Hierarchical Insurance Claims Modeling

Edward W. FREES and Emiliano A. VALDEZ

This work describes statistical modeling of detailed, microlevel automobile insurance records. We consider 1993–2001 data from a major insurance company in Singapore. By detailed microlevel records, we mean experience at the individual vehicle level, including vehicle and driver characteristics, insurance coverage, and claims experience, by year. The claims experience consists of detailed information on the type of insurance claim, such as whether the claim is due to injury to a third party, property damage to a third party, or claims for damage to the insured, as well as the corresponding claim amount. We propose a hierarchical model for three components, corresponding to the frequency, type, and severity of claims. The first model is a negative binomial regression model for assessing claim frequency. The driver's gender, age, and no claims discount, as well as vehicle age and type, turn out to be important variables for predicting the event of a claim. The second is a multinomial logit model to predict the type of insurance claim, whether it is third-party injury, third-party property damage, insured's own damage or some combination. Year, vehicle age, and vehicle type turn out to be important predictors for this component. Our third model is for the severity component. Here we use a generalized beta of the second kind of long-tailed distribution for claim amounts and also incorporate predictor variables. Year, vehicle age, and person's age turn out to be important predictors for this component. Not surprisingly, we show a significant dependence among the different claim types; we use a t-copula to account for this dependence. The three-component model provides justification for assessing the importance of a rating variable. When taken together, the integrated model allows more efficient prediction of automobile claims compared with than traditional methods. Using simulation, we demonstrate this by developing predictive distributions and calculating premiums under alternative coverage limitations.

KEY WORDS: Copula; Insurance claim; Long-tailed regression.

1. INTRODUCTION

A primary attribute of the actuary has been the ability to successfully apply statistical techniques in the analysis and interpretation of data. In this article we analyze a highly complex data structure and demonstrate the use of modern statistical techniques in solving actuarial problems. Specifically, we focus on a portfolio of motor (or automobile) insurance policies and, in analyzing the historical data drawn from this portfolio, we are able to revisit some of the classical problems faced by actuaries dealing with insurance data. This article explores statistical models that can be constructed when detailed, microlevel records of automobile insurance policies are available.

To an actuarial audience, the article provides a fresh look into the process of modeling and estimation of insurance data. For a statistical audience, we wish to emphasize the following:

- The highly complex data structure, making the statistical analysis and procedures interesting. Despite this complexity, the automobile insurance problem is common, and many readers will be able to relate to the data.
- The long-tailed nature of the distribution of insurance claims. This, along with the multivariate nature of different claim types, is of broad interest. Using the additional information provided by the frequency and type of claims, the actuary will be able to provide more accurate estimates of the claims distribution.
- The interpretation of the models and their results. We introduce a hierarchical, three-component model structure to help interpret our complex data.

In analyzing the data, we focus on two concerns of the actuary. First, there is a consensus, at least for motor insurance, of the importance of identifying key explanatory variables for rating purposes [see, e.g., Lemaire 1985 or the guide available from the General Insurance Association (GIA) of Singapore (*http://www.gia.org.sg*)]. Insurers often adopt a so-called "risk factor rating system" in establishing premiums for motor insurance, so that identifying these important risk factors is a crucial process in developing insurance rates. To illustrate, these risk factors include driver (e.g., age, gender) and vehicle (e.g., make/brand/model of car, cubic capacity) characteristics.

The second concern is one of the most important aspects of the actuary's job: to be able to predict claims as accurately as possible. Actuaries require accurate predictions for pricing, for estimating future company liabilities, and for understanding the implications of these claims to the solvency of the company. For example, in pricing, the actuary may attempt to quantify the effect of placing an upper limit on the amount reimbursed to a policyholder in the event of a claim, known as a "coverage limit." This process is important to ensure equity in the premium structure available to consumers.

In this article we consider policy exposure and claims experience data derived from vehicle insurance portfolios from a major general insurance company in Singapore. Our data are from the GIA of Singapore, an organization comprising most of the general insurers in Singapore. The observations are from each policyholder over a period of 9 years, January 1993–December 2001. Thus our data come from financial records of automobile insurance policies. In many countries, owners of automobiles are not free to drive their vehicles without some form of insurance coverage. Singapore is no exception; it requires drivers to have minimum coverage for personal injury to third parties.

We examined three databases: the policy, claims, and payment files. The policy file consists of all policyholders with vehicle insurance coverage purchased from a general insurer during the observation period. Each vehicle is identified with

Edward W. Frees is Assurant Health Professor of Actuarial Science, School of Business, University of Wisconsin, Madison, WI 53706 (E-mail: *jfrees@bus.wisc.edu*). Emiliano A. Valdez is Professor of Actuarial Science, Department of Mathematics, University of Connecticut, Storrs, CT 06269-3009 (E-mail: *valdez@math.uconn.edu*). The authors acknowledge the research assistance of Mitchell Wills, Shi Peng, and Katrien Antonio. The work of Frees was supported by the National Science Foundation (grant SES-0436274) and the Assurant Health Insurance Professorship. The work of Valdez was supported by the Australian Research Council through Discovery grant DP0345036 and the UNSW Actuarial Foundation of the Institute of Actuaries of Australia. The authors thank the associate editor and three referees for extensive comments that led us to reformulate much of the manuscript.

Journal of the American Statistical Association
December 2008, Vol. 103, No. 484, Applications and Case Studies
DOI 10.1198/016214508000000823

a unique code. This file provides characteristics of the policyholder, such as age and gender, and of the vehicle insured, such as type and age. The claims file provides a record of each accident claim filed with the insurer during the observation period and is linked to the policyholder file. The payment file consists of information on each payment made during the observation period and is linked to the claims file. It is common to see that a claim will have multiple payments made.

In predicting or estimating claims distributions, at least for motor insurance, we often associate the cost of claims with two components: the event of an accident and the amount of claim, if an accident occurs. Actuaries term these the claims frequency and severity components. This is the traditional way of decomposing this so-called "two-part" data, where one can think of a zero as arising from a vehicle without a claim. This decomposition easily allows us to incorporate having multiple claims per vehicle. Moreover, records from our databases show that when a claim payment is made, we also can identify the type of claim. For our data, there are three types: (a) claims for injury to a party other than the insured; (b) claims for damages to the insured, including injury, property damage, fire, and theft; and (c) claims for property damage to a party other than the insured. Identifying the type of claim is traditionally done through a "multidecrement" model, such as might be encountered in a competing-risks framework in biostatistics [see, e.g., Bowers, Gerber, Hickman, Jones, and Nesbitt 1997 for an actuarial introduction to two-part data (chap. 2) and multidecrement models (chap. 10)].

Thus, instead of a traditional univariate claim analysis, we potentially observe a trivariate claim amount, one claim for each type. For each accident, it is possible to have more than a single type of claim incurred; for example, an automobile accident can result in damages to a driver's own property, as well as damages to a third party involved in the accident. Thus modeling the joint distribution of the simultaneous occurrence of these claim types when an accident occurs provides the unique feature in this article. From a multivariate analysis standpoint, this is a nonstandard problem, in that we rarely observe all three claim types simultaneously (see Sec. 3.3 for the distribution of claim types). Not surprisingly, it turns out that claim amounts among types are related. To further complicate matters, it turns out that one type of claim is censored (see Sec. 2.1). We use copula functions to specify the joint multivariate distribution of the claims arising from these various claims types (see Frees and Valdez 1998; Nelsen 1999 for introductions to copula modeling).

To provide focus, we restrict our considerations to "nonfleet" policies; these comprise about 90% of the policies for this company. These are policies issued to customers whose insurance covers a single vehicle. In contrast, fleet policies are issued to companies that insure several vehicles, for example, coverage provided to a taxicab company, where several taxicabs are insured (see Angers, Desjardins, Dionne, and Guertin 2006; Desjardins, Dionne, and Pinguet 2001 for discussions of fleet policies). Thus unit of observation in our analysis is a registered vehicle insured, broken down according to exposure in each calendar year from 1993 to 2001. To investigate the full multivariate nature of claims, we further restrict our consideration to policies that offer comprehensive coverage, not just coverage for only third-party injury or property damage.

In constructing the models for our portfolio of policies, we focus on the development of the claims distribution according to three different components: claims frequency, conditional claim type, and conditional severity. The claims frequency provides the likelihood that an insured registered vehicle will have an accident and will make a claim in a given calendar year. Given that a claim is to be made when an accident occurs, the conditional claim type model describes the probability that it will be one of the three claim types, or any possible combination of them. The conditional severity component describes the claim amount structure according to the combination of claim types paid. In this article we provide appropriate statistical models for each component, emphasizing that the unique feature of this decomposition is the joint multivariate modeling of the claim amounts arising from the various claim types. Because of the short-term nature of the insurance coverages investigated here, we summarize the many payments per claim into a single claim amount (see Antonio, Beirlant, Hoedemakers, and Verlaak 2006 for a recent description of the claims "run-off" problem).

