Hierarchical Insurance Claims Modeling

Edward W. (Jed) Frees, University of Wisconsin - Madison
Emiliano A. Valdez, University of Connecticut

2009 Joint Statistical Meetings
Session 587 - Thu 8/6/09 - 10:30 AM to 12:20 PM
1-6 August 2009

Outline of presentation

- Motivation
- Data
- Three component (hierarchical) models:
  - claim frequency
  - type of claim
  - claim severity
- Summary and concluding remarks

Motivation for hierarchical model

- We are used to thinking that in predicting/estimating insurance claims distributions:

$$\text{Cost of Claims} = \text{Frequency} \times \text{Severity}$$

- Improvements can be made:
  - prediction on frequency: introducing heterogeneity
  - prediction on severity: using additional information such as types of claims
- It is in the second component that we are interested to further explore.

Motivation driven by data

- We have a portfolio of automobile insurance policies from Singapore.
  - detailed information on policies of registered cars, claims and payments settled.
  - period: 1 January 1993 until 31 December 2001 (nine years in total)
- Data provided by the General Insurance Association (GIA) of Singapore:
  - GIA has (29) member companies to promote their common interest and that of the industry (e.g. educating media, public awareness, interest to government)
  - check website: http://www.gia.org.sg.
  - may have similar function to the Insurance Services Office (ISO), although seems to be more of a depository of data.

Risk factor rating system

- Insurers adopt "risk factor rating system" in establishing premiums for motor insurance
- Some risk factors considered:
  - vehicle characteristics: make/brand/model, engine capacity, year of make (or age of vehicle), price/value
  - driver characteristics: age, sex, occupation, driving experience, claim history
  - other characteristics: what to be used for (private, corporate, commercial, hire), type of coverage
- The "no claims discount" (NCD) system:
  - rewards for safe driving
  - discount upon renewal of policy ranging from 0 to 50%, depending on the number of years of zero claims.

Data characteristics

- Individual records of 1,090,942 registered cars with policy and claims information over nine (9) years [1993 to 2001], from 46 companies.
- Policy file has 26 variables with 5,667,777 records; claims file has 12 variables with 786,678 records; payment file has 8 variables with 4,427,605 records.
- Gross premiums: 1999 = 3.7 bn; 2000 = 4.3 bn; 2001 = 4.7bn.
- In each year, about 5 to 10% are recorded fleets.
- To provide focus for our investigation, we selected non-fleet policies from just a single insurer.
  - The non-fleet policies provided for a more interesting model fits.
  - For this insurer, about 90% are non-fleet policies.

Data extraction

- The data available are disaggregated by risk class i (vehicle) and over time t (year). For each observational unit $\{it\}$, the responses are:
  - number of claims within a year: N_{it}
  - type of claim, available for each claim: $M_{it,j}$ for $j = 1, ..., N_{it}$
  - the loss amount, for each claim: $C_{it,jk}$ for $j = 1, ..., N_{it}$ and for type $k = 1, 2, 3$
  - exposure: e_{it}
  - vehicle characteristics: described by the vector $\mathbf{x}_{it}$
  - excess or deductible: d_{it}
- The data available therefore consist of

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

- That is, there are n subjects and each subject is observed T_i times. For our data, we have $n = 96,014$ and T_i has a maximum of 9 years. Total observation is 199,352 so that on average, we observe each vehicle for only 2.08 per vehicle.

Possible covariates

- The calendar year - 1993-2001; treated as continuous variable.
- The level of gross premium for the policy in the calendar year - continuous.
- The type of vehicle:
  - bus (B), car (C), or motor cycle (M)
- Cover type: comprehensive (C), third party fire and theft (F), and third party (T).
- The NCD applicable for the calendar year - 0%, 10%, 20%, 30%, 40%, and 50%.
- Some driver characteristics such as age and gender.

Claim types

- There were three (3) possible types of claims:
  1. claims for injury to a party other than the insured - I
  2. claims for property damage to a party other than the insured - P; and
  3. claims for damages to the insured, including injury, property damage, fire and theft. - O
- For each accident, it is not uncommon to have more than one type of claim incurred.
- For the first two types, claim amounts are available, but for "own damages" claims, only the loss amount is available (some censoring).
- Thus, it is possible to have a zero loss associated with an "own damage" claim. We assume that these deductibles apply on a per accident basis.

