Hierarchical Insurance Claims Modeling*

Edward W. Frees
University of Wisconsin

Emiliano A. Valdez
University of New South Wales

02-February-2007†

Abstract

This work describes statistical modeling of detailed, micro-level automobile insurance records. We consider 1993-2001 data from a major insurance company in Singapore. By detailed micro-level records, we refer to 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 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 long-tailed distribution for claim amounts and also incorporate predictor variables. Year, vehicle age and a person's age turn out to be important predictors for this component. Not surprisingly, we show that there is 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 an actuary to predict automobile claims more efficiently than traditional methods. Using simulation, we demonstrate this by developing predictive distributions and calculating premiums under alternative reinsurance coverages.

*Keywords: Long-tail regression and copulas.

†The authors acknowledge the research assistance of Mitchell Wills, Shi Peng and Katrien Antonio. The first author thanks the National Science Foundation (Grant Number SES-0436274) and the Assurant Health Insurance Professorship for funding to support this research. The second author thanks the Australian Research Council through the Discovery Grant DP0345036 and the UNSW Actuarial Foundation of the Institute of Actuaries of Australia for financial support. Please do not quote without the authors' permission.

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 paper, 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 re-visit some of the classical problems faced by actuaries dealing with insurance data. This paper explores statistical models that can be constructed when detailed, micro-level records of automobile insurance policies are available.

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

- 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-tail nature of the distribution of insurance claims. This, and 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 for example, LeMaire (1985) or the guide available from the General Insurance Association (G.I.A.) of Singapore[1]. 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.

Second 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 segregate the "good drivers" from the "bad drivers" and assess the proper increment in the insurance premium for those considered "bad drivers."

[1]See the organization's website at: http://www.gia.org.sg.

This process is important to ensure equity in the premium structure available to the consumers.

In this paper, 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 General Insurance Association of Singapore, an organization consisting of most of the general insurers in Singapore. The observations are from each policyholder over a period of nine years: January 1993 until December 2001. Thus, our data comes 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 a unique code. This file provides characteristics of the policyholder and the vehicle insured, such as age and gender and type and age of vehicle insured. The claims file provides a record of each accident claim that has been filed with the insurer during the observation period and is linked to the policyholder file. The payment file consists of information on each payment that has been 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, respectively. 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 can also identify the type of claim. For our data, there are three types: (1) claims for injury to a party other than the insured, (2) claims for damages to the insured including injury, property damage, fire and theft, and (3) 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 one might encounter in a competing risks framework in biostatistics. See, for example, Bowers et al. (1997), for an actuarial introduction to two-part data (Chapter 2) and multidecrement models (Chapter 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 who might be involved in the accident. Modelling therefore the joint distribution of the simultaneous occurrence of these claim types, when an accident occurs, provides the unique feature in this paper. From a multivariate analysis standpoint, this is a

nonstandard problem in that we rarely observe all three claim types simultaneously (see Section 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 Section 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) and Nelsen (1999) for introductions to copula modeling.

To provide focus, we restrict our considerations to "non-fleet" 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 insured several vehicles, for example, coverage provided to a taxicab company, where several taxicabs are insured. See Angers et al. (2006) and Desjardins et al. (2001) for discussions of fleet policies. The unit of observation in our analysis is therefore a registered vehicle insured, broken down according to their exposure in each calendar year 1993 to 2001. In order to investigate the full multivariate nature of claims, we further restrict our consideration to policies that offer comprehensive coverage, not merely for only third party injury or property damage.

In constructing the models for our portfolio of policies, we therefore focus on the development of the claims distribution according to three different components: (1) the claims frequency, (2) the conditional claim type, and (3) the 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 paper, 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 et al. (2006) for a recent description of the claims "run-off" problem.

The organization of the rest of the paper is 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 proceed with fitting 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 one can use the modeling construction and results. We provide concluding remarks in Section 5.

2 Modeling

2.1 Data Structure

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

- N_{it} - the number of claims within a year;
- $M_{it,j}$ - the type of claim, available for each claim, $j = 1, ..., N_{it}$; and
- $C_{it,jk}$ - the claim amount, available for each claim, $j = 1, ..., N_{it}$, 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: (1) claims for injury to a party other than the insured, (2) claims for damages to the insured, including injury, property damage, fire and theft, and (3) claims for property damage to a party other than the insured. Occasionally, we shall simply refer to them as "injury," "own damage" and "third party property." It is not uncommon to have more than one type of claim 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 zero if the loss is less than the deductible). For notation purposes, we will 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 throughout the calendar year for which the vehicle had insurance coverage. The various vehicle and policyholder characteristics are described by the vector $\mathbf{x}_{it}$ and will 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 for which each subject is observed T_i times. In principle, the maximum value of T_i is 9 years because our data consists of policies from 1993 up until 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 upon which actuarial reports are

based. However, our data set is from an insurance company where there is 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 on average 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:

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

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 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; and
3. conditional severity: $\mathrm{f}(\mathbf{C}|N, \mathbf{M})$ denotes the conditional density of the claim vector $\mathbf{C}$ given N and $\mathbf{M}$.

