Longitudinal Modeling of Singapore Motor Insurance*

Emiliano A. Valdez
University of New South Wales

Edward W. (Jed) Frees
University of Wisconsin

28-December-2005†

Abstract

This work describes longitudinal modeling of detailed, micro-level automobile insurance records. We consider 1993-2001 data from the General Insurance Association of Singapore, an organization of insurance companies. By detailed micro-level records, we refer to experience at the individual vehicle level. The data consists of vehicle characteristics, insurance coverage (including the premium) 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 statistical models for three components, corresponding to the frequency, type and severity of claims. The first is a random effects Poisson regression model for assessing claim frequency, using the policyholder file to calibrate the model. Vehicle type and no claims discount 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 damage or some combination. Premiums turn out to be important predictors for this component.

Our third model for the severity component is the most innovative. Here, we use a Burr XII long-tailed distribution for claim amounts and also incorporate predictor variables. 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 models provide 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. We demonstrate this by developing predictive distributions via simulation.

*Keywords: Insurance claims, long-tail regression, random effects models and copulas.

†The authors wish to acknowledge the research assistance of Mitchell Wills of UNSW. The first 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. The second author thanks the National Science Foundation (Grant Number SES-0436274) and the Assurant Health Insurance Professorship for provided funding to support this research.

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 longitudinal 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 outlook 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 three 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". This process is important to ensure equity in the premium distribution available to the consumers.

Third is the concern of establishing a rating system based on variables that are exogenous, or "outside," of the claims process. The actuary needs to account for the

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

behavioral aspects of consumers when designing an insurance system. For example, it would not be surprising to learn that consumers may be hiding important rating information in an attempt to acquire the most favorable insurance quotation. In a statistical modeling framework, actuaries employ variables that are useful determinants of insurance claims and that are not themselves determined by insurance claims. To illustrate, we will consider the use of gross premiums as a determinant of insurance claims. As we will see, a premium is a variable that is statistically positively related to an insurance claim (the larger the premium, the higher is the claim). One viewpoint is that a premium is simply the insurance company actuary's (possibly nonlinear) summary measure of several risk factors. If one is using the model to predict claims for solvency purposes, then premiums may be thought of as exogenous. Another viewpoint is that premiums evolve over time and that high insurance claims in one year lead to high premiums in subsequent years. From this viewpoint, premiums are only sequentially exogenous and special techniques are required. For the purposes of this paper, our examples treat all variables as exogenous. For further reading, Pinquet (2000) discusses exogeneity in an insurance rating context and Frees (2004, Chapter 6) provides an overview of exogeneity in longitudinal models.

In this paper, we consider policy exposure and claims experience data derived from vehicle insurance portfolios of general insurance companies in Singapore. The primary source of this data is 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 the level of gross premium, driver information, and brand or make of vehicle insured. There are well over 5 million records in the policy file, with each record corresponding to a vehicle. 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. There are over 700,000 recorded claims in the claims file, whereas the payment file has more than 4 million recorded payments.

To provide focus, we restrict our considerations to "fleet" policies from one insurance company. These are policies issued to customers whose insurance covers more than a single vehicle. A typical situation of "fleet" policies is motor insurance coverage provided to a taxicab company, where several taxicabs are insured. The unit of observation in our analysis is therefore a registered vehicle insured under a fleet policy. We further break down these registered vehicles according to their exposure

in each calendar year 1993 to 2001.

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. Further, 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. 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). Further, 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.

In constructing the longitudinal 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 modelling 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.

The organization of the rest of the paper 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, focusing on its usefulness for prediction purposes. We provide concluding remarks in Section 5.

2 Modeling

2.1 Data

As explained in the introduction, the data available are disaggregated by risk class i, denoting 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 loss 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, claims amounts are available. However, for claims for damages to the insured ("own damages"), only a loss amount is available. Here, we follow standard actuarial terminology and define the loss amount, $C_{it,2k}$, to be equal to the excess of a claim over a known deductible, d_{it} (and equal to zero if the claim is less than the deductible). For notation purposes, we will sometimes use $C^*_{it,2k}$ to denote the claim amount; this quantity is not known when it falls below the deductible. Thus, it is possible to have observed a zero loss 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 $\mathbf{C}_{it}$. Finally, the observable data available consist of the following information

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

In summary, there are $n = 9,409$ subjects for which each subject is observed T_i times. The maximum value of T_i is 9 years because our data consists only 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.