Decomposition of the joint distribution

- We can write the joint distribution of the observables as

$$\mathsf{f}(N, \mathbf{M}, \mathbf{C}) = \mathsf{f}(N) \times \mathsf{f}(\mathbf{M}|N) \times \mathsf{f}(\mathbf{C}|N, \mathbf{M}) .$$

- This leads us to a decomposition of the joint distribution into the following components:
  1. the frequency component $\mathsf{f}(N)$ - accounts for the number of claims made in the calendar year;
  2. the conditional claim type component $\mathsf{f}(\mathbf{M}|N)$ - accounts for the type of claim given the number of claims; and
  3. the conditional severity component $\mathsf{f}(\mathbf{C}|N, \mathbf{M})$ - accounts for the amount of loss incurred, conditional on claim count and claim types.
- Such natural decomposition allowed us to investigate each component separately.

The frequency component

- The frequency component, $\mathrm{f}(N)$, has been well analyzed in the actuarial literature and we use these developments:
  - Dionne and Vanasse (1989)
  - Pinquet (1997, 1998)
  - Pinquet, Guillén and Bolancé (2001) and Bolancé, Guillén and Pinquet (2003)
  - Purcaru and Denuit (2003)
- Standard random effects count models:
  - Poisson and Negative Binomial models
  - Diggle et al. (2002); or
  - Frees (2004)

Observed frequency of 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

Random effects count model

- Let $\lambda_{it} = e_{it} \exp\left(\alpha_{\lambda i} + \mathbf{x}_{it}' \beta_\lambda\right)$ be the conditional mean parameter for the $\{it\}$ observational unit, where $\alpha_{\lambda i}$ is a time-constant latent random variable for heterogeneity.
- With $\lambda_i = (\lambda_{i1}, ..., \lambda_{iT_i})'$, the frequency component likelihood for the i -th subject is $L_i = \int \Pr\left(N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i} | \lambda_i\right) \mathrm{f}\left(\alpha_{\lambda i}\right) d\alpha_{\lambda i}$.
- Typically one uses a normal distribution for $\mathrm{f}\left(\alpha_{\lambda i}\right)$.
- The conditional joint distribution for all observations from the i-th subject is

$$\Pr\left(N_{i1} = n_{i1}, ..., N_{iT_i} = n_{iT_i} | \lambda_i\right) = \prod_{t=1}^{T_i} \Pr\left(N_{it} = n_{it} | \lambda_{it}\right).$$

Random effects Poisson and N.B. count model

- Poisson distribution model:
  - $\Pr(N = n|\lambda) = \lambda^n e^{-\lambda}/n!$ using $\lambda = \lambda_{it}$ for the mean parameter.
- Negative binomial distribution model with parameters p and r:
  - $\Pr(N = n|r, p) = \binom{n + r - 1}{r - 1} p^r (1 - p)^n.$
  - Here, $\sigma = r^{-1}$ is the dispersion parameter and
  - $p = p_{it}$ is related to the mean through

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

The effect of calendar year

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

	Percentage by Year									Total	Total
Count	1993	1994	1995	1996	1997	1998	1999	2000	2001	Number	Percent
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

The effect of vehicle type and vehicle age

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	Total Number	Total Percent
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

The effect of gender, age and NCD discounts

Table 3.4. Number and Percentages of Claims, by Gender, Age and NCD

	Percentage by Count							
	Count =0	Count =1	Count =2	Count =3	Count =4	Count =5	Total Number	Total Percent
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

Comparison of the fitted frequency models

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

The conditional claim type component

- We recorded combinations of claim types (denoting by M the r.v. describing the combination observed) as

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

- Certain characteristics help to describe the types of claims that arise and to explain this feature, we use the multinomial logit of the form

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

where $V_{itj,r}=\mathbf{x}_{it}'\beta_{M,r}$.

- Known as a "selection" or "participation" equation in econometrics (see for example Jones, 2000).