The article is organized as follows. First, in Section 2 we introduce the observable data, summarize its important characteristics, and provide details of the statistical models chosen for each of the three components of frequency, conditional claim type, and conditional severity. In Section 3 we fit the statistical model to the data and interpreting the results. The likelihood function construction for the estimation of the conditional severity component is detailed in the Appendix. In Section 4 we describe how the modeling construction and results can be used. We provide concluding remarks in Section 5.

2. MODELING

2.1 Data Structure

As explained in Section 1, the data available are disaggregated by risk class i, denoting insured vehicle, and over time t, denoting calendar year. Then, for each observational unit $\{it\}$, the potentially observable responses consist of the following:

- N_{it}, the number of claims within a year
- $M_{it,j}$, the type of claim, available for each claim, $j = 1, \ldots, N_{it}$
- $C_{it,jk}$, the claim amount, available for each claim, $j = 1, \ldots, N_{it}$, and for each type of claim, $k = 1, 2, 3$.

When a claim is made, it is possible to have one or a combination of three types of claims. To reiterate, we consider (a) claims for injury to a party other than the insured; (b) claims for damages to the insured, including injury, property damage, fire and theft; and (c) claims for property damage to a party other than the insured. Occasionally, we refer to these simply "injury," "own damage," and "third-party property." It is not uncommon for more than one type of claim to be incurred with each accident.

For the two third-party types, loss amounts are available; however, for damages to the insured ("own damages"), only a claim amount is available. Here we follow standard actuarial terminology and define the claim amount, $C_{it,2k}$, to be equal to the excess of a loss over a known deductible, d_{it} (and equal to 0 if the loss is less than the deductible). For notation purposes, we

sometimes use $C^*_{it,2k}$ to denote the loss amount; this quantity is not known when it falls below the deductible. Thus it is possible to observe a zero claim associated with an "own damages" claim. For our analysis, we assume that the deductibles apply on a per accident basis.

We also have the exposure e_{it}, measured in (a fraction of) years, which provides the length of time in the calendar year during which the vehicle had insurance coverage. The various vehicle and policyholder characteristics are described by the vector $\mathbf{x}_{it}$ and serve as explanatory variables in our analysis. For notational purposes, let $\mathbf{M}_{it}$ denote the vector of claim types for an observational unit and similarly for $\mathbf{C}_{it}$. In summary, the observable data available consist of

$$\{d_{it}, e_{it}, N_{it}, \mathbf{M}_{it}, \mathbf{C}_{it}, \mathbf{x}_{it},\ t = 1, \ldots, T_i,\ i = 1, \ldots, n\}.$$

There are $n = 96{,}014$ subjects, each of which is observed T_i times. In principle, the maximum value of T_i is 9 years, because our data consist of policies from 1993 to 2001. Even though a policy issued in 2001 may well extend coverage into 2002, we ignore the exposure and claims behavior beyond 2001. The motivation is to follow standard accounting periods on which actuarial reports are based. However, our data set is from an insurance company that had a substantial turnover of policies. For the full data set, there are 199,352 observations arising from 96,014 subjects, for an average of only 2.08 observations per subject. When examining the weights e_{it}, there is an average of only 1.29 years of exposure per subject. Thus, although we model the longitudinal behavior of subjects, for this data set, the relationship among components turns out to be more relevant.

2.2 Decomposing the Joint Distribution Into Components

Suppressing the $\{it\}$ subscripts, we decompose the joint distribution of the dependent variables as

$$\begin{aligned} \mathrm{f}(N, \mathbf{M}, \mathbf{C}) &= \mathrm{f}(N) \times \mathrm{f}(\mathbf{M}|N) \times \mathrm{f}(\mathbf{C}|N, \mathbf{M}), \\ \text{joint} &= \text{frequency} \times \text{conditional claim type} \\ &\quad \times \text{conditional severity}, \end{aligned}$$

where $\mathrm{f}(N, \mathbf{M}, \mathbf{C})$ denotes the joint distribution of $(N, \mathbf{M}, \mathbf{C})$. This joint distribution equals the product of the following three components:

1. Claims frequency: $\mathrm{f}(N)$ denotes the probability of having N claims.
2. Conditional claim type: $\mathrm{f}(\mathbf{M}|N)$ denotes the probability of having a claim type of $\mathbf{M}$, given N.
3. Conditional severity: $\mathrm{f}(\mathbf{C}|N, \mathbf{M})$ denotes the conditional density of the claim vector $\mathbf{C}$ given N and $\mathbf{M}$.

In the actuarial literature it is customary to condition on the frequency component when analyzing the joint frequency and severity distributions (see, e.g., Klugman, Panjer, and Willmot 2004). As described in Section 2.2.2, we incorporate an additional claims type layer. An alternative approach was taken by Pinquet (1998). He was interested in two lines of business, claims at fault and not at fault with respect to a third party. For each line, he hypothesized a frequency component and a severity component that were allowed to be correlated to one another. In particular, the claims frequency distribution was assumed to be bivariate Poisson. In contrast, our approach is to have a univariate claims number process and then decompose each claim by claim type. As we show in Section 2.2.3, we also allow for dependent claim amounts arising from the different claim types using the copula approach. Under this approach, a wide range of possible dependence structure can be flexibly specified.

We next discuss each of the three components.

2.2.1 Frequency Component. The frequency component, $\mathrm{f}(N)$, has been well analyzed in the actuarial literature. The modern approach of fitting a claims number distribution to longitudinal data can be attributed to the work of Dionne and Vanasse (1989), who applied a random-effects Poisson count model to automobile insurance claims. Here a (time-constant) latent variable was used to represent the heterogeneity among the claims, which also implicitly induces a constant correlation over time. For their data, Dionne and Vanasse established that a random-effects Poisson model provided a better fit than the usual Poisson and negative binomial models. Pinquet (1997, 1998) extended this work, considering severity as well as frequency distributions. He also allowed for different lines of business, as well as an explicit correlation parameter between the frequency and the severity components. Later, Pinquet, Guillén, and Bolancé (2001) and Bolancé, Guillén, and Pinquet (2003) introduced a dynamic element into the observed latent variable. Claims frequency was modeled using Poisson distribution, conditional on a latent variable that was lognormally distributed with an autoregressive order structure. Examining claims from a Spanish automobile insurer, they found evidence of positive serial dependencies. Purcaru and Denuit (2003) studied the type of dependence introduced through correlated latent variables and suggested using copulas to model the serial dependence of latent variables.

For our purposes, we explore the use of standard random-effects count models (see, e.g., Diggle, Heagarty, Liang, and Zeger 2002; Frees 2004). For these models, we use $\lambda_{it} = e_{it} \exp(\alpha_{\lambda i} + \mathbf{x}'_{it}\beta_\lambda)$ to be the conditional mean parameter for the $\{it\}$ observational unit. Here $\alpha_{\lambda i}$ is a time-constant latent random variable to account for the time dependencies and e_{it} is the amount of exposure, because a driver may have insurance coverage for only part of the year. With this, the frequency component likelihood for the ith subject can be expressed as

$$L_{F,i} = \int \Pr\bigl(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}|\alpha_{\lambda i}\bigr)\mathrm{f}(\alpha_{\lambda i})\, d\alpha_{\lambda i}.$$

Typically, a normal distribution is used for $\mathrm{f}(\alpha_{\lambda i})$, and this is our distributional choice. Furthermore, we assume that $(N_{i1}, \ldots, N_{iT_i})$ are independent, conditional on $\alpha_{\lambda i}$. Thus the conditional joint distribution for all observations from the ith subject is given by

$$\Pr\bigl(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}|\alpha_{\lambda i}\bigr) = \prod_{t=1}^{T_i} \Pr(N_{it} = n_{it}|\alpha_{\lambda i}).$$

With the Poisson distribution for counts, recall that we have $\Pr(N = k) = \lambda^k e^{-\lambda}/k!$, using $\lambda = \lambda_{it}$ for the mean parameter. We also use the negative binomial distribution with parameters p and r, so that $\Pr(N = k) = \binom{k+r-1}{r-1} p^r (1-p)^k$. Here $\sigma = r^{-1}$ is the dispersion parameter, and $p = p_{it}$ is related to the mean through $(1 - p_{it})/p_{it} = \lambda_{it}\sigma = \exp(\alpha_{\lambda i} + \mathbf{x}'_{it}\beta_\lambda)\sigma$.