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. 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. Here, 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 use the standard random effects Poisson count model. See, for example, Diggle et al. (2002) or Frees (2004). For this model, one uses $\exp(\alpha_{\lambda i} + \mathbf{x}_{it}'\boldsymbol{\beta}_\lambda)$ to be the Poisson parameter for the $\{it\}$ observational unit, where $\alpha_{\lambda i}$ is a time-constant latent random variable to account for the heterogeneity. We also allow for the fact that an observational unit may be exposed only partially during the year. If we denote by e_{it} the length of exposure, then we adjust the Poisson mean parameter to be $\lambda_{it} = e_{it}\exp(\alpha_{\lambda i} + \mathbf{x}_{it}'\boldsymbol{\beta}_\lambda)$. 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 $\Pr(N = n|\lambda) = \lambda^n e^{-\lambda}/n!$.

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 25,440 observations of which 93.2% did not have any claims. There are a total of 1,872 (=1,600×1 + 121×2 + 10×3) claims, from 1,731 (=1,600+121+10) subject-year $\{it\}$ observations.

Table 2.1. Frequency of Claims

Count	0	1	2	3	Total
Number	23,709	1,600	121	10	25,440
Percentage	93.2	6.3	0.5	0.1	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: "injury", "third party property", and "own damages". 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. Here, 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	22	959	681	3	10	196	1	1,872
Percentage	1.2	51.2	36.4	0.2	0.5	10.5	0.1	100.0

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

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

where $V_{itj,r} = \mathbf{x}_{it}' \boldsymbol{\beta}_{M,r}$. 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 r although we allow parameters ($\boldsymbol{\beta}_{M,r}$) to depend on r.

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, suggesting the long-tail nature of the data. Third party injury claims, although the least frequent, have the strongest potential for large consequences. There are 97 losses 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	36	1,062	1,159	888
Mean	7,303	2,631	2,410	2,803
Standard Deviation	12,297	3,338	3,277	3,415
Median	1,716	1,459	1,267	1,768
Minimum	14	3	0	5
Maximum	51,000	32,490	32,490	35,544

Note: Censored "own damages" losses have values of zero.

To accommodate the long-tail nature of claims, we use the Burr XII marginal distribution for each claim type. This has distribution function

$$F_C(c) = 1 - \left(\frac{\gamma}{\gamma + c^\tau}\right)^\eta, c \geq 0, \tag{2}$$

where τ is a shape parameter and γ is a scale parameter. This distribution is well known in actuarial modeling of univariate loss distributions (see for example, Klugman, Panjer and Willmot, 2004). Not only is the Burr XII useful in handling long-tail distributions, but from equation (2) we see that the inverse distribution function is readily computable; this feature will turn out to be computationally useful in our residual analysis that follows.

We will use this distribution but will allow parameters to vary by type and thus consider γ_k and η_k for $k = 1, 2, 3$. Further, recently Beirlant et al. (1998) have demonstrated the usefulness of the Burr XII distribution in regression applications by allowing covariates to appear through the shape parameter τ. For a more general approach, Beirlant et al. (2004) suggested allowing all the parameters to depend on covariates. We use the simpler specification and, following the work of Beirlant et al. (1998), parameterize the shape parameter as $\tau_{it,k} = \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta}_{C,k}\right)$. With this notation, we define the distribution function for the $\{it\}$ observational unit and the kth type of claim as

$$\mathrm{F}_{it,k}\left(c\right) = 1 - \left(\frac{\gamma_k}{\gamma_k + c^{\tau}}\right)^{\eta_k}, \text{where} \quad \tau = \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta}_{C,k}\right).$$

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

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

Here, the marginal distribution of C_j is given by $\mathrm{F}_j(\cdot)$ and $\mathrm{H}\left(\cdot\right)$ is the copula. 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

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

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

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

and the corresponding copula as

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

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 of the vehicles were available to explain and predict automobile accident frequency, type and severity. Because only fleets policies are being considered, driver characteristics (e.g. age, gender) do not appear in our analysis. Table 3.1 summarizes these characteristics.

Table 3.1 Description of Covariates

Covariate	Description
Year	The calendar year. This varies from 1993-2001.
Premium	The level of gross premium for the policy in the calendar year.
Cover Type	The type of coverage on the insurance policy. It is either comprehensive (C), third party fire and theft (F) or third party (T).
NCD	No Claims Discount. This is a categorical variable based on the previous accident record of a vehicle. The categories are 0%, 10%, 20%, 30%, 40% and 50%. The higher the discount, the better is the prior accident record.
Vehicle Type	The type of vehicle being insured, either automobile (A) or other (O).

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 1993-1999, inclusive. We reserved observations in years 2000 and 2001 for out-of-sample validation, the topic of Section 4.

3.2 Fitting the Frequency Component Model

We begin by displaying summary statistics to suggest the effects of each Table 3.1 covariates on claim frequency. We then show a fitted model that summarizes all of these effects in a single model.

Table 3.2 displays the claims frequency distribution over time, showing no strong trends.

