increases, which suggests that it will work for large insurance claim datasets.



(a) $(p_1, p_2) = (0.25, 0.15)$

(b) $(p_1, p_2) = (0.35, 0.25)$

(c) $(p_1, p_2) = (0.25, 0.15)$

(d) $(p_1, p_2) = (0.35, 0.25)$

Figure 5: 95% credible intervals for the probabilities p_1 and p_2 for the negative binomial loss frequency model.

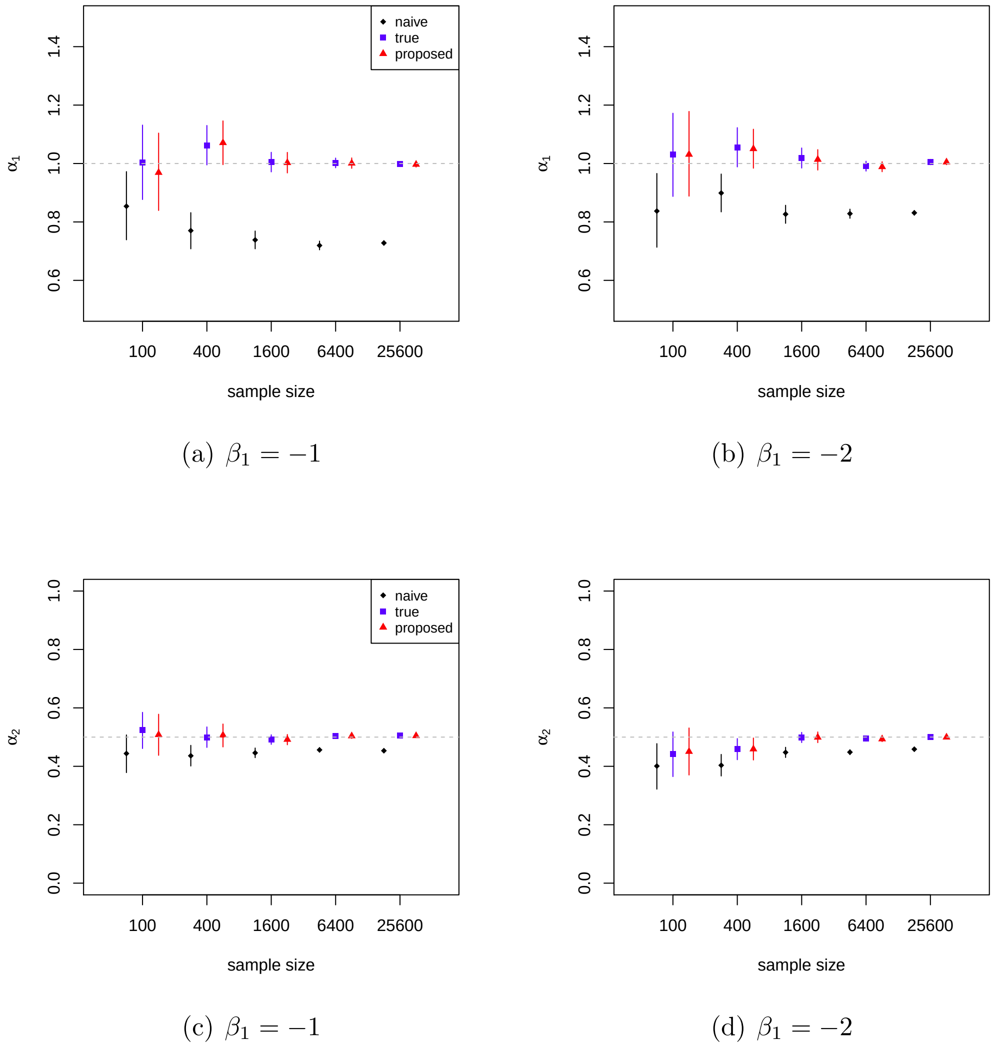
Figure 5 presents the credible intervals of the probabilities p_1 and p_2 from the proposed model. The results on the five prevalence of misrepresentation q_j's are similar. The credible interval becomes narrower as the sample size increases, with all the intervals covering the true value of the probability. In both figures, there is larger variability in the estimation for the case with $(p_1,\, p_2) = (0.35, 0.25)$, where the issue of misrepresentation is more severe.