Table 1. Frequency of claims

	Count						
	0	1	2	3	4	5	Total
Number	178,080	19,224	1,859	177	11	1	199,352
Percentage	89.3	9.6	.9	.1	0	0	100.0

To get a sense of the empirical observations for claim frequency, Table 1 shows the frequency of claims during the entire observation period. According to this table, there were a total of 199,352 observations, of which 89.3% had no claims. There are a total of 23,522 ($= 19{,}224 \times 1 + 1{,}859 \times 2 + 177 \times 3 + 11 \times 4 + 1 \times 5$) claims.

2.2.2 Claims Type Component. In Section 2.1 we described the three types of claims that may occur in any combination for a given accident: third-party injury, own damage, and third-party property. Conditional on having observed at least one type of claim, the random variable M describes the combination observed. Table 2 provides the distribution of M. We see that, for example, third-party injury (C_1) is the least prevalent. Moreover, Table 2 shows that all combinations of claims occurred in our data.

To incorporate explanatory variables, we model the claim type as a multinomial logit of the form

$$\Pr(M = m) = \frac{\exp(V_m)}{\sum_{s=1}^{7} \exp(V_s)}, \tag{1}$$

where $V_{itj,m} = \mathbf{x}_{it}' \beta_{M,m}$. This is known as a "selection" or "participation" equation in econometrics (see, e.g., Jones 2000). Note that for our application, the covariates do not depend on the accident number j nor on the claim type m, although we allow parameters ($\beta_{M,m}$) to depend on m.

2.2.3 Severity Component. Table 3 provides a first look at the severity component of our data. For each type of claim, we see that the standard deviation exceeds the mean. For nonnegative data, this suggests using distributions with fatter tails than the normal. Third-party injury claims, although the least frequent, have the strongest potential for large consequences. A total of 2,529 ($= 20{,}503 - 17{,}974$) claims for damages to the insured ("own damages") are censored, indicating that a formal mechanism for handling the censoring is important.

To accommodate the long-tailed nature of claims, we use the generalized beta of the second kind (GB2) for each claim type. This has density function

$$f_C(c) = \frac{\exp(\alpha_1 z)}{c|\sigma|\mathrm{B}(\alpha_1, \alpha_2)[1 + \exp(z)]^{\alpha_1 + \alpha_2}}, \qquad c \geq 0, \tag{2}$$

where $z = (\ln c - \mu)/\sigma$ and $\mathrm{B}(\alpha_1, \alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1 + \alpha_2)$, the usual Beta function. Here μ is a location parameter, σ is a scale parameter, and α_1 and α_2 are shape parameters. This distribution is well known in actuarial modeling of univariate loss distributions (see, e.g., Klugman et al. 2004). With four parameters, the distribution has great flexibility for fitting heavy-tailed data. Many distributions useful for fitting long-tailed distributions can be written as special or limiting cases of the GB2 distribution (see, e.g., McDonald and Xu 1995).

We use this distribution but allow scale and shape parameters to vary by type and thus consider α_{1k}, α_{2k}, and σ_k for $k = 1, 2, 3$. Despite the prominence of the GB2 in fitting distributions to univariate data, relatively few applications use the GB2 in a regression context. The first related work was from McDonald and Butler (1990) who used the GB2 with regression covariates to examine the duration of welfare spells. Beirlant, Goegebeur, Verlaak, and Vynckier (1998) demonstrated the usefulness of the Burr XII distribution, a special case of the GB2 with $\alpha_1 = 1$, in regression applications. Recently, Sun, Frees, and Rosenberg (2008) used the GB2 in a longitudinal data context to forecast nursing home utilization. For a general approach on how to include covariates, Beirlant, Goegebeur, Segers, and Teugels (2004) suggested allowing all of the parameters (of a Burr XII distribution) to depend on covariates. We use a simpler specification and parameterize the location parameter as $\mu_k = \mathbf{x}' \beta_{C,k}$. This is due in part to the interpretability of parameters; if C is a GB2 random variable, then straightforward calculations show that $\beta_{C,k,j} = \partial \ln \mathrm{E}(C|\mathbf{x})/\partial x_j$, meaning that we may interpret the regression coefficients as proportional changes.

To accommodate dependencies among claim types, we use a parametric copula (see Frees and Valdez 1998 for an introduction to copulas). Suppressing the $\{it\}$ subscripts, we can write the joint distribution of claims (C_1, C_2, C_3) as

$$\begin{aligned} \mathrm{F}(c_1, c_2, c_3) &= \Pr(C_1 \leq c_1, C_2 \leq c_2, C_3 \leq c_3) \\ &= \Pr\big(\mathrm{F}_1(C_1) \leq \mathrm{F}_1(c_1), \mathrm{F}_2(C_2) \leq \mathrm{F}_2(c_2), \\ &\quad \mathrm{F}_3(C_3) \leq \mathrm{F}_3(c_3)\big) \\ &= \mathrm{H}\big(\mathrm{F}_1(c_1), \mathrm{F}_2(c_2), \mathrm{F}_3(c_3)\big). \end{aligned}$$

Table 2. Distribution of claims, by claim type observed

	Value of M, claim type							
	1 (C_1)	2 (C_2)	3 (C_3)	4 (C_1, C_2)	5 (C_1, C_3)	6 (C_2, C_3)	7 (C_1, C_2, C_3)	Total
Number	102	17,216	2,899	68	18	3,176	43	23,522
Percentage	.4	73.2	12.3	.3	.1	13.5	.2	100.0

Table 3. Summary statistics of claim losses, by type of claim

Statistic	Third-party injury (C_1)	Own damage (C_2)		Third-party property (C_3)
		Noncensored	All	
Number	231	17,974	20,503	6,136
Mean	12,781.89	2,865.39	2,511.95	2,917.79
Standard deviation	39,649.14	4,536.18	4,350.46	3,262.06
Median	1,700	1,637.40	1,303.20	1,972.08
Minimum	10	2	0	3
Maximum	336,596	367,183	367,183	56,156.51

NOTE: Censored own damages claims have values of 0.

Here the marginal distribution of C_j is given by $\mathrm{F}_j(\cdot)$, and $\mathrm{H}(\cdot)$ is the copula linking the marginals to the joint distribution. We use a trivariate t-copula with an unstructured correlation matrix. The multivariate t-copula has been shown to work well on loss data (see Frees and Wang 2005). As a member of the elliptical family of distributions, it has the important property of preserving the family under the marginals (see Landsman and Valdez 2003), so that when we observe only a subset of the three types, we still can use the t-copula.

The likelihood, developed formally in the Appendix, depends on the association among claim amounts. To see this, suppose that all three types of claims are observed ($M = 7$) and that each is uncensored. In this case the joint density is

$$\mathrm{f}_{uc,123}(c_1, c_2, c_3) = \mathrm{h}_3\big(\mathrm{F}_{it,1}(c_1), \mathrm{F}_{it,2}(c_2), \mathrm{F}_{it,3}(c_3)\big) \prod_{k=1}^{3} \mathrm{f}_{it,k}(c_k), \quad (3)$$

where $\mathrm{f}_{it,k}$ is the density associated with the $\{it\}$ observation and the kth type of claim and $\mathrm{h}_3(\cdot)$ is the probability density function for the trivariate t-copula. Specifically, we can define the density for the trivariate t-distribution as

$$t_3(\mathbf{z}) = \frac{\Gamma(\frac{r+3}{2})}{(r\pi)^{3/2}\Gamma(\frac{r}{2})\sqrt{\det(\Sigma)}} \left(1 + \frac{1}{r}\mathbf{z}'\Sigma^{-1}\mathbf{z}\right)^{-(r+3)/2}, \quad (4)$$

and the corresponding copula as

$$\mathrm{h}_3(u_1, u_2, u_3) = t_3\big(\mathrm{G}_r^{-1}(u_1), \mathrm{G}_r^{-1}(u_2), \mathrm{G}_r^{-1}(u_3)\big) \prod_{k=1}^{3} \frac{1}{\mathrm{g}_r(\mathrm{G}_r^{-1}(u_k))}. \quad (5)$$

Here G_r is the distribution function for a t-distribution with r degrees of freedom, G_r^{-1} is the corresponding inverse, and g_r is the probability density function. Using the copula in (3) allows us to compute the likelihood. We also consider the case where $r \to \infty$, so that the multivariate t-copula becomes the well-known Normal copula.

3. DATA ANALYSIS

3.1 Covariates

As noted in Section 2.1, several characteristics are available to explain and predict automobile accident frequency, type, and severity. These characteristics include vehicle characteristics, such as type and age, as well as person-level characteristics, such as age, gender, and previous driving experience. Table 4 summarizes these characteristics.