Table 3.2. Number and Percentages of Claims, by Count and Year

Count	1993	1994	1995	1996	1997	1998	1999	2000	2001	Total
0	4,015	3,197	1,877	2,291	2,517	2,711	2,651	2,624	1,826	23,709
	93.4	94.0	94.1	94.5	92.3	93.7	91.7	93.1	91.7	93.2
1	267	195	108	127	196	168	216	175	148	1,600
	6.2	5.7	5.4	5.2	7.2	5.8	7.5	6.2	7.5	6.3
2	17	10	9	5	12	15	21	17	15	121
	0.4	0.3	0.5	0.2	0.4	0.5	0.7	0.6	0.8	0.5
3	0	0	0	0	1	0	4	3	2	10
	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.1	0.1
Total	4,299	3,402	1,994	2,423	2,726	2,894	2,892	2,819	1,991	25,440

Table 3.3 shows the distribution of premiums, given in thousands of Singaporean dollars (not adjusted for inflation), by claims count. Not surprisingly, we see that policies with zero claims have lower premiums, suggesting a positive relation between premiums and claims count. Insurance companies compute premiums based on policy type, vehicle type, driving history and so forth. By restricting considerations to a single company, we do not introduce the heterogeneity of different expense loadings by different companies; this suggests using a single parameter to incorporate premiums.

Table 3.3. Summary Statistics of Premiums, by Claims Count

Count	Number	Mean	Median	Standard Deviation	Minimum	Maximum
0	23,709	0.623	0.388	0.663	0.001	11.634
1	1,600	0.830	0.626	0.703	0.001	6.849
2	121	0.888	0.723	0.643	0.004	3.554
3	10	0.674	0.587	0.459	0.006	1.566

Tables 3.4-3.6 show the effects of coverage type, no claims discount (NCD) and vehicle type on the frequency distribution. The comprehensive coverage type is the most widely used (57.9%) and also is the most likely to incur a claim (9.2%). Most policies did not have a no claims discount (85.1%). Somewhat surprisingly, those with a positive NCD did not seem to have a lower accident frequency. Finally, automobiles have slightly higher accident rates than the other category (7.5% versus 5.9%).

Table 3.4. Number and Percentages of Claims, by Cover Type

Cover Type	Count				
	0	1	2	3	Total
Comprehensive (C)	13,375	1,238	107	10	14,730
	90.8	8.4	0.7	0.1	
Third Party	627	16	1	0	644
Fire&Theft (F)	97.4	2.5	0.1	0.0	
Third Party (T)	9,707	346	13	0	10,066
	96.4	3.4	0.1	0.0	
Total	23,709	1,600	121	10	25,440

Table 3.5. Number and Percentages of Claims, by No Claims Discount

NCD	Count				
	0	1	2	3	Total
0%	20,205	1,326	105	10	21,646
	93.3	6.1	0.5	0.1	
10%	683	65	2	0	750
	91.0	8.9	0.3	0.0	
20%	1,365	77	7	0	1,449
	94.2	5.3	0.5	0.0	
30%	361	34	5	0	400
	90.3	8.5	1.3	0.0	
40%	234	29	0	0	263
	89.0	11.0	0.0	0.0	
50%	861	69	2	0	932
	92.4	7.4	0.2	0.0	
Total	23,709	1,600	121	10	25,440

Table 3.6. Number and Percentages of Claims, by Vehicle Type

Vehicle Type	Count				
	0	1	2	3	Total
Auto	13,204	993	76	7	14,280
	92.5	7.0	0.5	0.1	
Others	10,505	607	45	3	11,160
	94.1	5.5	0.4	0.1	
Total	23,709	1,600	121	10	25,440

After additional examination of the data, the Section 2.2.1 random effects Poisson model was fit. As part of the examination process, we investigated interaction terms

among the covariates and nonlinear fits with regard to year and premiums. The final fitted model, summarized in Table 3.7, does not include a time effect (year). Premium enters the systematic component linearly, and we used binary variables for comprehensive coverage (Cover=Comp), automobile vehicle type (Vtype=Auto) and zero no claims discount (NCD=0). The random effects component was treated as normally distributed with mean 0 and variance σ_λ^2.

Table 3.7 shows that comprehensive coverage, automobile vehicle type and zero no claims discount are each associated with a higher tendency for accidents. Further, higher premiums are associated with more accidents, except for the case of automobile vehicle type.

Table 3.7. Fitted Claims Frequency Model

Variable	Regression coefficient ($\boldsymbol{\beta}_\lambda$) estimate	Standard error	t-statistic	p-value
intercept	-4.215	0.152	-27.81	<0.001
Premium	0.329	0.231	1.42	0.155
Cover=Comp	1.600	0.128	12.49	<0.001
VType=Auto	0.217	0.117	1.86	0.064
NCD=0	0.062	0.094	0.66	0.510
(Cover=Comp)*Premium	-0.321	0.231	-1.39	0.165
(VType=Auto)*Premium	-0.511	0.133	-3.96	<0.001
Log σ_λ^2	0.633	0.030	21.16	<0.001

3.3 Fitting the Claim Type Model

Table 3.8 shows the relation between claim type and premiums. Here, we see that third party damages to property (C_3) are associated with small premiums and an insured's own damages (C_2) are associated with large premiums. Not surprisingly, larger premiums are associated with accidents having more than one claim type.