3.4 Embedded model on prevalence of misrepresentation

The last case we study is when there is a correctly measured risk factor X that affects the prevalence of misrepresentation q. The process of data simulation differs, as we need to determine the value of q for each sample X. We directly simulate samples of V^* from a Bernoulli trial with a probability θ^*, and use them to obtain those of V based on the calculated values of q. The samples of V and X are then used to obtain those for the outcome Y. The proposed model uses simulated samples of (Y, V^*, X) to estimate the parameters in the Poisson distributional structure given in Example 2.3.

We use the negative binomial model in Example 2.3 to generate a single sample of size n for the reported risk status V^*, using a Bernoulli trial with the probability $\theta^* = 0.5$. The samples of the additional risk factor X are generated from a gamma distribution with the shape and scale parameters being 2 and 0.5, respectively. For the true model, we generate the corresponding samples of V, assuming that the prevalence of misrepresentation is given by $\text{logit}(q) = \beta_0 + \beta_1 X$. We assume the regression coefficients in logistic regression, $(\beta_0,\, \beta_1)$, take two sets of values $(0, -1)$ and $(0, -2)$. For each sample of X, we calculate the prevalence of misrepresentation and obtain the corresponding true samples of V based on those of V^*. The corresponding samples of negative binomial counts Y are then generated from those of V and X, with regression coefficients $(\alpha_0,\, \alpha_1,\, \alpha_2)$ being $(1.2, 1, 0.5)$.

We compare the results from the proposed model with naive estimates from negative binomial regression using the observed values of V^*, pretending there to be no misrepresentation. We denote the true model as negative binomial regression using the corresponding values of V. The proposed model is based on the mixture representation using the observed risk status V^*. For all the models, independent normal priors with mean 0 and variance 10 are used for all the regression coefficients and the logarithm of the dispersion parameter. For the probability parameter θ, a uniform prior on (0, 1) is used. Other MCMC details are the same as those for the gamma model.



(a) $\beta_1 = -1$

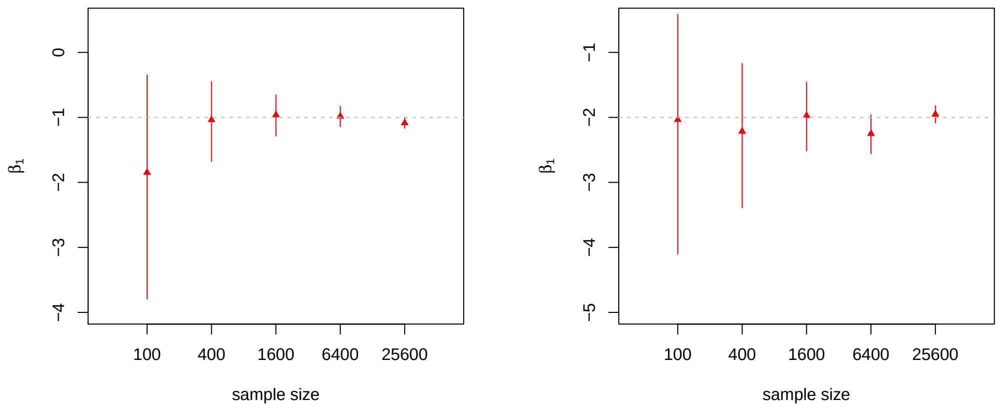
(b) $\beta_1 = -2$

(c) $\beta_1 = -1$

(d) $\beta_1 = -2$

Figure 6: Credible intervals for the risk effects of V (top) and X (bottom) for the Poisson loss frequency model. The dashed line marks the true value.

Figure 6 presents the 95% credible intervals for the regression coefficients α_1 and α_2 of the true risk status V and X, for each of the five sample sizes. For both α_1 and α_2, we observe that the naive estimates are biased downward compared to those from the true models using the corresponding values of V. This means misrepresentation in one risk factor may cause bias in the estimates of effects for both the risk factor itself (i.e., the attenuation effect), as well as for other risk factors. The centers of the posterior distributions for the regression effects α_1 and α_2 from the proposed model are very close to those from the true model. The proposed model seems to give a little larger posterior standard deviation, acknowledging the uncertainty due to the existence of misrepresentation.



(a) $\beta_1 = -1$

(b) $\beta_1 = -2$

Figure 7: 95% credible intervals for the risk effect β_1 on the prevalence of misrepresentation for the Poisson loss frequency model.