The description in Section 2 uses a generic vector $\mathbf{x}$ to indicate the availability of covariates that are common to the three outcome variables. In our investigation we found that the usefulness of covariates depended on the type of outcome and used a parsimonious selection of covariates for each type. In the following sections we describe how the covariates can be used to fit our frequency, type, and severity models. For congruence with Section 2, the data summaries refer to the full data set comprising the years 1993–2001 inclusive; however, when fitting models, we used only 1993–2000 inclusive. We reserved observations in year 2001 for out-of-sample validation, as discussed in Section 4.

3.2 Fitting the Frequency Component Model

We begin by displaying summary statistics to suggest the effects of each covariate in Table 4 on claim frequency. We then compare fitted models that summarize all of these effects in a single framework.

Table 5 displays the claims frequency distribution over time. For this company, the number of insurance policies increased

Table 4. Description of covariates

Covariate	Description
Year	The calendar year, from 1993–2001, inclusive
Vehicle type	The type of vehicle being insured, either automobile (A) or other (O)
Vehicle age	The age of the vehicle, in years, grouped into seven categories
Gender	The policyholder's gender, either male or female
Age	The age of the policyholder, in years, grouped into seven categories
NCD	No claims discount, based on the previous accident record of the policyholder; the higher the discount, the better is the prior accident record

Table 5. Number and percentages of claims, by count and year

Count	Percentage by year									Number	Percent of total
	1993	1994	1995	1996	1997	1998	1999	2000	2001		
0	91.5	89.5	89.8	92.6	92.8	90.8	88.0	89.2	87.8	178,080	89.3
1	7.9	9.6	9.2	7.0	6.7	8.4	10.6	9.8	11.0	19,224	9.6
2	.5	.9	.9	.4	.5	.7	1.3	.9	1.1	1,859	.9
3	.1	.1	.1	0	0	.1	.1	.1	.1	177	.1
4		0					0	0	0	11	0
5			0							0	0
Number by year	4,976	5,969	5,320	8,562	19,344	19,749	28,473	44,821	62,138	199,352	100.0

significantly from 1993 to 2001. We also note that the percentage of no accidents was lower in later years. This is not an uncommon situation in the insurance industry, where a company may decide to relax its underwriting standards to gain additional business in a competitive marketplace. Typically, relaxed underwriting standards means acceptance of more business at the price of poorer overall experience.

Table 6 shows the effects of vehicle characteristics on claim count. The "Automobile" category has lower overall claims experience. The "Other" category consists primarily of (commercial) goods vehicles, as well as weekend and hire cars. The vehicle age shows nonlinear effects. Here we see low claims for new cars, with initially increasing accident frequency over time. But the accident frequencies are relatively low for vehicles in operation for long periods. There also are some important interaction effects between vehicle type and age not shown here. Nonetheless, Table 6 clearly suggests the importance of these two variables on claim frequencies.

Table 7 shows the effects of person-level characteristics [gender, age, and no claims discount (NCD)] on the frequency distribution. Person-level characteristics were largely unavailable for commercial use vehicles, and so Table 7 presents summary statistics for only those observations having automobile coverage with the requisite gender and age information. When we restricted consideration to (private use) automobiles, relatively few policies contained no gender and age information.

Table 7 suggests that driving experience was roughly similar in males and females. This company insured very few young drivers, so the young male driver category, which typically has extremely high accident rates in most automobile studies, is less important for our data. Nonetheless, Table 7 suggests strong age effects, with older drivers having better driver experience. Table 7 also demonstrates the importance of the NCD. As anticipated, drivers with better previous driving records enjoy a higher NCD and have fewer accidents. Although the results are not reported here, we also considered interactions among these three variables.

As part of the examination process, we investigated interaction terms among the covariates and nonlinear specifications. After examining the data in more detail, we report five fitted count models: a basic Poisson model without covariates, Poisson and negative binomial models with covariates, and their counterparts that incorporate random effects. For the latter four models, we used the same covariates to form the systematic component, $\mathbf{x}_{it}'\beta_\lambda$. We used maximum likelihood to fit each model and empirical Bayes to predict the random intercepts.

Table 8 compares these five fitted models, providing predictions for each level of the response variable. To summarize the overall fit, we report a Pearson chi-squared goodness-of-fit statistic. As anticipated, the Poisson performed much better with covariates, and the negative binomial fit better than the Poisson. Somewhat surprisingly, the random-effects models fared poorly

Table 6. Number and percentages of claims, by vehicle type and age

	Percentage by count						Number	Percent of total
	Count = 0	Count = 1	Count = 2	Count = 3	Count = 4	Count = 5		
Vehicle type								
Other	88.6	10.1	1.1	.1	0	0	43,891	22.0
Automobile	89.5	9.5	.9	.1	0		155,461	78.0
Vehicle age (in years)								
0	91.4	7.9	.6	0	0		58,301	29.2
1	86.3	12.2	1.3	.2	0		44,373	22.3
2	88.8	10.1	1.1	.1			20,498	10.3
3–5	89.2	9.7	1.0	.1	0		41,117	20.6
6–10	90.1	8.9	.9	.1		0	33,121	16.6
11–15	91.4	7.6	.7	.2			1,743	.9
16 and older	89.9	8.5	1.5				199	.1
Number by count	178,080	19,224	1,859	177	11	1	199,352	100.0

Table 7. Number and percentages of claims, by gender, age, and NCD, for automobile policies

	Percentage by count							
	Count = 0	Count = 1	Count = 2	Count = 3	Count = 4	Count = 5	Number	Percent of total
Gender								
Female	89.7	9.3	.9	.1	0		34,190	22.0
Male	89.5	9.5	.9	.1	0	0	121,271	78.0
Person age (in years)								
21 and younger	86.9	12.4	.7				153	.1
22–25	85.5	12.9	1.4	.2			3,202	2.1
26–35	88.0	10.8	1.1	.1	0	0	44,134	28.4
36–45	90.1	9.1	.8	.1	0		63,135	40.6
46–55	90.4	8.8	.8	.1	0		34,373	22.1
56–65	90.7	8.4	.9	.1			9,207	5.9
66 and over	92.8	7.0	.2	.1			1,257	.8
NCD								
0	87.7	11.1	1.1	.1	0		37,139	23.9
10	87.8	10.8	1.2	.1	0		13,185	8.5
20	89.1	9.8	1.0	.1			14,204	9.1
30	89.1	10.0	.9	.1			12,558	8.1
40	89.8	9.3	.9	.1	0		10,540	6.8
50	91.0	8.3	.7	.1		0	67,835	43.6
Number by count	139,183	14,774	1,377	123	3	1	155,461	100.0

compared with the negative binomial model. Recall, however, that our data set is from an insurance company with a substantial turnover of policies. Thus we interpret the findings of Table 8 to mean that the negative binomial distribution well captures the heterogeneity in the accident frequency distribution and that the NCD variable captures claims history. Thus for this particular company, the additional complexity of the random-effects portion of each model is not warranted.

3.3 Fitting the Claim Type Model

The claim type model helps the analyst assess the type of claim, when a claim has occurred. As described in Section 2.2.2, we considered seven combinations of third-party injury (C_1), own damages (C_2), and third-party property (C_3). For our data set, we found that person-level characteristics (gender, age, and NCD) did not seem to influence claim type significantly.

Vehicle characteristics and year did seem to influence claim type significantly, as suggested by Table 9. As with the frequency model, we found that whether or not a vehicle is an automobile is an important determinant of claim type. For automobiles, vehicle age was important, although we do not require as much detail as in the frequency model. For claims type, we considered automobiles with vehicle age >2 to be "old" and all others to be "new." When examining a detailed distribution of type over time, we found a sharp break between 1996 and 1997. Table 9 provides the distribution of claims types by level of these variables, suggesting their importance as determinants. For example, we see that overall, 73.2% of claims fell into the own damages (C_2) category, although only 63.4% for nonautomobiles compared with 76.3% for automobiles.