Table 3.8. Summary Statistics of Premiums, by Claims Type

M	Type	Number	Mean	Median	Standard Deviation	Minimum	Maximum
1	C_1	22	0.665	0.367	0.742	0.188	1.314
2	C_2	959	0.986	0.784	0.735	0.004	6.849
3	C_3	681	0.592	0.393	0.554	0.001	3.132
4	C_1, C_2	3	0.869	0.632	0.441	0.597	3.661
5	C_1, C_3	10	0.713	0.629	0.605	0.091	1.720
6	C_2, C_3	196	0.962	0.804	0.674	0.004	4.724
7	C_1, C_2, C_3	1	0.775	0.775	0	0.775	0.775

Table 3.9 reports the results from a fitted multinomial logit model using premiums (in thousands of Singaporean dollars) as the explanatory variable. Here, C_2 is the

omitted category. Using equation (1), these parameter estimates provide predictions of claim type. The usual interpretations are also available. To illustrate, comparing two policies that differ by 10 Singaporean dollars, we interpret the slope for type $M = 1$ to mean that the policy is 1.05% ($=e^{1.047/100} - 1$) times more likely to have an injury (C_1) compared to an insureds own damage claim (C_2).

Table 3.9. Fitted Multinomial Logit Model

		Intercept			Slope		
			Standard			Standard	
M	Type	Estimate	Error	p-value	Estimate	Error	p-value
1	C_1	-2.031	0.135	<.001	1.047	0.130	<.001
3	C_3	-6.950	1.621	<.001	0.630	1.650	0.703
4	C_1, C_2	-3.612	0.347	<.001	0.287	0.410	0.485
5	C_1, C_3	-4.510	0.514	<.001	0.446	0.567	0.432
6	C_2, C_3	-0.486	0.086	<.001	1.090	0.098	<.001
7	C_1, C_2, C_3	-6.040	0.946	<.001	0.860	0.863	0.319

Notes: Response is claim type, explanatory variable is premium (in thousands). Omitted category is C_2.

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 Burr XII distribution as a model for losses. Figure 2 provides QQ plots, described in Beirlant et al. (1998). Here, we see that this distribution fits the data well, even without the use of covariates. 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 Burr 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 3.10 under the "Independence" column. As noted in Section 2.2.3, all three parameters of the Burr distribution varied by claim type. In the interest of parsimony, no covariates were used for the 30 injury claims, whereas an intercept, Year and Premium were used for the Third Party Property and automobile coverage type was used for Own Damage. 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 and the t-copula was an 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 t-copula indeed provides a better fit to the data than the Normal copula.

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 3.10, 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.

4 Prediction

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 prediction process in two different ways: (1) prediction based on an individual observation and (2) out-of-sample validation for a portfolio of claims. For both types of prediction problems, 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, $N = 1$.

It is common for actuaries to examine one or more "test cases" when setting premium scales or reserves. To illustrate what an actuary can learn when predicting based on an individual observation, we chose an observation from our out-of-sample period consisting of years 2000 and 2001. Claim number 215 from our database involves a policy for a 1996 BMW with a premium of \$215.83 for two months during 2001, and with a \$750 deductible for comprehensive coverage.

Table 3.10. Fitted Copula Model

Parameter	Type of Copula		
	Independence	Normal copula	t-copula
		Third Party Injury	
η_1	1,488 (13,313)	1,501 (6,041)	1,499 (6,770)
γ_1	2,557 (22,888)	2,601 (10,647)	2,636 (11,892)
τ_1	0.474 (0.0693)	0.483 (0.067)	0.486 (0.068)
		Own Damage	
η_2	4.174 (1.384)	4.313 (1.443)	4.073 (1.310)
γ_2	6.953 (2.368)	7.038 (2.418)	6.643 (2.199)
$\beta_{C,2,1}$ (intercept)	-0.066 (0.087)	-0.700 (0.086)	-0.065 (0.086)
$\beta_{C,2,2}$ (Year)	0.027 (0.010)	0.027 (0.010)	0.028 (0.010)
$\beta_{C,2,3}$ (Premium)	-0.626 (0.330)	-0.678 (0.326)	-0.653 (0.327)
		Third Party Property	
η_3	4.059 (0.959)	3.713 (0.787)	3.568 (0.741)
γ_3	9.107 (2.461)	7.638 (1.898)	7.295 (1.787)
$\beta_{C,3,1}$(Cover=Comp)	0.111 (0.046)	0.135 (0.046)	0.151 (0.047)
		Copula	
ρ_{12}	-	-0.653 (0.263)	-0.620 (0.288)
ρ_{13}	-	0.276 (0.335)	0.223 (0.341)
ρ_{23}	-	0.315 (0.054)	0.330 (0.060)
r	-	-	11.805 (9.041)
Model Fit Statistics			
log-likelihood	-3,158	-3,142.6	-3,141.6
number of parms	11	14	15
AIC	6,338	6,313.3	6,253.2

Note: Standard errors are in parenthesis.