Figure 7 presents the credible intervals of the risk effect β_1 on the prevalence of misrepresentation from the proposed model. The credible interval becomes narrower as the sample size increases, with all the intervals covering the true value of the coefficient. In both figures, there is larger variability in the estimation for the case with $\beta_1 = -2$, where the prevalence of misrepresentation q varies more widely with the factor X. When compared with Figure 6, we observe that the credible intervals for the regression coefficients β_1 are wider than those for α_1 and α_2 that have a similar scale. This indicates that latent models may require a larger sample size to learn the parameters with the same precision. Note that the naive and true models do not assume misrepresentation in the risk factor, so there is no inference for the coefficient β_1 from the latent model on the prevalence of misrepresentation.

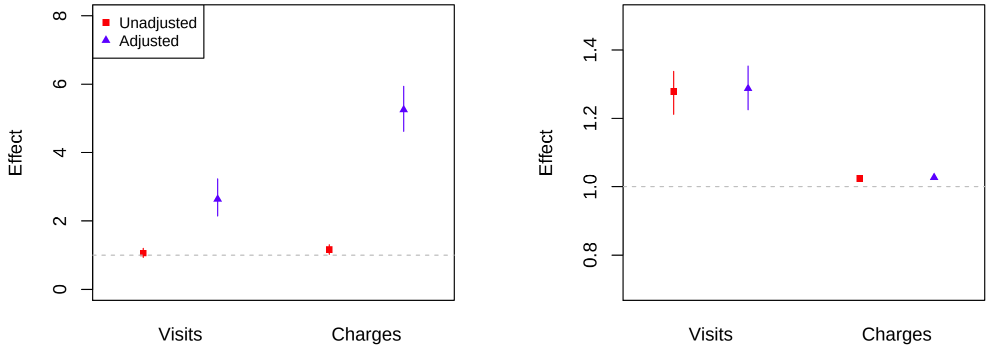
4 Case study on medical expenditures

The Medical Expenditure Panel Survey (MEPS, AHRQ [2013]) is a set of national surveys on medical expenditures and frequencies of healthcare utilization by the Americans. In the actuarial and statistical literature, the MEPS data have been used by earlier papers such as Frees et al. [2013], Xia and Gustafson [2014, 2016], Hua and Xia [2014], Hua [2015] for exploring patterns concerning healthcare loss frequency and severity. For the case study, we use the 2013 MEPS consolidated data to illustrate the proposed GLM loss frequency and severity models that embed a predictive analysis on the misrepresentation risk.

In the case study, we choose two response variables, the office-based visits and total medical charges, respectively for our loss frequency and severity models. According to the Patient Protection and Affordable Care Act (PPACA), health insurance premiums can account for the five risk factors of age, location, tobacco use, plan type (individual vs. family) and plan category based on the level of coverage. For the MEPS data, we choose the two factors of age and smoking status that are available in the dataset. In particular, the self-reported smoking status may be subject to misrepresentation, owing to social desirability concerns. For the empirical analysis, we include insured individuals from the age of 18 to 60 who were the reference person in their household. The sample sizes for the office-based visits and the total medical charges are 3,249 and 2,948, respectively. The sample size is smaller for the total medical charges variable, as we only include individuals with a positive expenditure.

Due to the over-dispersed feature of the office-based visits variable, we specify a negative binomial GLM for the loss frequency model. For the total medical charges variable, we use a gamma GLM for the loss severity model. We first perform an unadjusted analysis, using regular GLM ratemaking models, without adjusting for misrepresentation in the self-reported smoking status. For the adjusted analysis, we adopt the same regression structures as those in Example

2.3, despite differences in the distributional form of Y. The MCMC settings are similar to those in the previous section. For the regression coefficients α_0, α_1, α_2, β_0 and β_1, we specify vague normal priors with mean 0 and variance 10 (100 for the gamma model to account for a larger scale of Y). For the probability θ, we assume a beta prior with both parameters being 2 (corresponding to prior mean 0.5 and standard deviation 0.224). At the current sample sizes, the slightly more concentrate beta prior seems to help with the convergence of the negative binomial model. For the probabilities p and q, the prior distributions are transformations of those for β_0, β_1 and θ, according to the relationships of the parameters.