Table 10 summarizes the performance of some alternative multinomial logit models used to fit claim types. Here the number of parameters from the model fit and minus twice the log-likelihood is reported. Table 10 shows that the binary variable indicating whether or not the vehicle is an automobile was an important determinant, whereas the addition of gender does not seem to contribute much. Using Year as a continuous variable was not as useful as the automobile variable, although the binary variable Year1996 (that distinguishes between before 1997

Table 8. Comparison of fitted frequency models based on the 1993–2000 in-sample data

Count	Observed	No covariates	Poisson	Negative binomial	Random effects Poisson	Random effects negative binomial
0	123,528	123,152.6	123,190.9	123,543.0	124,728.4	125,523.4
1	12,407	13,090.4	13,020.1	12,388.1	11,665.7	7,843.1
2	1,165	920.6	946.7	1,164.1	775.5	2,189.5
3	109	48.3	53.6	107.8	42.3	854.1
4	4	2.0	2.5	10.0	2.1	374.4
5	1	1.6	2.0	.9	1.6	178.8
Chi-squared goodness of fit		125.2	101.8	9.0	228.4	73,626.7

Table 9. Distribution of claim type, by vehicle characteristics and year

M	Claim type	Nonauto (other)	Auto	Old vehicle	New vehicle	Before 1997	After 1996	Overall
1	C_1	.7	.4	.6	.3	1.3	.3	.4
2	C_2	63.4	76.3	69.4	75.4	62.5	74.4	73.2
3	C_3	23.7	8.8	15.1	10.7	21.2	11.3	12.3
4	C_1, C_2	.2	.3	.4	.2	.5	.3	.3
5	C_1, C_3	.1	.1	.1	0	.3	0	.1
6	C_2, C_3	11.8	14.0	14.2	13.1	14.0	13.4	13.5
7	C_1, C_2, C_3	.1	.2	.2	.2	.1	.2	.2
	Counts	5,608	17,914	8,750	14,772	2,421	21,101	23,522

and after 1996) was important. Similarly, the binary variable VehAge2, which distinguishes between a vehicle age <3 and >2, was useful. We also explored interactions and other combinations of variables, not reported here. Using the three binary variables, A, VehAge2, and Year1996, provided the best fit.

3.4 Fitting the Severity Component Model

As noted in Section 2.2.3, it is important to consider long-tailed distributions when fitting models of insurance claims. Table 3 provides some evidence, and Figure 1 reinforces this concept with an empirical histogram for each type of claim suggesting the importance of long-tailed distributions.

In Section 2.2.3 we discussed the appropriateness of the GB2 distribution as a model for losses. Figure 2 provides *qq* plots, based on residuals after the introduction of covariates. Here we see that this distribution fits the data well. The poorest part of the fit is in the lower quantiles. But for insurance applications, most of the interest is in the upper tails of the distribution (corresponding to large claim amounts), so that poor fit in the lower quantiles is of less concern.

An advantage of the copula construction is that each of the marginal distributions can be specified in isolation of the others and then be joined by the copula. Thus we fit each type of claim amount using the GB2 regression model described in Section 2.2.3. Standard variable selection procedures were used for each marginal, and the resulting fitted parameter estimates are summarized in Table 11 in the "Independence" column. As noted in Section 2.2.3, all three parameters of the GB2 distribution varied by claim type. In the interest of parsimony, no covariates were used for the 231 injury claims, whereas an intercept, Year, and vehicle age (VehAge2) were used for third-party property, and an intercept, year (Year1996), vehicle age (VehAge2), and insured's age were used for own damage. For insured's age, Age2 is a binary variable indicating that a driver is 26–55, and Age3 indicates that a driver is 56 over. For own damage, a censored likelihood was used. All parameter estimates were calculated by maximum likelihood; see the Appendix for a detailed description.

Using the parameter estimates from the independence model as initial values, we then estimated the full copula model by maximum likelihood. We used two choices of copulas, the standard normal (Gaussian) copula and the t-copula. An examination of the likelihood and information statistics shows that the normal copula model was an improvement over the independence model; however, the t-copula showed little improvement over the normal copula. These models are embedded within one another in the sense that the normal copula with zero correlation parameters reduces to the independence model and the t-copula tends to the normal copula as the degrees of freedom r tends to infinity. Thus it is reasonable to compare the likelihoods and argue that the normal copula is statistically significantly better than the independence copula using a likelihood ratio test. Furthermore, although a formal hypothesis test is not readily available, a quick examination of the information statistics shows that the extra complexity from the t-copula is not warranted, and the simpler normal copula is preferred.

Table 10. Comparison of fit of alternative claim type models

Model Variables	Number of parameters	-2 Log-likelihood
Intercept only	6	25,465.3
Automobile (A)	12	24,895.8
A and Gender	24	24,866.3
Year	12	25,315.6
Year1996	12	25,259.9
A and Year1996	18	24,730.6
VehAge2 (old vs. new)	12	25,396.5
VehAge2 and A	18	24,764.5
A, VehAge2, and Year1996	24	24,646.6

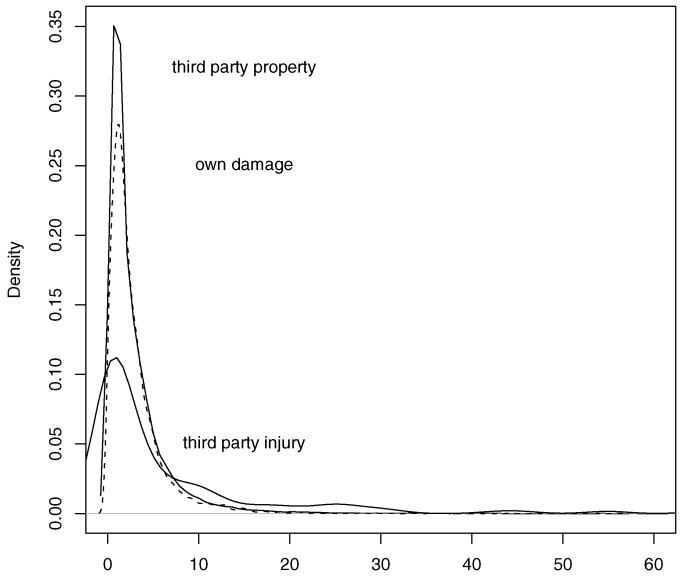

Figure 1. Density of losses by claim type, in thousands.

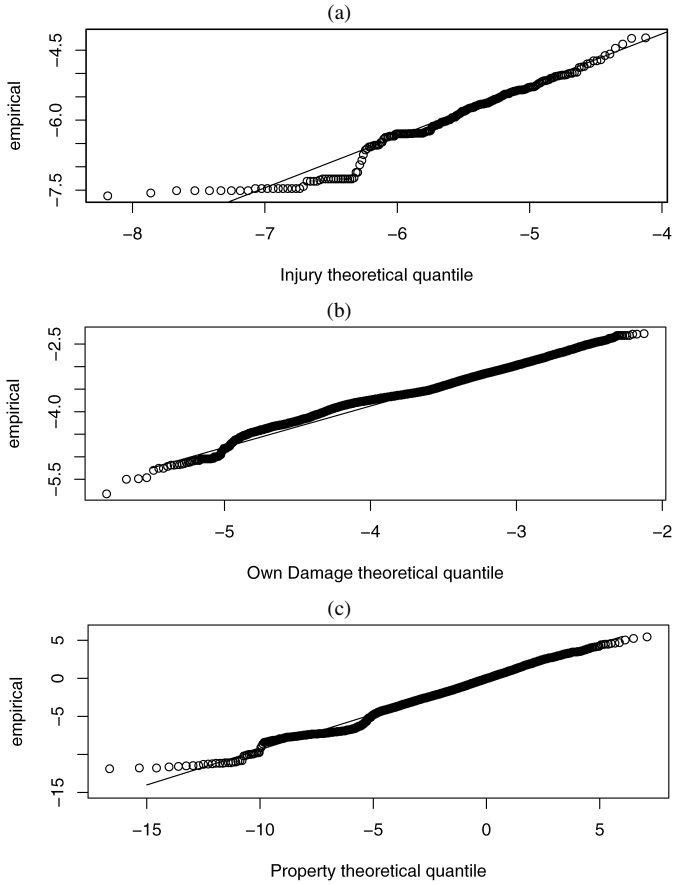

Figure 2. Quantile–quantile plots for fitting the GB2 distributions. (a) Injury theoretical quantile. (b) Own damage theoretical quantile. (c) Property theoretical quantile.

We note that there are different perspectives on the choice of the degrees of freedom for the t-copula. One argument is to choose the degrees of freedom as one would for a standard analysis of variance procedure, as the number of observations minus the number of parameters. Alternatively, the degrees of freedom could be chosen to maximize the likelihood but restricted to be an integer. Because of the widespread availability of modern computational tools, we determined the degrees of freedom parameter, r, by maximum likelihood without restricting it to be an integer.

From Table 11, we also see that parameter estimates are qualitatively similar under each copula. Interestingly, the correlation coefficient estimates indicate significant relationships among the three claim types. Although the data are not presented here, it turns out that these relationships were not evident when just the raw statistical summaries were examined. We also note, for the own damage and third-party property claims, that estimators of first shape parameter, α_{21} and α_{31}, were statistically different from 1. This indicates that the additional complexity of the GB2 compared with the simpler Burr XII is warranted. There is not a statistically significant difference for injury possibly due to the fact that we had only 231 injury claims to assess.