It is customary in the actuarial literature to condition on the frequency component when analyzing the joint frequency and severity distributions. See, for example, 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). Pinquet was interested in two lines of business, claims at fault and not at fault with respect to a third party. For each line, Pinquet hypothesized a frequency and 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 via claim type. As will be seen 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 now discuss each of the three components in the following subsections.

2.2.1 Frequency Component

The frequency component, $\mathrm{f}(N)$, has been well analyzed in the actuarial literature and we will use these developments. 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 Vannasse 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 log-normally 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; they 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, for example, Diggle et al. (2002) or Frees (2004). For these models, one uses $\lambda_{it} = e_{it} \exp(\alpha_{\lambda i} + \mathbf{x}_{it}'\boldsymbol{\beta}_\lambda)$ 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 i-th subject can be expressed as

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

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

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

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 $(1 - p_{it})/p_{it} = \lambda_{it}\sigma = \exp(\alpha_{\lambda i} + \mathbf{x}_{it}'\boldsymbol{\beta}_\lambda)\sigma$.

To get a sense of the empirical observations for claim frequency, we present Table 2.1 showing 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% did not have

any claims. There are a total of 23,522 (=19,224×1 + 1,859×2 + 177×3 + 11×4 + 1×5) claims.

Table 2.1. 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	0.9	0.1	0.0	0.0	100.0

2.2.2 Claims Type Component

In Section 2.1, we described the three types of claims which 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 2.2 provides the distribution of M. For example, we see that third party injury (C_1) is the least prevalent. Moreover, Table 2.2 shows that all combinations of claims occurred in our data.

Table 2.2. Distribution of Claims, by Claim Type Observed

Value of M	1	2	3	4	5	6	7	Total
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)	
Number	102	17,216	2,899	68	18	3,176	43	23,522
Percentage	0.4	73.2	12.3	0.3	0.1	13.5	0.2	100.0

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 $V_{itj,m} = \mathbf{x}_{it}'\boldsymbol{\beta}_{M,m}$. This is known as a "selection" or "participation" equation in econometrics; see, for example, 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 ($\boldsymbol{\beta}_{M,m}$) to depend on m.

2.2.3 Severity Component

Table 2.3 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. There are 2,529 (=20,503-17,974) claims for damages to the insured ("own damages") that are censored, indicating that a formal mechanism for handling the censoring is important.

Table 2.3. Summary Statistics of Claim Losses, by Type of Claim

Statistic	Third Party Injury (C_1)	Own Damage (C_2)		Third Party Property (C_3)
		non-censored	*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 zero.

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

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

where $z = (\ln c - \mu)/\sigma$ and $B(\alpha_1,\alpha_2) = \Gamma(\alpha_1)\Gamma(\alpha_2)/\Gamma(\alpha_1+\alpha_2)$, 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 for example, Klugman, Panjer and Willmot, 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, for example, 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, there are relatively few applications that use the GB2 in a regression context. The first paper is due to McDonald and Butler (1990) who used the GB2 with regression covariates to examine the duration of welfare spells. Beirlant et al. (1998) demonstrated the usefulness of the Burr XII distribution, a special case of the GB2 with $\alpha_1 = 1$, in regression applications. Recently, Sun et al. (2006) used the GB2 in a longitudinal data context to forecast nursing home utilization. For a general approach on how to include covariates, Beirlant et al. (2004) suggested allowing all the parameters (of a Burr XII distribution) to depend on covariates. We use a simpler specification and parameterize the location parameter as $\mu_k = \mathbf{x}'\boldsymbol{\beta}_{C,k}$. Part of this is due 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 may write the joint distribution of claims (C_1, C_2, C_3) as

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

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, an important property is that the family is preserved under the marginals (see Landsman and Valdez, 2003) so that when we observe only a subset of the three types, one can still 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 are uncensored. In this case, the joint density would be

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

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

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

and the corresponding copula as

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

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 equation (3) allows us to compute the likelihood. We will 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 were available to explain and predict automobile accident frequency, type and severity. These characteristics include vehi-

cle characteristics, such as type and age, as well as person level characteristics, such as age, gender and prior driving experience. Table 3.1 summarizes these characteristics.

Table 3.1 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. This is based on the previous accident record of the policyholder. The higher the discount, the better is the prior accident record.

The Section 2 description 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. The following subsections 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 that comprise years 1993-2001, inclusive. However, when fitting models, we only used 1993-2000, inclusive. We reserved observations in year 2001 for out-of-sample validation, discussed in Section 4.