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	1.24	48.96	38.95	0.16	0.56	10.09	0.05	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 or excess, 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 \$10,438.68, corresponding to the 98th percentile of the predictive distribution.

For the out-of-sample validation procedure, we consider 375 claims that were observed during 2000 and 2001. For each claim, we generated 5,000 simulated values of total claims as described above. Figure 4 summarizes the relationship between the simulated predicted and the actual held-out values. Depending on the purposes of the prediction, the actuary can select a predicted value from the predictive distribution generated for each of the 375 claims. In the figure, we show the resulting comparison with the actual held-out values using either the average, the median, the 90th percentile or the 95th percentile of the predictive distributions. For each of the 375 claims, the actual values are on the vertical axis and the predicted value from the 5,000 simulations are on the horizontal axis. Not surprisingly, these graphs show the large variability in actual losses compared to the point prediction. They also demonstrate that choosing a high percentile for a predictive value leads to higher probability of covering most claims.

With the entire predictive distribution, the actuary need not restrict him or herself to using the mean. Table 4.2 summarizes results using alternative summary measures of the simulated distribution, including not only the mean of the 5,000 simulations but also the median, the 90th and the 95th percentiles. These alternative measures can be used for different business purposes. For example, one might use the 95th percentile to set reserves for statutory purposes, where the concern is to set aside a sufficient amount to safeguard against insolvency. Alternatively, one might use the median to set reserves for generally accepted accounting purposes (GAAP in the USA). Here, the actuary uses a "best" estimate of future liabilities when setting the reserve to balance competing interests of shareholders (who would like company liabilities to be low) and policyholders (who would like the company to retain a larger reserve in order to be in a better position to pay future losses).

Table 4.2. Comparison of Actual Loss Distribution to Summary Measures of the Simulated Predictive Distribution

	Mean	Median	Standard Deviation	Minimum	Maximum
Actual Loss Distribution	3,047.5	1,773.9	3,407.5	5.2	19,528.9
Predictive Distribution Summary Measure					
Mean	2,255.3	2,320.4	307.2	715.2	2,864.0
Median	337.1	357.3	211.6	0.0	1,045.1
90th Percentile	6,240.7	6,341.9	790.1	1,867.8	7,727.7
95th Percentile	9,421.2	9,395.3	1,060.5	3,216.8	11,862.6

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.

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.

Further, our main contribution in this paper is the introduction of a multivariate claims distribution for handling long-tailed, related claims using covariates. As in the work of Beirlant et al. (1998), we used the Burr XII 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.

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 data 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 claim and $c_2 = \max(0, c_2^* - d)$ be the observed loss. 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

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

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 $f_{uc,123}$ is given in equation (3) and

$$H_{c,123}(c_1, d, c_3) = \int_0^d h_3(F_1(c_1), F_2(z), F_3(c_3)) f_1(c_1) f_3(c_3) f_2(z) dz.$$

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

References

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

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

[3] 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.

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

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

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

[7] 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.

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

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

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

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

[12] 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.

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

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

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

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

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

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

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

[20] 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.