4. PREDICTION

As noted in Section 1, an important application of the modeling process for the actuary involves predicting claims arising from insurance policies. We illustrate the process in three different ways: (a) prediction based on an individual observation, (b) determination of expected functions of claims over different policy scenarios, and (c) a predictive distribution for a portfolio of policies.

4.1 Test Case

It is common for actuaries to examine one or more "test cases" when setting premium scales or reserves. These test cases also help actuaries in enhance their understanding of the suitability of the models considered. The first step is to generate a prediction of the claims frequency model that we fit in Section 3.2. Because this problem has been well discussed in the literature (see, e.g., Bolancé et al. 2003), we focus on prediction conditional on the occurrence of a claim, that is, $N = 1$. To illustrate what an actuary can learn when predicting based on an individual observation, we chose an observation from our out-of-sample period year 2001. Claim number 1,901 from our database involves a policy for a 53-year-old male driving a 1999 Mercedes Benz with a 1,998–cubic inch capacity engine. The driver enjoys the largest NCD, 50, and has a comprehensive policy with a $750 deductible for the own damage portion.

Using the claim type model in Section 3.3, it is straightforward to generate predicted probabilities for claim type, as shown in Table 12.

We then generated 5,000 simulated values of total claims. For each simulation, we used three random variates to generate a realization from the trivariate joint distribution function of claims (see, e.g., DeMarta and McNeil 2005 for techniques on simulating realizations using t-copulas). After adjusting for the own damage deductible, we then combined these three random claims using an additional random variate for the claim type into a single predicted total claim for the policy. Figure 3 summarizes the result of this simulation. This figure underscores the long-tailed nature of this predictive distribution, an important point for the actuary when pricing policies and setting reserves. For reference, it turned out that the actual claim for this policy was $2,453.95, corresponding to the 56th percentile of the predictive distribution.

4.2 Coverage Limits

For a second anticipated application of our models, we present several measures of expected functions of claims over different policy scenarios. As financial analysts, actuaries become involved in setting policy coverage parameters and the relationship of these parameters with premiums and reserves; for example, it is common for automobile policies to have upper limits beyond which the insurer is no longer responsible. These coverage limits may depend on claim type or the total amount of claims. For example, if 50,000 is such an upper limit and C represents total claims, then the insurance company will be responsible for the limited value, $Y = \min(C, 50{,}000)$. Naturally, the random variable Y has a distribution, and the actuary must set a contract price based on this distribution. The differences between the claims amount C and limited value Y are largely influenced by the tails of the claims distribution; this is one motivation for considering a long-tailed distributions such as the GB2.

Table 11. Fitted copula model

Parameter	Type of copula		
	Independence	Normal copula	t-copula
Third-party injury			
σ_1	$1.316_{(.124)}$	$1.320_{(.138)}$	$1.320_{(.120)}$
α_{11}	$2.188_{(1.482)}$	$2.227_{(1.671)}$	$2.239_{(1.447)}$
α_{12}	$500.069_{(455.832)}$	$500.068_{(408.440)}$	$500.054_{(396.655)}$
$\beta_{C,1,1}$ (intercept)	$18.430_{(2.139)}$	$18.509_{(4.684)}$	$18.543_{(4.713)}$
Own damage			
σ_2	$1.305_{(.031)}$	$1.301_{(.022)}$	$1.302_{(.029)}$
α_{21}	$5.658_{(1.123)}$	$5.507_{(.783)}$	$5.532_{(.992)}$
α_{22}	$163.605_{(42.021)}$	$163.699_{(22.404)}$	$170.382_{(59.648)}$
$\beta_{C,2,1}$ (intercept)	$10.037_{(1.009)}$	$9.976_{(.576)}$	$10.106_{(1.315)}$
$\beta_{C,2,2}$ (VehAge2)	$.090_{(.025)}$	$.091_{(.025)}$	$.091_{(.025)}$
$\beta_{C,2,3}$ (Year1996)	$.269_{(.035)}$	$.274_{(.035)}$	$.274_{(.035)}$
$\beta_{C,2,4}$ (Age2)	$.107_{(.032)}$	$.125_{(.032)}$	$.125_{(.032)}$
$\beta_{C,2,5}$ (Age3)	$.225_{(.064)}$	$.247_{(.064)}$	$.247_{(.064)}$
Third-party property			
σ_3	$.846_{(.032)}$	$.853_{(.031)}$	$.853_{(.031)}$
α_{31}	$.597_{(.111)}$	$.544_{(.101)}$	$.544_{(.101)}$
α_{32}	$1.381_{(.372)}$	$1.534_{(.402)}$	$1.534_{(.401)}$
$\beta_{C,3,1}$ (intercept)	$1.332_{(.136)}$	$1.333_{(.140)}$	$1.333_{(.139)}$
$\beta_{C,3,2}$ (VehAge2)	$-.098_{(.043)}$	$-.091_{(.042)}$	$-.091_{(.042)}$
$\beta_{C,3,3}$ (Year1)	$.045_{(.011)}$	$.038_{(.011)}$	$.038_{(.011)}$
Copula			
ρ_{12}		$.018_{(.115)}$	$.018_{(.115)}$
ρ_{13}		$-.066_{(.112)}$	$-.066_{(.111)}$
ρ_{23}		$.259_{(.024)}$	$.259_{(.024)}$
r			$193.055_{(140.648)}$
Model fit statistics			
Log-likelihood	−31,006.505	−30,955.351	−30,955.281
Number of parameters	18	21	22
Akaike information criterion	62,049.010	61,952.702	61,954.562

NOTE: Standard errors are in parentheses.

Table 13 shows the results of sample calculations for our illustrative policy, observation number 1,901. We now allow for the possibility of multiple claims, however. Using the negative binomial model developed in Section 3.2, we predicted the probability of up to five claims. The probability of six or more claims was negligible for this case. We generated $25{,}000 = 5 \times 5{,}000$ claims of each type, using the process described in our test scenario. We applied upper limits to each claim type (representative values are given in Table 13) and applied an overall upper limit to each potential realization of the total amount (over multiple claims). Finally, we weighted each realization by probabilities of the number of claims.

Table 13 gives summary statistics based on 5,000 simulations. This table provides the mean, 25th, median, and 75th percentiles for each simulated distribution, showing the effect of different coverage limits. For the three upper limits by type, the injury upper limit had the greatest effect due to the long-tailed nature of its severity relative to the other types. This is interesting, because injury also has the least probability of occurring, and so this result was not obvious a priori. An overall limit has a greater effect on the limited value variable than any individual limit, because an overall limit applies to all three types as well as to the total amount of claims, not just on a per claim occurrence. We also see that a limit of $50,000 on each of three types of claim produced approximately the same limited value variable distribution compared with an overall limit of $50,000. This is not surprising, because (as shown Fig. 1) events of this size are relatively rare. This means that the insurance company can substitute a single overall contractual limit that is easier to administer and enjoy the same protection as if it had required three separate limits.

Table 12. Prediction of claim type for claim number 1,901

	Claim type							
	(C_1)	(C_2)	(C_3)	(C_1,C_2)	(C_1, C_3)	(C_2, C_3)	(C_1, C_2, C_3)	Total
Percentage	1.17	53.08	33.79	.24	.31	11.32	.09	100.0

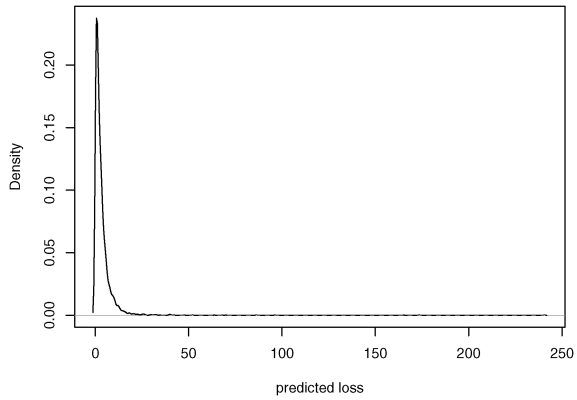

Figure 3. Simulated predictive distribution for observation 1,901 from our data set.