3.2 Fitting the Frequency Component Model

We begin by displaying summary statistics to suggest the effects of each Table 3.1 covariate on claim frequency. We then compare fitted models that summarize all of these effects in a single framework.

Table 3.2 displays the claims frequency distribution over time. For this company, the number of insurance policies increased significantly over 1993 to 2001. We also note that the percentage of no accidents was lower in later years. This is not uncommon in the insurance industry, where a company may decide to relax its underwriting standards in order 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 3.2. Number and Percentages of Claims, by Count and Year

	Percentage by Year										Percent
Count	1993	1994	1995	1996	1997	1998	1999	2000	2001	Number	of Total
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	0.5	0.9	0.9	0.4	0.5	0.7	1.3	0.9	1.1	1,859	0.9
3	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.1	177	0.1
4		0.0					0.0	0.0	0.0	11	0.0
5			0.0							0.0	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

Table 3.3 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 of the age of the vehicle. Here, we see low claims for new cars with initially increasing accident frequency over time. However, for vehicles in operation for long periods of time, the accident frequencies are relatively low. There are also some important interaction effects between vehicle type and age not shown here. Nonetheless, Table 3.3 clearly suggests the importance of these two variables on claim frequencies.

Table 3.3. Number and Percentages of Claims, by Vehicle Type and Age

	Percentage by Count							
	Count =0	Count =1	Count =2	Count =3	Count =4	Count =5	Number	Percent of Total
Vehicle Type								
Other	88.6	10.1	1.1	0.1	0.0	0.0	43,891	22.0
Automobile	89.5	9.5	0.9	0.1	0.0		155,461	78.0
Vehicle Age (in years)								
0	91.4	7.9	0.6	0.0	0.0		58,301	29.2
1	86.3	12.2	1.3	0.2	0.0		44,373	22.3
2	88.8	10.1	1.1	0.1			20,498	10.3
3 to 5	89.2	9.7	1.0	0.1	0.0		41,117	20.6
6 to 10	90.1	8.9	0.9	0.1		0.0	33,121	16.6
11 to 15	91.4	7.6	0.7	0.2			1,743	0.9
16 and older	89.9	8.5	1.5				199	0.1
Number by Count	178,080	19,224	1,859	177	11	1	199,352	100.0

Table 3.4 shows the effects of person level characteristics, gender, age and no claims discount, on the frequency distribution. Person level characteristics were largely unavailable for commercial use vehicles and so Table 3.4 present 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 did not contain gender and age information.

Table 3.4 suggests that driving experience was roughly similar between males and females. This company insured very few young drivers, so the young male driver category that typically has extremely high accident rates in most automobiles studies is less important for our data. Nonetheless, Table 3.4 suggests strong age effects,

with older drivers having better driver experience. Table 3.4 also demonstrates the importance of the no claims discounts (NCD). As anticipated, drivers with better previous driving records enjoy a higher NCD and have fewer accidents. Although not reported here, we also considered interactions among these three variables.

Table 3.4. 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	0.9	0.1	0.0		34,190	22.0
Male	89.5	9.5	0.9	0.1	0.0	0.0	121,271	78.0
Person Age (in years)								
21 and younger	86.9	12.4	0.7				153	0.1
22-25	85.5	12.9	1.4	0.2			3,202	2.1
26-35	88.0	10.8	1.1	0.1	0.0	0.0	44,134	28.4
36-45	90.1	9.1	0.8	0.1	0.0		63,135	40.6
46-55	90.4	8.8	0.8	0.1	0.0		34,373	22.1
56-65	90.7	8.4	0.9	0.1			9,207	5.9
66 and over	92.8	7.0	0.2	0.1			1,257	0.8
No Claims Discount (NCD)								
0	87.7	11.1	1.1	0.1	0.0		37,139	23.9
10	87.8	10.8	1.2	0.1	0.0		13,185	8.5
20	89.1	9.8	1.0	0.1			14,204	9.1
30	89.1	10.0	0.9	0.1			12,558	8.1
40	89.8	9.3	0.9	0.1	0.0		10,540	6.8
50	91.0	8.3	0.7	0.1		0.0	67,835	43.6
Number by Count	139,183	14,774	1,377	123	3	1	155,461	100.0

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

Table 3.5 compares these five fitted models, providing predictions for each level of the response variable. To summarize the overall fit, we report a Pearson chi-square 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 compared to the negative binomial model. However, recall that our data set is from an insurance company where there is substantial turnover of policies. Thus, we interpret the findings of Table 3.5 to mean that the negative binomial distribution well captures the heterogeneity in the accident frequency distribution and that the no claims discount variable captures claims

history. Thus, for this particular company, the additional complexity of the random effects portion of each model is not warranted.