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

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

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

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

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

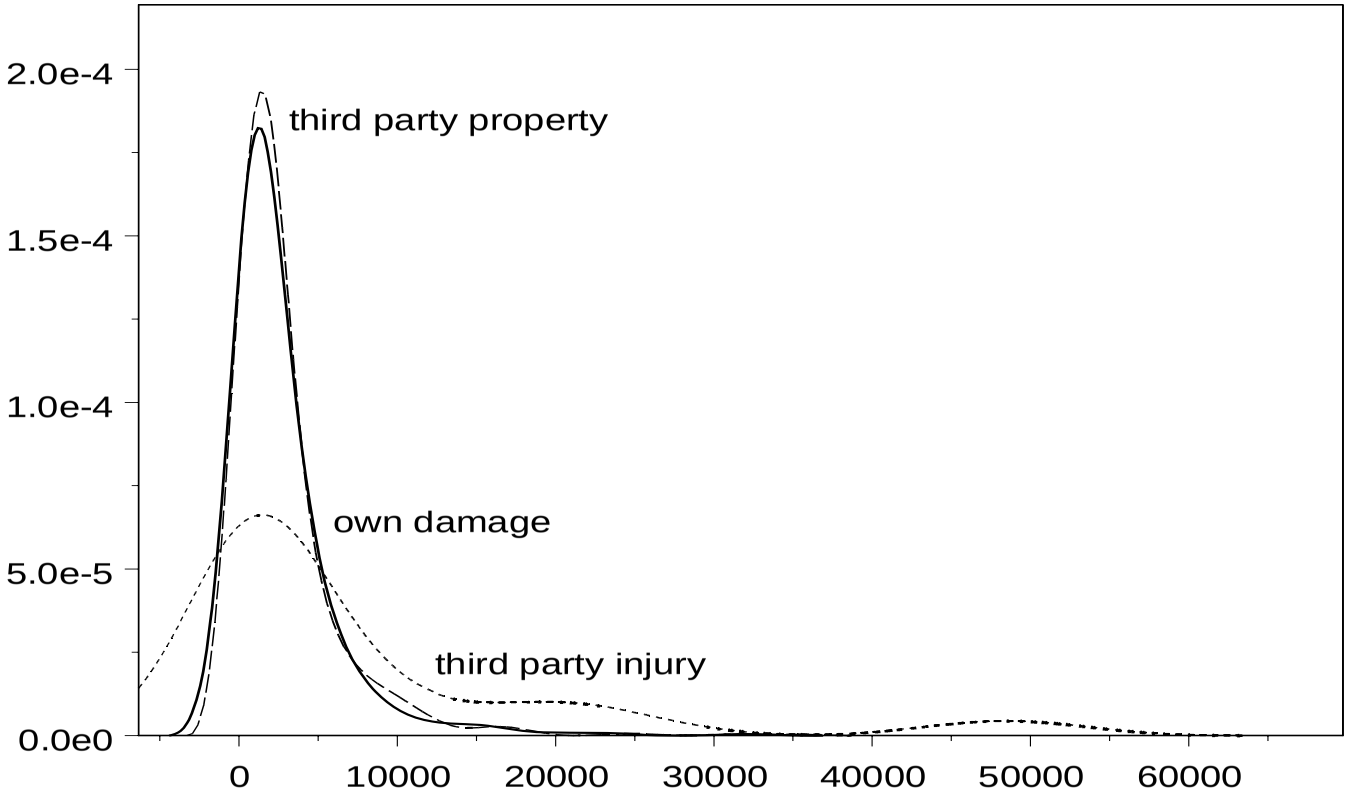

Figure 1: Density of losses by claim type

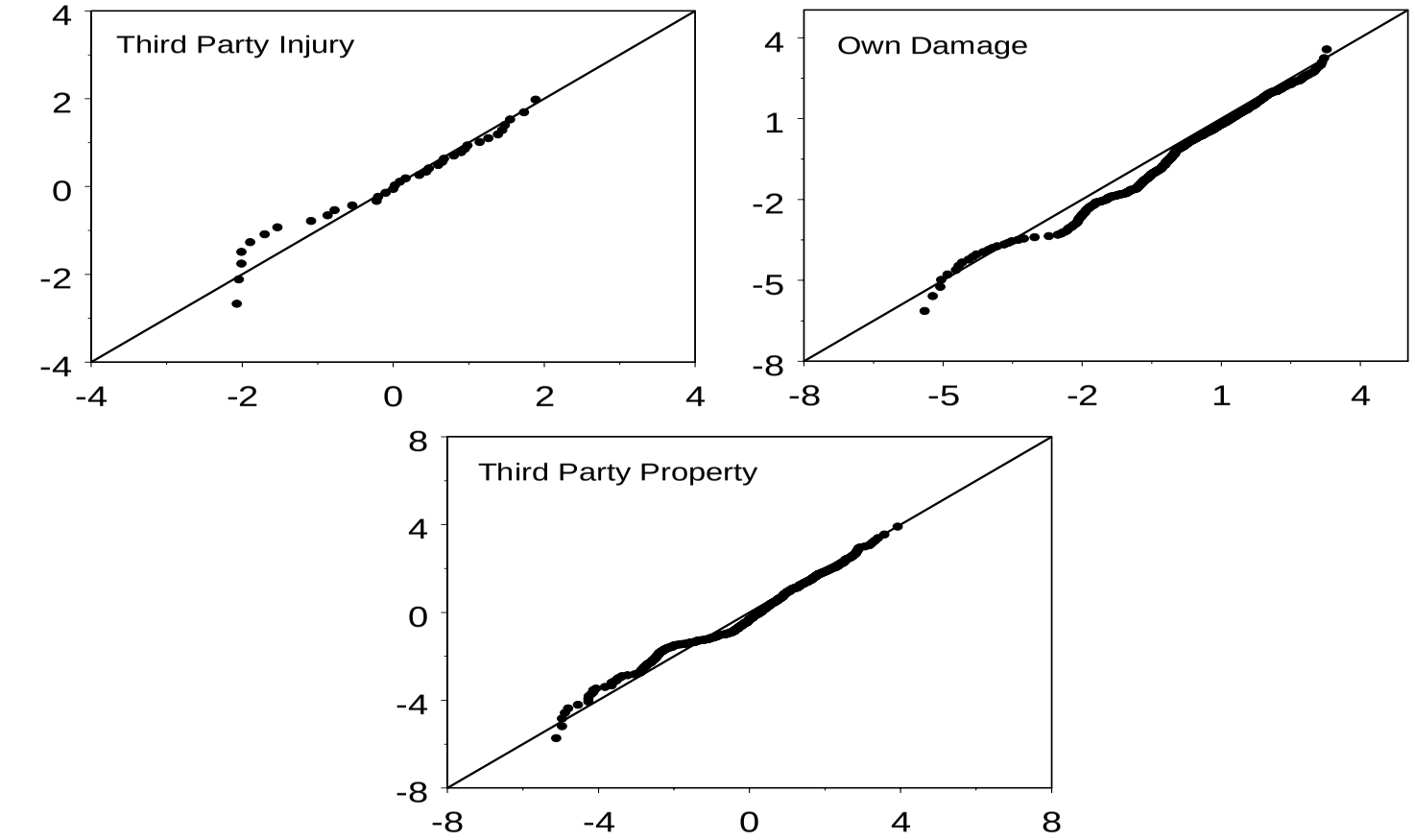

Figure 2: Quantile-quantile plots for fitting the Burr XII distributions

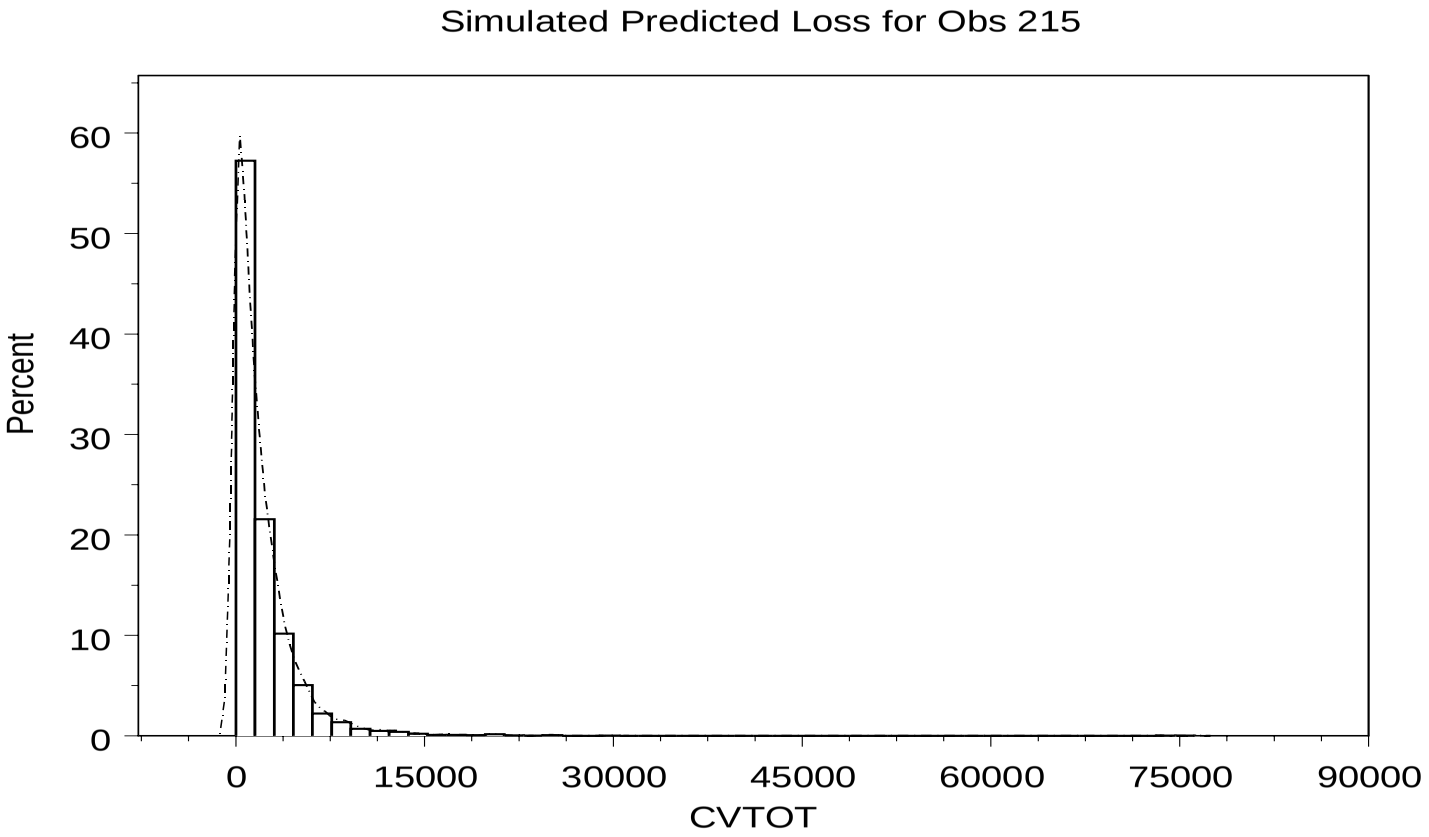