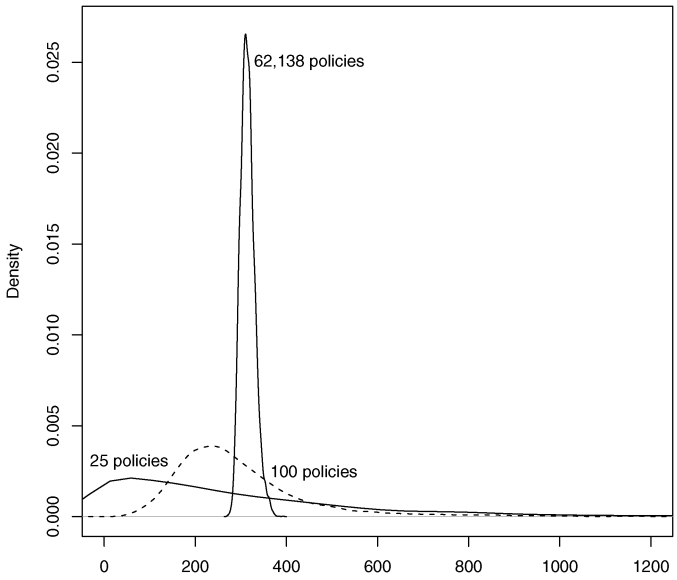

Figure 4. Predictive distributions of average claims for three 2001 portfolios. The long-tailed, flat distribution is for the portfolio of 25 policies, the peaked distribution is for the portfolio of 62,138 policies, and the dashed line gives the distribution for the intermediate case of 100 policies.

4.3 Portfolio Prediction

As shown in Table 5, in 2001 a total of 62,138 policies were reserved for out-of-sample validation. From these policies, we created two portfolios by randomly selecting 25 and 100 policies. The full set of 62,138 policies provided our third portfolio.

For each policy in a portfolio, we used the covariates known at the beginning of the year to randomly generate a number of claims. Then for each claim generated, we used the procedures described in Section 4.2 to generate the type and amount of claims. We generated 5,000 simulations for each policy.

Figure 4 compares the predictive distributions of the three portfolios. The distributions are kernel density estimates of the predictive density based on the 5,000 simulations. So that the portfolios would be roughly comparable, we divided by the number of policies in each portfolio (25, 100, and 62,138); thus Figure 4 gives the predictive distribution of the average portfolio claim. In the figure, the solid line represents the distribution for the 25 policies. This distribution is long-tailed and skewed to the right, reflecting the fact that it is an average of only 25 policies. As the portfolio size increases (to 100 and then to 62,138), the effect of the central limit theorem becames more apparent: the distribution becomes more symmetric, and the variability grows smaller.

Figure 5 presents more detail of the full 2001 portfolio of 62,138 policies. Here we no longer rescale by the number of policies, and we express potential claims in millions of Singaporean dollars. Although based on highly skewed individual policies distributions, the distribution of the portfolio sum is roughly symmetric, with a slight right skewness. For reference, it turned out that the actual amount of out-of-sample claims was 22.195 million Singaporean dollars, corresponding to the 96th percentile of the simulated distribution. Our estimate for this company may have been high, but we found that this company's business grew considerably in 2000 and 2001. In a competitive market, this is probably due to loosening of underwriting standards, which could have had a spiraling effect on claims in later years. The fact that we could give a predictive distribution in accordance with reality when knowing only data provided to a professional association (the GIA) and knowing so little about the actual competitive insurance environment in Singapore is a great tribute to the potential of the statistical methods introduced in this article.

Table 13. Summary statistics of the excess: Simulation based on 5,000 replications

Limit							
Injury	Own damage	Property	Overall	Mean	25th percentile	Median	75th percentile
None	None	None	None	542.33	91.96	293.04	652.87
25,000	None	None	None	507.04	91.40	292.61	649.69
50,000	None	None	None	518.04	91.96	292.80	652.51
None	25,000	None	None	540.80	91.96	293.04	652.87
None	50,000	None	None	542.33	91.96	293.04	652.87
None	None	25,000	None	537.19	91.60	293.04	651.91
None	None	50,000	None	540.26	91.96	293.04	652.87
None	None	None	25,000	499.31	91.37	292.61	649.37
None	None	None	50,000	515.68	91.96	292.80	652.51
None	None	None	100,000	528.07	91.96	293.04	652.51
50,000	50,000	50,000	None	515.97	91.96	292.80	652.51

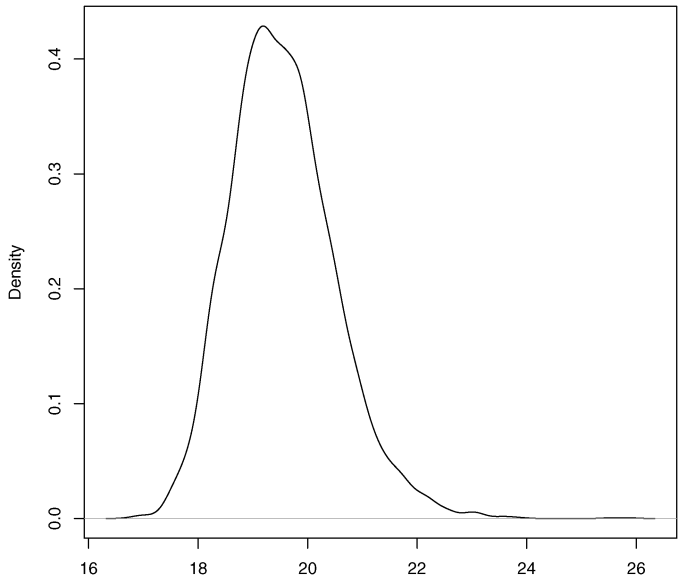

Figure 5. Predictive distribution of total losses for a portfolio of 62,138 policies. This portfolio corresponds to all 2001 policies reserved for out-of-sample validation.

5. SUMMARY AND CONCLUDING REMARKS

One way to think of the insurance claims data used in this article is as a set of multivariate longitudinal responses with covariate information. The longitudinal nature is because vehicles are observed over time. For each vehicle, there are three responses in a given year: claims amount for injury, own damage, and third-party property damage. One approach to modeling this data set would be to use techniques from multivariate longitudinal data (see, e.g., Fahrmeir and Tutz 2001); however, as we have pointed out, in most years policyholders do not incur a claim, resulting in many repeated 0's (see, e.g., Olsen and Shafer 2001), and when a claim does occur, the distribution is long-tailed. Both of these features are not readily accommodated using standard multivariate longitudinal data models that generally assume that data are from an exponential family of distributions.

Other possible approaches to modeling the data set include Bayesian predictive modeling (see de Alba 2002; Verrall 2004 for recent actuarial applications). Another approach would be to model the claims count for each of the three types jointly and thus consider a trivariate Poisson process. This was the approach taken by Pinquet (1998) when considering two types of claims, those at fault and no-fault. This approach is comparable to that taken in this article, in that linear combinations of Poisson process are also Poisson processes. We have chosen to reorganize this multivariate count data into count and type events because we feel that this approach is more flexible and easier to implement, especially when the dimension of the types of claims increases.

The main contribution in this article is the introduction of a multivariate claims distribution for handling long-tailed, related claims using covariates. We used the GB2 distribution to accommodate the long-tailed nature of claims while at the same time, allowing for covariates. As an innovative approach, we introduced copulas to allow for relationships among different types of claims.

The focus of our illustrations in Section 4 was on predicting total claims arising from an insurance policy on a vehicle. We note that our model is sufficiently flexible to allow the actuary to focus on a single type of claim. This would be of interest when, for example, an actuary is designing an insurance contract and is interested in the effect of different deductibles or policy limits on "own damages" types of claims. We also have shown how to predict the distribution for a portfolio of insurance policies that would be of concern to an insurer for pricing or reserving considerations.

The modeling approach developed in this article is sufficiently flexible to handle our complex data. Nonetheless, we acknowledge that many improvements can be made. In particular, we did not investigate potential explanations for the lack of balance in our data; we implicitly assumed that the lack of balance in our longitudinal framework was due to data that were missing at random (Little and Rubin 1987). It is well known in longitudinal data modeling that attrition and other sources of imbalance may seriously affect statistical inference. This is an area for future investigation.

APPENDIX: SEVERITY LIKELIHOOD

Consider the seven different combinations of claim types arising when a claim is made. For claim types $M = 1, 3, 5$, no censoring is involved, and we may simply integrate out the effects of the types not observed. Thus, for example, for $M = 1, 3$, we have the likelihood contributions $\mathrm{L}_1(c_1) = \mathrm{f}_1(c_1)$ and $\mathrm{L}_3(c_3) = \mathrm{f}_3(c_3)$. The subscript of the likelihood contribution L refers to the claim type. For claim type $M = 5$, there also is no own damage amount, so that the likelihood contribution is given by

$$\begin{aligned}\mathrm{L}_5(c_1, c_3) &= \int_0^\infty \mathrm{h}_3\big(\mathrm{F}_1(c_1), \mathrm{F}_2(z), \mathrm{F}_3(c_3)\big)\mathrm{f}_1(c_1)\mathrm{f}_3(c_3)\mathrm{f}_2(z)\,dz \\ &= \mathrm{h}_2\big(\mathrm{F}_1(c_1), \mathrm{F}_3(c_3)\big)\mathrm{f}_1(c_1)\mathrm{f}_3(c_3) \\ &= \mathrm{f}_{uc,13}(c_1, c_3),\end{aligned}$$

where h_2 is the density of the bivariate t-copula, which has the same structure as the trivariate t-copula given in (5). Note that we are using the important property that a member of the elliptical family of distributions (and thus elliptical copulas) is preserved under the marginals.