Table 3.5. Comparison of Fitted Frequency Models Based on the 1993-2000 Insample Data

Count	Observed	No Covariates	Poisson	Negative Binomial	RE Poisson	RE Neg 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	0.9	1.6	178.8
ChiSquare Goodness of Fit		125.2	101.8	9.0	228.4	73,626.7

3.3 Fitting the Claim Type Model

The claim type model helps the analyst assess the type of claim, given that a claim has occurred. As described in Section 2.2.2, there are seven combinations of third party injury (C_1), own damages (C_2) and third party property (C_3) under consideration. For our data set, we found that person level characteristics, gender, age and no claims discount, did not seem to influence claim type significantly.

Vehicle characteristics and year do seem to influence claim type significantly, as suggested by Table 3.6. As with the frequency model, we find that whether or not a vehicle is an automobile is an important determinant of claim type. For automobiles, the vehicle age is important although we do not require as much detail as in the frequency model. For claims type, we consider automobiles with vehicle age greater than 2 to be "old" and otherwise "new." When examining a detailed distribution of type over time, we found a sharp break between 1996 and 1997. Table 3.6 provides the distribution of claims types by level of these variables, suggesting their importance as a determinant. For example, we see that overall 73.2% of claims fell into the own damages (C_2) category, although only 63.4% for non-autos compared to 76.3% for automobiles.

Table 3.6. Distribution of Claim Type, by Vehicle Characteristics and Year

M	Claim Type	Non-Auto (Other)	Auto	Old Vehicle	New Vehicle	Before 1997	After 1996	Overall
1	C_1	0.7	0.4	0.6	0.3	1.3	0.3	0.4
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	0.2	0.3	0.4	0.2	0.5	0.3	0.3
5	C_1, C_3	0.1	0.1	0.1	0.0	0.3	0.0	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	0.1	0.2	0.2	0.2	0.1	0.2	0.2
Counts		5,608	17,914	8,750	14,772	2,421	21,101	23,522

Table 3.7 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 3.7 shows that the the binary variable indicating whether or not the vehicle is an automobile is an important determinant whereas the addition of gender does not seem to contribute much. Using Year as a continuous variable is not as useful as the automobile variable although the binary variable Year1996 (that distinguishes between before 1997 and after 1996) is important. Similarly, the binary variable VehAge2, that distinguishes between a vehicle age less than 3 and greater than 2, is useful. We also explored interactions and other combinations of variables, not reported here. Using the three binary variables, "A," "VehAge2" and "Year1996" provides the best fit.

Table 3.7. 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

3.4 Fitting the Severity Component Model

As noted in Section 2.2.3, it is important to consider long-tail distributions when fitting models of insurance claims. Table 2.3 provided some evidence and Figure 1 reinforces this concept with an empirical histogram for each type of claim; this figure also suggests the importance of long-tail 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. However, 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 3.8 under 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 the 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 that indicates if a driver is 26-55, and Age3 indicates if a driver is 56 and over. For own damage, a censored likelihood was used. All parameter estimates were calculated via 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 via maximum likelihood. Two choices of copulas were used, the standard normal (Gaussian) copula and the t-copula. An examination of the likelihood and information statistics show 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.

We remark 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. One could also choose the degrees of freedom to maximize the likelihood but restrict it to be an integer. Because of the widespread availability of modern computational tools, we determined the degrees of freedom parameter, r, via maximum likelihood without restricting it to be an integer.

From Table 3.8, one also sees that parameter estimates are qualitatively similar under each copula. Interestingly, the correlation coefficient estimates indicate significant relationships among the three claim types. Although not presented here, it turns out that these relations were not evident when simply examining the raw statistical summaries. We also note, for the own damage and property claims, that estimators of first shape parameter, α_{21} and α_{31} respectively, are statistically significant than 1. This indicates that the additional complexity of the GB2 compared to the simpler Burr XII is warranted. For injury, there is not a statistically significant difference. This may be due to the fact that we had only 231 injury claims to assess.