The effect of vehicle characteristics and calendar year

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

Comparison of fit of alternative claim type models

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

The conditional severity component

- For each given accident, we are able to observe a triplet of loss variables (C_1, C_2, C_3) where each loss corresponds to the type of the claim as discussed previously.
- Suppress the $\{it\}$ subscripts and consider the joint distribution of claims (C_1, C_2, C_3):

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

  Here, the marginal distribution of C_j is given by $\mathrm{F}_j(\cdot)$ with inverse $\mathrm{F}_j^{-1}(\cdot)$, and $\mathrm{H}(\cdot)$ is the copula.
- Copula: Sklar's Theorem.
- Modeling the joint distribution of the simultaneous occurrence of the claim types, when an accident occurs, provides the unique feature of our work.
- Some references are: Frees and Valdez (1998), Nelsen (1999).

Choice of copula models

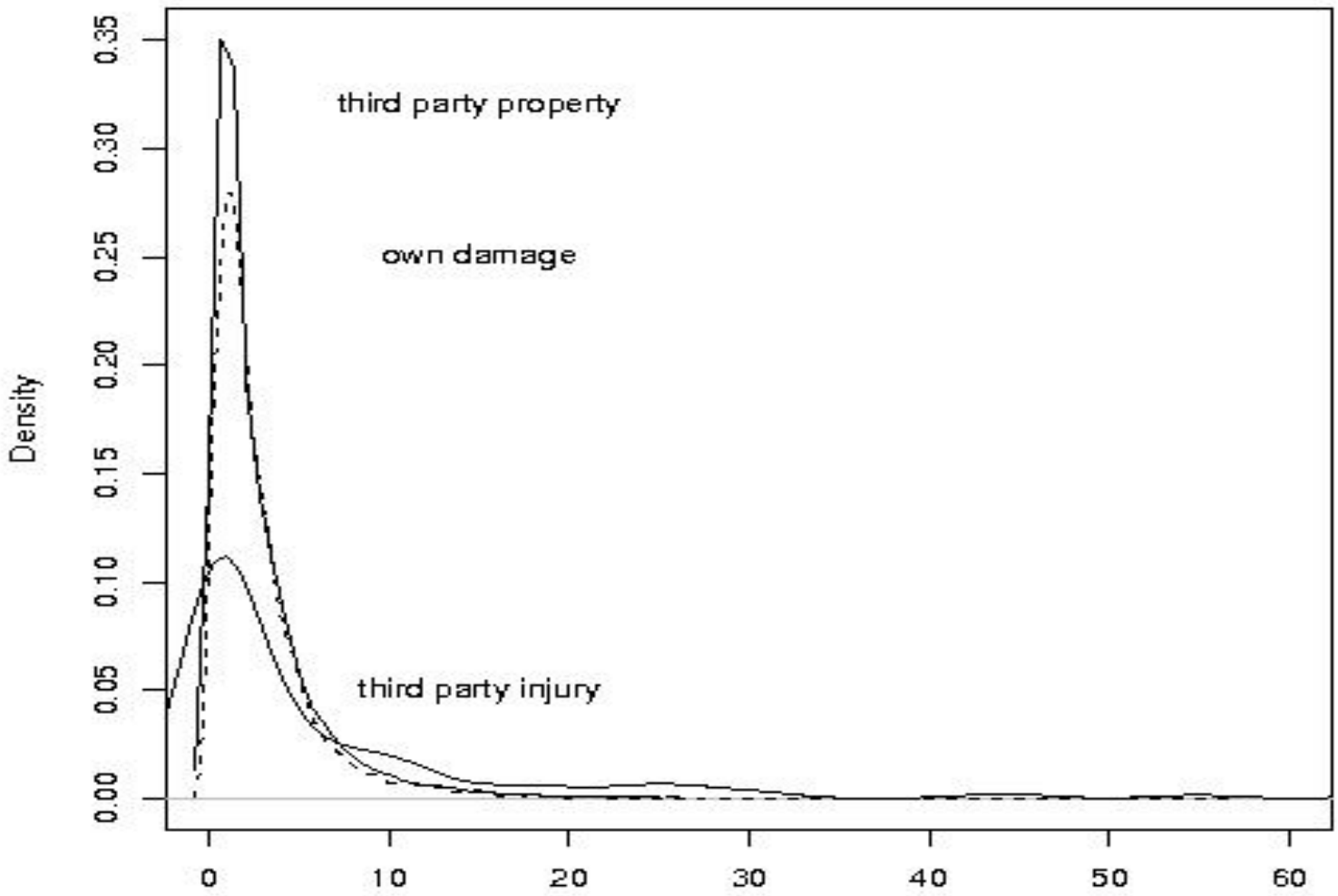
- Elliptical copulas:
  - independence copula: $\mathrm{C}(u_1, ..., u_n) = u_1 \cdots u_n$
  - Normal copula: $\mathrm{C}(u_1, ..., u_n) = \mathrm{H}\left(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_n)\right)$ where H is the joint df of a standard Normal.
  - Student-t copula: $\mathrm{C}(u_1, ..., u_n) = \mathrm{T}_r\left(t_r^{-1}(u_1), ..., t_r^{-1}(u_n)\right)$ where T is the joint df of a standard Student-t with r degrees of freedom.
    - When $r \to \infty$, we have the special case of the Normal copula.
- Frees and Wang (2005) - credibility
- Landsman and Valdez (2003) - application in finance, multivariate elliptical but with elliptical margins