The cases $M = 2, 4, 6, 7$ involve own damage claims; thus we need to allow for the possibility of censoring. Let c_2^* be the unobserved loss and let $c_2 = \max(0, c_2^* - d)$ be the observed claim. Furthermore, define

$$\delta = \begin{cases} 1 & \text{if } c_2^* \le d \\ 0 & \text{otherwise} \end{cases}$$

to be a binary variable that indicates censoring. Thus the familiar case where $M = 2$ is given by

$$\begin{aligned}\mathrm{L}_2(c_2) &= \begin{cases} \mathrm{f}_2(c_2 + d)/(1 - \mathrm{F}_2(d)) & \text{if } \delta = 0 \\ \mathrm{F}_2(d) & \text{if } \delta = 1 \end{cases} \\ &= \left\{ \left[\frac{\mathrm{f}_2(c_2 + d)}{1 - \mathrm{F}_2(d)} \right]^{1-\delta} (\mathrm{F}_2(d))^{\delta}. \right.\end{aligned}$$

For the case where $M = 6$, we have

$$\mathrm{L}_6(c_2, c_3) = \left[\frac{\mathrm{f}_{uc,23}(c_2 + d, c_3)}{1 - \mathrm{F}_2(d)} \right]^{1-\delta} (\mathrm{H}_{c,23}(d, c_3))^{\delta},$$

where

$$\mathrm{H}_{c,23}(d, c_3) = \int_0^d \mathrm{h}_2\big(\mathrm{F}_2(z), \mathrm{F}_3(c_3)\big)\mathrm{f}_3(c_3)\mathrm{f}_2(z)\,dz.$$

It is not difficult to show that this also can be expressed as

$$\mathrm{H}_{c,23}(d, c_3) = \mathrm{f}_3(c_3)\mathrm{H}_2\big(\mathrm{F}_2(d), \mathrm{F}_3(c_3)\big).$$

The case where $M = 4$ follows in the same fashion, reversing the roles of types 1 and 3. The more complex case where $M = 7$ is given by

$$\mathrm{L}_7(c_1, c_2, c_3) = \left[\frac{\mathrm{f}_{uc,123}(c_1, c_2 + d, c_3)}{1 - \mathrm{F}_2(d)}\right]^{1-\delta} (\mathrm{H}_{c,123}(c_1, d, c_3))^{\delta},$$

where $\mathrm{f}_{uc,123}$ is as given in (3) and

$$\mathrm{H}_{c,123}(c_1, d, c_3) = \int_0^d \mathrm{h}_3\big(\mathrm{F}_1(c_1), \mathrm{F}_2(z), \mathrm{F}_3(c_3)\big)\mathrm{f}_1(c_1)\mathrm{f}_3(c_3)\mathrm{f}_2(z)\,dz.$$

With these definitions, the total severity log-likelihood for each observational unit is $\log(\mathrm{L}_S) = \sum_{j=1}^{7} I(M = j)\log(\mathrm{L}_j)$.

[Received December 2005. Revised December 2007.]

REFERENCES

Angers, J.-F., Desjardins, D., Dionne, G., and Guertin, F. (2006), "Vehicle and Fleet Random Effects in a Model of Insurance Rating for Fleets of Vehicles," *ASTIN Bulletin*, 36, 25–77.

Antonio, K., Beirlant, J., Hoedemakers, T., and Verlaak, R. (2006), "Lognormal Mixed Models for Reported Claims Reserves," *North American Actuarial Journal*, 10, 30–48.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004), *Statistics of Extremes: Theory and Applications*, New York: Wiley.

Beirlant, J., Goegebeur, Y., Verlaak, R., and Vynckier, P. (1998), "Burr Regression and Portfolio Segmentation," *Insurance: Mathematics and Economics*, 23, 231–250.

Bolancé, C., Guillén, M., and Pinquet, J. (2003), "Time-Varying Credibility for Frequency Risk Models: Estimation and Tests for Autoregressive Specifications on the Random Effects," *Insurance: Mathematics and Economics*, 33, 273–282.

Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A., and Nesbitt, C. J. (1997), *Actuarial Mathematics*, Schaumburg, IL: Society of Actuaries.

Cameron, A. C., and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge, U.K.: Cambridge University Press.

De Alba, E. (2002), "Bayesian Estimation of Outstanding Claim Reserves," *North American Actuarial Journal*, 6, 1–20.

Demarta, S., and McNeil, A. J. (2005), "The *t* Copula and Related Copulas," *International Statistical Review*, 73, 111–129.

Desjardins, D., Dionne, G., and Pinquet, J. (2001), "Experience Rating Schemes for Fleets of Vehicles," *ASTIN Bulletin*, 31, 81–105.

Diggle, P. J., Heagarty, P., Liang, K.-Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data* (2nd ed.), Oxford, U.K.: Oxford University Press.

Dionne, G., and Vanasse, C. (1989), "A Generalization of Actuarial Automobile Insurance Rating Models: The Negative Binomial Distribution With a Regression Component," *ASTIN Bulletin*, 19, 199–212.

Fahrmeir, L., and Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag.

Frees, E. W. (2004), *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences*, Cambridge, U.K.: Cambridge University Press.

Frees, E. W., and Valdez, E. A. (1998), "Understanding Relationships Using Copulas," *North American Actuarial Journal*, 2, 1–25.

Frees, E. W., and Wang, P. (2005), "Credibility Using Copulas," *North American Actuarial Journal*, 9, 31–48.

Jones, A. M. (2000), "Health Econometrics," in *Handbook of Health Economics*, eds. A. J. Culyer and J. P. Newhouse, Amsterdam: Elsevier, pp. 265–344.

Klugman, S., Panjer, H., and Willmot, G. (2004), *Loss Models: From Data to Decisions* (2nd ed.), New York: Wiley.

Landsman, Z. M., and Valdez, E. A. (2003), "Tail Conditional Expectations for Elliptical Distributions," *North American Actuarial Journal*, 7, 55–71.

Lemaire, J. (1985), *Automobile Insurance: Actuarial Models*, Huebner International Series on Risk, Insurance and Economic Security, Wharton, Pennsylvania: Kluwer Academic.

Lindskog, F., and McNeil, A. J. (2003), "Common Poisson Shock Models: Applications to Insurance and Credit Risk Modelling," *ASTIN Bulletin*, 33, 209–238.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: Wiley.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

McDonald, J. B., and Butler, R. J. (1990), "Regression Models for Positive Random Variables," *Journal of Econometrics*, 43, 227–251.

McDonald, J. B., and Xu, Y. J. (1995), "A Generalization of the Beta Distribution With Applications," *Journal of Econometrics*, 66, 133–152.

Nelsen, R. (1999), *An Introduction to Copulas*, New York: Springer.

Olsen, M. K., and Shafer, J. L. (2001), "A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data," *Journal of the American Statistical Association*, 96, 730–745.

Pinquet, J. (1997), "Allowance for Cost of Claims in Bonus-Malus Systems," *ASTIN Bulletin*, 27, 33–57.

——— (1998), "Designing Optimal Bonus-Malus Systems From Different Types of Claims," *ASTIN Bulletin*, 28, 205–229.

——— (2000), "Experience Rating Through Heterogeneous Models," in *Handbook of Insurance*, ed. G. Dionne, Dordrecht: Kluwer Academic Publishers.

Pinquet, J., Guillén, M., and Bolancé, C. (2001), "Allowance for Age of Claims in Bonus-Malus Systems," *ASTIN Bulletin*, 31, 337–348.

Purcaru, O., and Denuit, M. (2003), "Dependence in Dynamic Claim Frequency Credibility Models," *ASTIN Bulletin*, 33, 23–40.

Sun, J., Frees, E. W., and Rosenberg, M. A. (2008), "Heavy-Tailed Longitudinal Data Modeling Using Copulas," *Insurance: Mathematics and Economics*, 42, 817–830. Available at *http://research3.bus.wisc.edu/course/view.php?id=129*.

Verrall, R. J. (2004), "A Bayesian Generalized Linear Model for the Bornhuetter–Ferguson Method of Claims Reserving," *North American Actuarial Journal*, 8, 67–89.