Table 3.8. Fitted Copula Model

Parameter	Type of Copula		
	Independence	Normal copula	t-copula
Third Party Injury			
σ_1	1.316 (0.124)	1.320 (0.138)	1.320 (0.120)
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 (0.031)	1.301 (0.022)	1.302 (0.029)
α_{21}	5.658 (1.123)	5.507 (0.783)	5.532 (0.992)
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 (0.576)	10.106 (1.315)
$\beta_{C,2,2}$ (VehAge2)	0.090 (0.025)	0.091 (0.025)	0.091 (0.025)
$\beta_{C,2,3}$ (Year1996)	0.269 (0.035)	0.274 (0.035)	0.274 (0.035)
$\beta_{C,2,4}$ (Age2)	0.107 (0.032)	0.125 (0.032)	0.125 (0.032)
$\beta_{C,2,5}$ (Age3)	0.225 (0.064)	0.247 (0.064)	0.247 (0.064)
Third Party Property			
σ_3	0.846 (0.032)	0.853 (0.031)	0.853 (0.031)
α_{31}	0.597 (0.111)	0.544 (0.101)	0.544 (0.101)
α_{32}	1.381 (0.372)	1.534 (0.402)	1.534 (0.401)
$\beta_{C,3,1}$ (intercept)	1.332 (0.136)	1.333 (0.140)	1.333 (0.139)
$\beta_{C,3,2}$ (VehAge2)	-0.098 (0.043)	-0.091 (0.042)	-0.091 (0.042)
$\beta_{C,3,3}$ (Year1)	0.045 (0.011)	0.038 (0.011)	0.038 (0.011)
Copula			
ρ_{12}	-	0.018 (0.115)	0.018 (0.115)
ρ_{13}	-	-0.066 (0.112)	-0.066 (0.111)
ρ_{23}	-	0.259 (0.024)	0.259 (0.024)
r	-	-	193.055 (140.648)
Model Fit Statistics			
log-likelihood	-31,006.505	-30,955.351	-30,955.281
number of parms	18	21	22
AIC	62,049.010	61,952.702	61,954.562

Note: Standard errors are in parenthesis.

4 Inference

As noted in the introduction, an important application of the modeling process for the actuary involves predicting claims arising from insurance policies. We illustrate the process in two different ways: (1) prediction based on an individual observation and (2) determination of expected functions of claims over different policy scenarios.

It is common for actuaries to examine one or more "test cases" when setting premium scales or reserves. 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, for example, 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 that has a 1,998 cubic inch capacity engine. The driver enjoys the largest no claims discount (NCD) equal to 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 4.1.

Table 4.1. Prediction of Claim Type for Claim Number 215

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	0.24	0.31	11.32	0.09	100.0

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, for example, 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.

For another 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. Similarly, some insurance companies will accept the possibility of large claims and purchase insurance from another insurance company; this process is known as "reinsurance." For example, if 1,000 is such an upper limit and C represents total claims, then the reinsurance company will be responsible for the excess, $Y = \max(0, C - 1000)$. Naturally, the random variable Y has a distribution; actuaries working for the selling insurance company would like to price Y based on a smaller amount compared to actuaries working for the buying reinsurance company. The important point is that summary measures of 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 4.2 shows the results of sample calculations for our illustrative policy, observation number 1,901. However, we now allow for the possibility of multiple claims. 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 appear in Table 4.2) 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 4.2 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 limits on the excess variable Y that is the financial obligation of a reinsuring company. For the three upper limits by type, we see that the injury upper limit has the least effect due to the low probability of an injury claim, relative to the other types. The property upper limit has the most effect due to the $750 deductible for own damage on this particular test case. An overall limit has a greater effect on the excess 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 occurence. We also see that a limit of $1,000 all three types of claim produces a smaller excess variable than an overall limit of $1,000 because the event of a claim may produce more than one type of claim. An examination of Table 4.1 shows that 11.96%(=0.24+0.31+11.32+0.09) of claims are due to more than one type. Imposing both claim type limits and a larger overall limit means that the distribution of Y has a high percentage of zeros, the typical case where a reinsurer does not become involved in reimbursing claims by the insurance company.

Table 4.2. Summary Statistics of the Excess
Simulation based on 5,000 replications

Limit					25th		75th
Injury	Own Damage	Property	Overall	Mean	Percentile	Median	Percentile
0	0	0	0	542.33	91.96	293.04	652.87
1,000	0	0	0	539.93	93.67	292.28	652.51
0	1,000	0	0	481.52	38.66	225.54	597.22
0	0	1,000	0	467.15	39.37	202.62	549.39
0	0	0	1,000	414.04	9.35	131.53	474.39
1,000	1,000	1,000	0	403.93	9.06	124.45	465.24
1,000	1,000	1,000	1,000	313.99	0.42	30.79	292.89
1,000	1,000	1,000	3,000	203.70	0.00	0.26	58.22

5 Summary and Concluding Remarks

One way to think of the insurance claims data used in this paper 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; the claims amount for injury, own damage and property damage. One approach to modeling this dataset would be to use techniques from multivariate longitudinal data (see, for example, Fahrmeir and Tutz, 2001). However,

as we have pointed out, in most years policyholders do not incur a claim, resulting in many repeated zeroes (see, for example, 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 data are from an exponential family of distributions.

Other possible approaches to modeling the dataset include Bayesian predictive modeling; see de Alba (2002) and 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 the one taken in this paper in that linear combinations of Poisson process are also Poisson processes. We have chosen to re-organize 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 paper 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, this paper introduces 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 also note that our model is sufficiently flexible to allow the actuary to focus on a single type of claim. For example, this would be of interest when the actuary is designing an insurance contract and is interested in the effect of different deductibles or policy limits on "own damages" types of claims. It is also straightforward to extend this type of calculation to a block of insurance policies, such as might be priced in a reinsurance agreement.