Figure 3: Simulated predictive distribution for observation 215

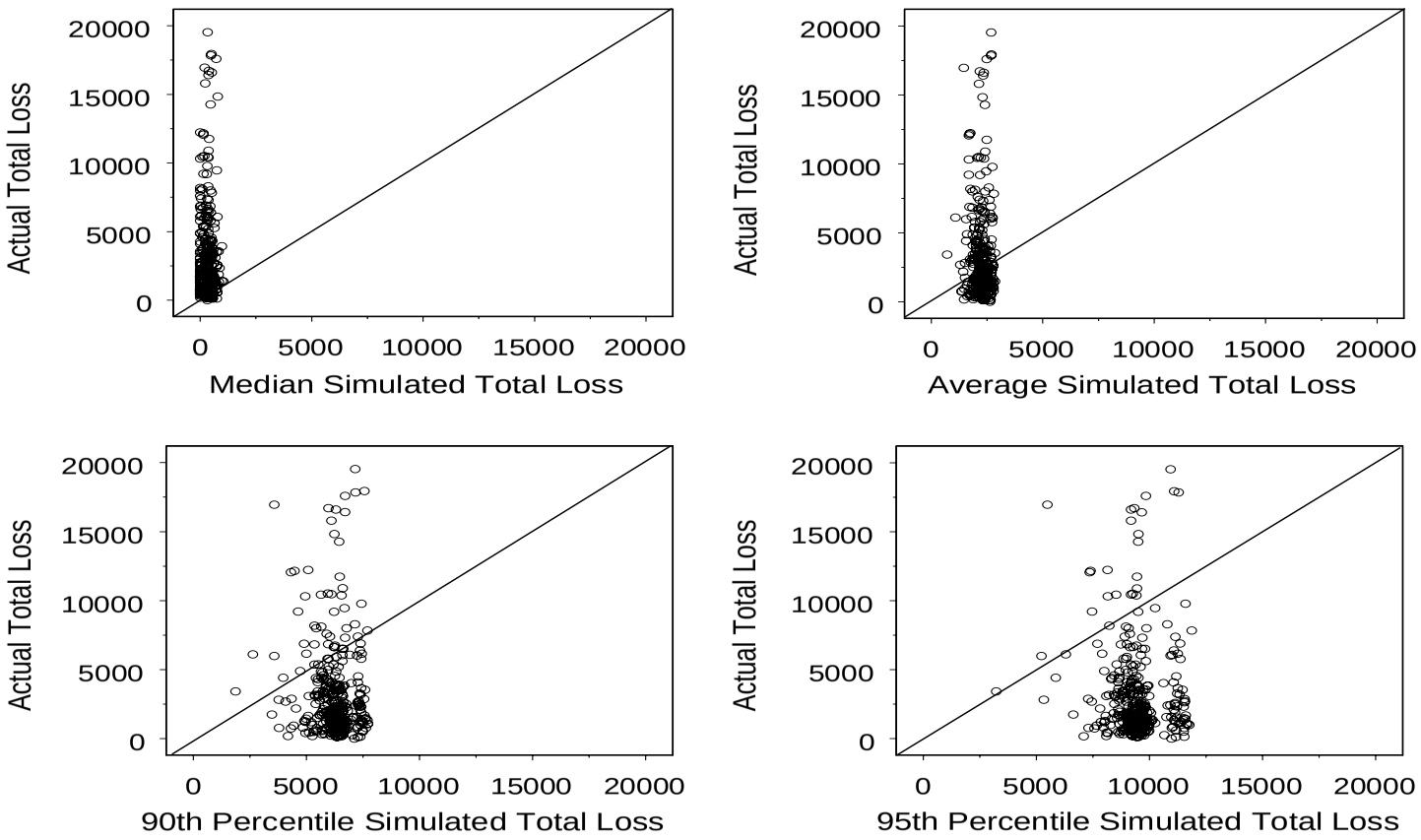

Figure 4: Comparison of the actual to the predicted held-out losses

Author information:

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

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