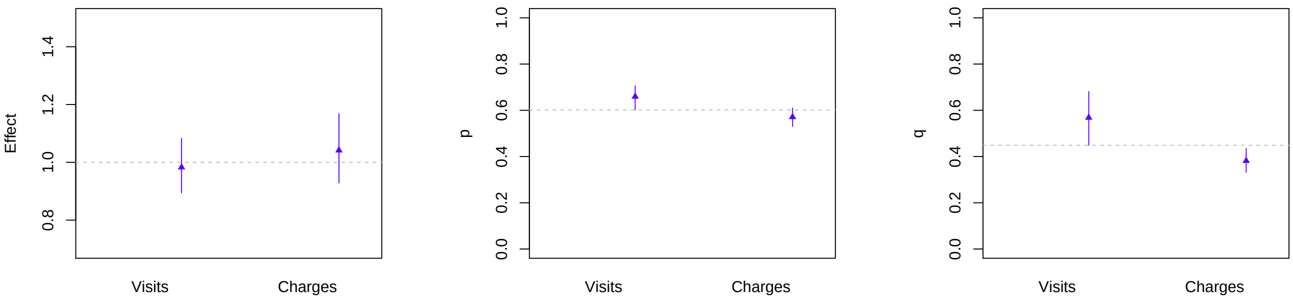


(a) Smoking

(b) Age

Figure 8: Credible intervals for the relativity of smoking and age, $\exp(\alpha_1)$ and $\exp(\alpha_2)$, for the negative binomial model on the office-based visits (left column) and the gamma model on total medical charges (right column). The age effect corresponds to the increase of age by one standard deviation (i.e., 12 years).

In Figure 8, we present the 95% equal-tailed credible intervals for the relativity $\exp(\alpha_1)$ and $\exp(\alpha_2)$ concerning the smoking and age effects on the average number of office-based visits and the average total medical charges. We observe that for the smoking risk factor, the adjusted models give estimated relativities that are substantially higher than those from the unadjusted models. Note that the above difference in the estimated relativity is very large, as the relativity is the exponential of the regression coefficients. When we look at the regression coefficients, the estimates are comparable to those from the simulation studies in the previous section. With the estimated prevalence q ranging from 0.38 to 0.57, such an estimated difference is likely to be attributed to misrepresentation. From a practical standpoint, the estimated smoking relativity from the unadjusted analysis is very close to one, contradictory to clinical findings on the health risks associated with smoking. For the current study, the estimates from the adjusted model are more likely to reflect the true smoking effect on the health outcomes. In general, however, business knowledge needs to be used when interpreting results from empirical studies. In the case of observation studies, there is a possibility of heterogeneity due to confounding from other risk factors. In such cases, the embedded analysis help us identify heterogeneity in the data that may require inclusion of other risk factors. The predictive model on the misrepresentation risk could provide insights to the underwriting department that would help optimize the cost/benefit tradeoff of undertaking interventions to minimize the occurrence of such frauds. For the age effect, the adjustment results in no noticeable difference in the estimated relativity. Both the smoking and age effects are significant from the adjusted model, confirming our intuition.



(a) Age

(b) $p(\bar{x})$

(c) $q(\bar{x})$

Figure 9: Credible intervals for the age effect $\exp(\beta_1)$ on the odds of misrepresentation, the predicted misrepresentation probability $p(\bar{x})$, and the prevalence of misrepresentation $q(\bar{x})$ for individuals at the average age of 42. In each panel, the left column corresponds to the negative binomial model on office-based visits, and the right column corresponds to the gamma model on total medical charges.

In Figure 9, we present the 95% equal-tailed credible intervals for the relative age effect on the odds of misrepresentation (i.e., $q/(1-q)$), the predicted misrepresentation probability $p(\bar{x})$, and the predicted prevalence of misrepresentation $q(\bar{x})$ for individuals at the average age of 42.

We observe that the age effect is insignificant in predicting the prevalence of misrepresentation q regarding both outcomes on the office-based visits and total medical charges. For individuals at the average age, the predicted misrepresentation probabilities are 66% and 57% for office-based visits and total medical charges. The credible intervals overlap for the two models using samples on different outcomes, indicating no statistical difference in the two probabilities. The predicted prevalence of misrepresentation is about 57% and 38%, with a larger difference caused by difference in the percentage of smokers in the gamma model that excludes individuals with no medical charge. Among people with an average age who identified themselves as non-smokers, about 48% of them are estimated to have misrepresented their smoking status.