The modeling approach developed in this paper 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 of future investigation.

A 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 to be $\mathrm{L}_1(c_1) = \mathrm{f}_1(c_1)$ and $\mathrm{L}_3(c_3) = \mathrm{f}_3(c_3)$, respectively. The subscript of the likelihood contribution L refers to the claim type. For claim type $M = 5$, there is also no own damage amount, so that the likelihood contribution is given by

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

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

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

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

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

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

For the $M = 6$ case, 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

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

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

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

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

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

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

References

[1] Angers, Jean-François, Denise Desjardins, Georges Dionne and Francois Guertin (2006). Vehicle and fleet random effects in a model of insurance rating for fleets of vehicles. *ASTIN Bulletin* 36(1): 25-77.

[2] Antonio, Katrien, Jan Beirlant, Tom Hoedemakers and Robert Verlaak (2006). Lognormal mixed modesl for reported claims reserves. *North American Actuarial Journal* 10(1): 30–48.

[3] Beirlant, Jan, Yuri Goegebeur, Johan Segers and Jozef Teugels (2004). *Statistics of Extremes: Theory and Applications* Wiley, New York.

[4] Beirlant, Jan, Yuri Goegebeur, Robert Verlaak and Petra Vynckier (1998). Burr regression and portfolio segmentation. *Insurance: Mathematics and Economics* 23, 231-250.

[5] Bolancé, Catalina, Montserratt Guillén and Jean Pinquet (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.

[6] Bowers, Newton L., Hans U. Gerber, James C. Hickman, Donald A. Jones and Cecil J. Nesbitt (1997). *Actuarial Mathematics*. Society of Actuaries, Schaumburg, IL.

[7] Cameron, A. Colin and Pravin K. Trivedi. (1998) *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

[8] De Alba, Enrique (2002). Bayesian estimation of outstanding claim reserves. *North American Actuarial Journal* 6(4): 1–20.

[9] Demarta, Stefano and Alexander J. McNeil (2005). The t copula and related copulas. *International Statistical Review* 73(1), 111-129.

[10] Desjardins, Denise, Georges Dionne and Jean Pinquet (2001). Experience rating schemes for fleets of vehicles. *ASTIN Bulletin* 31(1): 81-105.

[11] Diggle, Peter J., Patrick Heagarty, K.-Y. Liang and Scott L. Zeger, (2002). *Analysis of Longitudinal Data*. Second Edition. Oxford University Press.

[12] Dionne, Georges and C. Vanasse (1989). A generalization of actuarial automobile insurance rating models: the negative binomial distribution with a regression component. *ASTIN Bulletin* 19, 199-212.

[13] Fahrmeir, Ludwig and Gerhard Tutz. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag.

[14] Frees, Edward W. (2004). *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences*. Cambridge University Press.

[15] Frees, Edward W. and Emiliano A. Valdez (1998). Understanding relationships using copulas. *North American Actuarial Journal* 2(1), 1-25.

[16] Frees, Edward W. and Ping Wang (2005). Credibility using copulas. *North American Actuarial Journal* 9(2), 31-48.

[17] Jones, Andrew M. (2000). Health econometrics. Chapter 6 of the *Handbook of Health Economics, Volume 1*. Edited by Antonio.J. Culyer, and Joseph.P. Newhouse, Elsevier, Amersterdam. 265-344.

[18] Klugman, Stuart, Harry Panjer and Gordon Willmot (2004). *Loss Models: From Data to Decisions* (Second Edition), Wiley, New York.

[19] Landsman, Zinoviy M. and Emiliano A. Valdez (2003). Tail conditional expectations for elliptical distributions. *North American Actuarial Journal* 7(4), 55-71.

[20] Lemaire, Jean (1985) *Automobile Insurance: Actuarial Models,* Huebner International Series on Risk, Insurance and Economic Security, Wharton, Pennsylvania.

[21] Lindskog, Filip and Alexander J. McNeil (2003). Common Poisson shock models: Applications to insurance and credit risk modelling. *ASTIN Bulletin* 33(2): 209–238.

[22] Little, R.J.A., and Rubin, Donald B. (1987). *Statistical Analysis with Missing Data*. New York, NY: Wiley.

[23] McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models* (Second Edition). Chapman and Hall, London.

[24] McDonald, James B. and Richard J. Butler (1990). Regression models for positive random variables.*Journal of Econometrics* 43, 227-251.

[25] McDonald, James B. and Yexiao J. Xu (1995). A generalization of the beta distribution with applications. *Journal of Econometrics* 66, 133-152.

[26] Nelsen, Roger (1999). *An Introduction to Copulas*. Springer, New York.