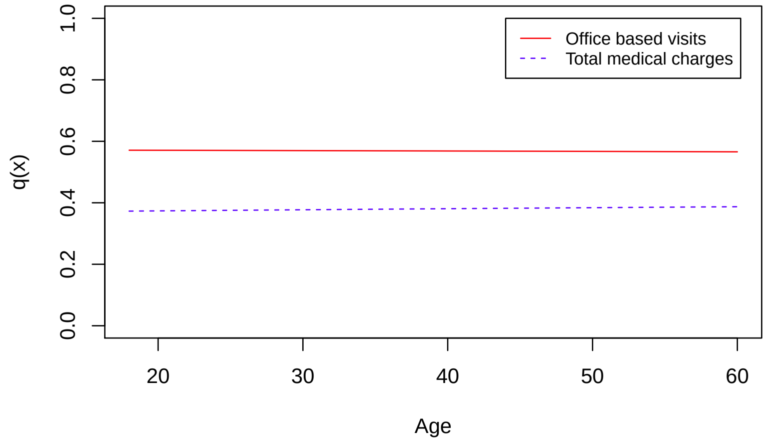


Figure 10: Predicted misrepresentation probability $q(x)$ by age for individuals who reported nonsmoking.

In Figure 10, we present the predicted prevalence of misrepresentation by age for individuals who identified themselves as non-smokers. For both the office-based visits and total medical charges, the predicted prevalence of misrepresentation does not seem to vary with age. For both models, we can predict the prevalence of misrepresentation for individuals with a specific age. For example, for respondents who were 60 years old, the predicted prevalence of misrepresentation is 56.6% and 38.7% respectively for the office-based visits and the total medical charges. It is not surprising for the predicted prevalence of misrepresentation to differ, as the percentages of smokers seem to differ in the sub-sample of individuals used for the gamma model who had positive medical charges. The predicted prevalence parameters of misrepresentation constitute risk scores concerning the misrepresentation risk. Here the age effect in Figure 9 is an insignificant risk factor concerning the misrepresentation probability, as there is no evidence that people will become more honest or less honest over time. With real insurance data, we may be able to identify other significant factors when we have a much larger sample size as well as a larger number of risk factors.

5 Conclusions

In the paper, we proposed an embedded predictive analysis of misrepresentation risk in GLM ratemaking models. Under the GLM ratemaking structure, we derived the mixture regression form for the conditional distribution of the claim outcome given the observed risk factors when some of them are subject to misrepresentation. The mixture regression form ensures the model identifiability, so that all parameters including the true relativities and the prevalence of misrepresentation can be estimated consistently using regular ratemaking data. Based on mixture regression models with concomitant variables, we embedded a predictive analysis of misrepresentation risk by a latent logistic regression model on the prevalence of misrepresentation. For insurance companies that have information on various risk factors concerning an insured policy, such an embedded model on the prevalence of misrepresentation allows the underwriting department to generate a misrepresentation risk profile based on models fitted from historical data. Using the risk profiles, the underwriting department may choose to undertake investigations on certain policies, in order to minimize the occurrence of misrepresentation frauds. By concentrating on policies with a higher misrepresentation risk, the analysis will help enhance the efficiency of underwriting practices. For the claims department, risk profiles on misrepresentation on policy applications may be further used to identify fraudulent claims.

Acknowledgments

The authors are grateful to the editor, associate editor and anonymous referees for their valuable comments and suggestions that help significantly improve the quality of the paper. The authors are grateful to the Casualty Actuarial Society (CAS) for their generous support at the 2016 Individual Grant Competition, and the travel support for presenting the work at the 2016 CAS annual meeting. The work was funded earlier by the Research and Artistry Opportunity Grant at the Northern Illinois University (NIU). Last but not least, we would like to thank Lauren Anglin for her work in obtaining the observable moments funded through the Student Engagement Funding (SEF) at NIU, and her helpful comments that help improve the slides for the presentation at the CAS annual meeting.

Appendix

A.1 Additional derivation for Section 2.2

Here we derive the forms of the prevalence of misrepresentation, the q_j's, for the model in Equation (2.4). Using Bayes's Theorem, we have

$$\begin{aligned} q_1 &= \mathrm{P}(V_1 = 1,\, V_2 = 1 \,|\, V_1^* = 0,\, V_2^* = 1) \\ &= \frac{\mathrm{P}(V_1^* = 0,\, V_2^* = 1 \,|\, V_1 = 1,\, V_2 = 1)\mathrm{P}(V_1 = 1,\, V_2 = 1)}{\sum_{v_1 \in \{0,1\};\, v_2 \in \{0,1\}} \mathrm{P}(V_1^* = 0,\, V_2^* = 1 \,|\, V_1 = v_1,\, V_2 = v_2)\mathrm{P}(V_1 = v_1,\, V_2 = v_2)} \\ &= \frac{p_1(1-p_2)\theta_1\theta_2}{0 + 0 + (1-p_2)(1-\theta_1)\theta_2 + p_1(1-p_2)\theta_1\theta_2} = \frac{\theta_1 p_1}{1 - \theta_1(1-p_1)} \end{aligned}$$

$$\begin{aligned} q_2 &= \mathrm{P}(V_1 = 1,\, V_2 = 1 \,|\, V_1^* = 1,\, V_2^* = 0) \\ &= \frac{\mathrm{P}(V_1^* = 1,\, V_2^* = 0 \,|\, V_1 = 1,\, V_2 = 1)\mathrm{P}(V_1 = 1,\, V_2 = 1)}{\sum_{v_1 \in \{0,1\};\, v_2 \in \{0,1\}} \mathrm{P}(V_1^* = 1,\, V_2^* = 0 \,|\, V_1 = v_1,\, V_2 = v_2)\mathrm{P}(V_1 = v_1,\, V_2 = v_2)} \\ &= \frac{(1-p_1)p_2\theta_1\theta_2}{0 + 0 + (1-p_1)\theta_1(1-\theta_2) + (1-p_1)p_2\theta_1\theta_2} = \frac{\theta_2 p_2}{1 - \theta_2(1-p_2)} \end{aligned}$$

$$\begin{aligned}
q_3 &= \mathrm{P}(V_1 = 1,\, V_2 = 1 \,|\, V_1^* = 0,\, V_2^* = 0) \\
&= \frac{\mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = 1,\, V_2 = 1)\mathrm{P}(V_1 = 1,\, V_2 = 1)}{\sum_{v_1\in\{0,1\};\, v_2\in\{0,1\}} \mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = v_1,\, V_2 = v_2)\mathrm{P}(V_1 = v_1,\, V_2 = v_2)} \\
&= \frac{p_1 p_2 \theta_1 \theta_2}{p_1 p_2 \theta_1 \theta_2 + p_1\theta_1(1-\theta_2) + p_2(1-\theta_1)\theta_2 + (1-\theta_1)(1-\theta_2)}
\end{aligned}$$

$$\begin{aligned}
q_4 &= \mathrm{P}(V_1 = 0,\, V_2 = 1 \,|\, V_1^* = 0,\, V_2^* = 0) \\
&= \frac{\mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = 0,\, V_2 = 1)\mathrm{P}(V_1 = 0,\, V_2 = 1)}{\sum_{v_1\in\{0,1\};\, v_2\in\{0,1\}} \mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = v_1,\, V_2 = v_2)\mathrm{P}(V_1 = v_1,\, V_2 = v_2)} \\
&= \frac{p_2(1-\theta_1)\theta_2}{p_1 p_2 \theta_1 \theta_2 + p_1\theta_1(1-\theta_2) + p_2(1-\theta_1)\theta_2 + (1-\theta_1)(1-\theta_2)}
\end{aligned}$$

$$\begin{aligned}
q_5 &= \mathrm{P}(V_1 = 1,\, V_2 = 0 \,|\, V_1^* = 0,\, V_2^* = 0) \\
&= \frac{\mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = 1,\, V_2 = 0)\mathrm{P}(V_1 = 1,\, V_2 = 0)}{\sum_{v_1\in\{0,1\};\, v_2\in\{0,1\}} \mathrm{P}(V_1^* = 0,\, V_2^* = 0 \,|\, V_1 = v_1,\, V_2 = v_2)\mathrm{P}(V_1 = v_1,\, V_2 = v_2)} \\
&= \frac{p_1\theta_1(1-\theta_2)}{p_1 p_2 \theta_1 \theta_2 + p_1\theta_1(1-\theta_2) + p_2(1-\theta_1)\theta_2 + (1-\theta_1)(1-\theta_2)}.
\end{aligned}$$

References

AHRQ. Agency for Healthcare Research and Quality (AHRQ). Medical Expenditure Panel Survey. *Rockville (MD): U.S. Department of Health and Human Services*, 2013. URL `http://meps.ahrq.gov/mepsweb/`.