[27] Olsen, Maren K. and Joseph L. Shafer (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* 96, 730-745.

[28] Pinquet, Jean (1997). Allowance for cost of claims in bonus-malus systems. *ASTIN Bulletin* 27(1): 33–57.

[29] Pinquet, Jean (1998). Designing optimal bonus-malus systems from different types of claims. *ASTIN Bulletin* 28(2): 205-229.

[30] Pinquet, Jean (2000). Experience rating through heterogeneous models. In *Handbook of Insurance*, editor by G. Dionne. Kluwer Academic Publishers.

[31] Pinquet, Jean, Montserratt Guillén and Catalina Bolancé (2001). Allowance for age of claims in bonus-malus systems. *ASTIN Bulletin* 31(2): 337-348.

[32] Purcaru, Oana and Michel Denuit (2003). Dependence in dynamic claim frequency credibility models. *ASTIN Bulletin* 33(1), 23-40.

[33] Sun, Jiafeng, Edward W. Frees and Marjorie A. Rosenberg (2007). Heavy-tailed longitudinal data modeling using copulas. University of Wisconsin working paper, available at http://research3.bus.wisc.edu/course/view.php?id=129.

[34] Verrall, Richard J. (2004). A Bayesian generalized linear model for the Bornhuetter-Ferguson method of claims reserving. *North American Actuarial Journal* 8(3): 67–89.

Author information:

Edward W. Frees
School of Business
University of Wisconsin
Madison, Wisconsin 53706 USA
e-mail: jfrees@bus.wisc.edu

Emiliano A. Valdez
School of Actuarial Studies
Faculty of Commerce & Economics
University of New South Wales
Sydney, Australia 2052
e-mail: e.valdez@unsw.edu.au

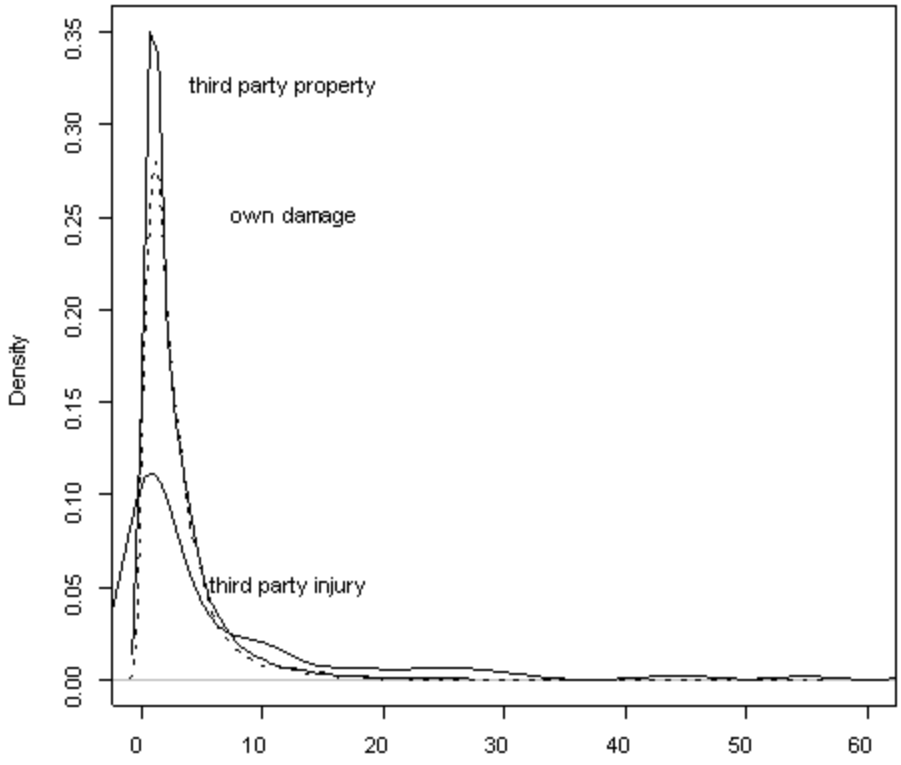

Figure 1: Density of losses by claim type

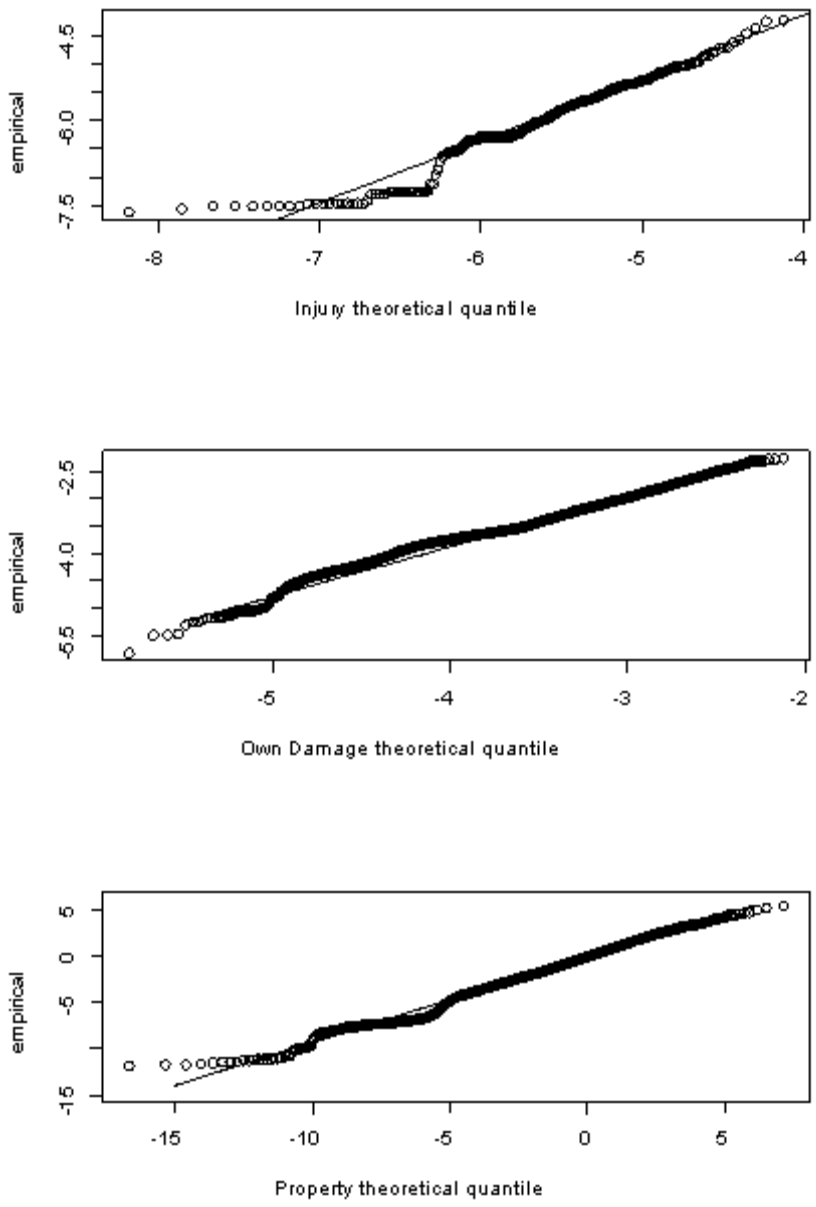

Figure 2: Quantile-quantile plots for fitting the GB2 distributions

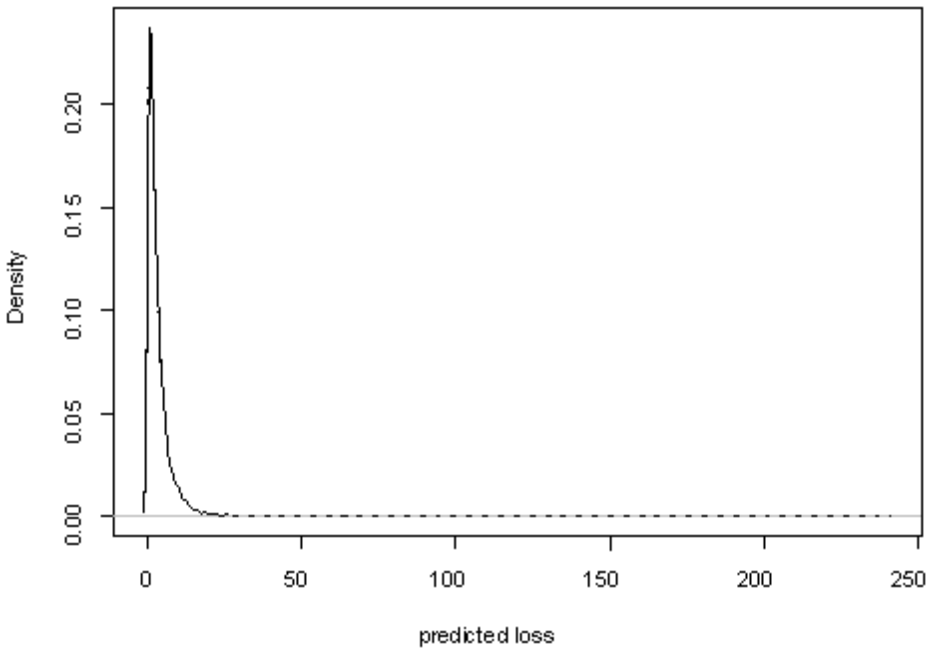

Figure 3: Simulated predictive distribution for observation 1901