N. Atienza, J. Garcia-Heras, and J. Munoz-Pichardo. A new condition for identifiability of finite mixture distributions. *Metrika*, 63(2):215–221, 2006.

N. Atienza, J. Garcia-Heras, J. Munoz-Pichardo, and R. Villa. On the consistency of MLE in finite mixture models of exponential families. *Journal of Statistical Planning and Inference*, 137(2):496–505, 2007.

L. Bermúdez and D. Karlis. A posteriori ratemaking using bivariate Poisson models. *Scandinavian Actuarial Journal*, pages 1–11, 2015.

M. J. Brockman and T. S. Wright. Statistical motor rating: making effective use of your data. *Journal of the Institute of Actuaries*, 119:457–543, 1992. ISSN 0.

M. David. Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, 20:147–156, 2015.

E. W. Frees, X. Jin, and X. Lin. Actuarial applications of multivariate two-part regression models. *Annals of Actuarial Science*, 7(2):258–287, 2013.

E. W. Frees, R. A. Derrig, and G. Meyers. *Predictive modeling applications in actuarial science*, volume 1. Cambridge University Press, 2014.

B. Grün and F. Leisch. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28:1–35, 2008a.

B. Grün and F. Leisch. Finite mixtures of generalized linear regression models. In *Recent advances in linear models and related areas*, pages 205–230. Springer, 2008b.

P. Gustafson. Bayesian statistical methodology for observational health sciences data. *Statistics in Action: A Canadian Outlook*, pages 163–176, 2014.

S. Haberman and A. E. Renshaw. Generalized linear models and actuarial science. *The Statistician*, 45(4):407–436, 1996.

P. R. Hahn, J. S. Murray, and I. Manolopoulou. A Bayesian partial identification approach to inferring the prevalence of accounting misconduct. *Journal of the American Statistical Association*, 111:14–26, 2016. doi: 10.1080/01621459.2015.1084307.

C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2):273–296, 2000.

L. Hua. Tail negative dependence and its applications for aggregate loss modeling. *Insurance: Mathematics and Economics*, 61:135–145, 2015.

L. Hua and M. Xia. Assessing high-risk scenarios by full-range tail dependence copulas. *North American Actuarial Journal*, 18(3):363–378, 2014.

N. Klein, M. Denuit, S. Lang, and T. Kneib. Nonlife ratemaking and risk management with bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55:225–249, 2014.

J. Lemaire, S. C. Park, and K. C. Wang. The use of annual mileage as a rating variable. *Astin Bulletin*, 46(01):39–69, 2016.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

D. P. Scollnik. Actuarial modeling with MCMC and BUGS. *North American Actuarial Journal*, 5(2):96–124, 2001.

D. P. Scollnik. Modeling size-of-loss distributions for exact data in WinBUGS. *Journal of Actuarial Practice*, 10:202–227, 2002.

P. Shi. Insurance ratemaking using a copula-based multivariate Tweedie model. *Scandinavian Actuarial Journal*, 2016(3):198–215, 2016.

P. Shi and E. A. Valdez. A copula approach to test asymmetric information with applications to predictive modeling. *Insurance: Mathematics and Economics*, 49(2):226–239, 2011.

L. Sun, M. Xia, Y. Tang, and P. G. Jones. Bayesian adjustment for unidirectional misclassification in ordinal covariates. *Journal of Statistical Computation and Simulation*, 87(18): 3440–3468, 2017. doi: 10.1080/00949655.2017.1370649.

R. S. Winsor. *Misrepresentation and non Disclosure on Applications for Insurance.* Blaney McMurtry LLP, 1995.

M. Xia and P. Gustafson. Bayesian sensitivity analyses for hidden sub-populations in weighted sampling. *Canadian Journal of Statistics*, 42(3):436–450, 2014.

M. Xia and P. Gustafson. Bayesian regression models adjusting for unidirectional covariate misclassification. *The Canadian Journal of Statistics*, 44:198–218, 2016. doi: 10.1002/cjs.11284.

M. Xia and P. Gustafson. Bayesian inference for unidirectional misclassification of a binary response trait. *Statistics in Medicine*, (in press):1–15, 2017.

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