Summary statistics of observed losses by claim type

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

Figure 1: Density of losses by claim type



Fitting the marginals

- We are particularly interested in accommodating the long-tail nature of claims.
- We use the generalized beta of the second kind (GB2) for each claim type with density

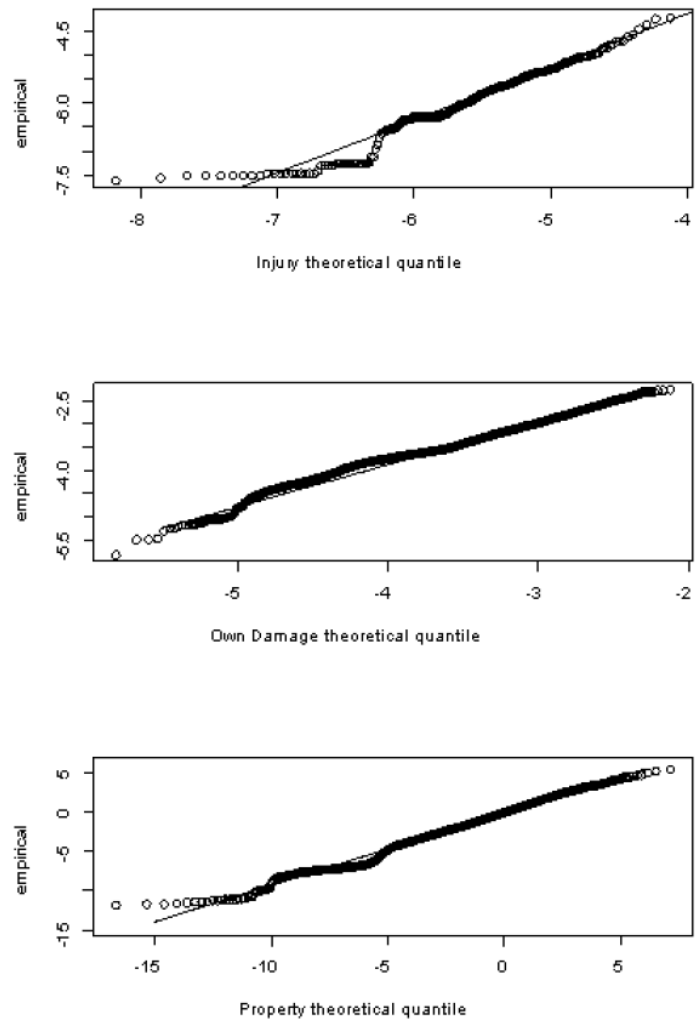
$$f_C(c) = \frac{\exp(\alpha_1 z)}{c|\sigma|B(\alpha_1,\alpha_2)\left[1+\exp(z)\right]^{\alpha_1+\alpha_2}}, \; c \geq 0,$$

  where $z = (\ln c - \mu)/\sigma$.
- μ is a location parameter, σ is a scale parameter and α_1 and α_2 are shape parameters.
- 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).

GB2 regression

- We allowed scale and shape parameters to vary by type and thus consider α_{1k}, α_{2k} and σ_k for $k = 1, 2, 3$.
- Despite its prominence, there are relatively few applications that use the GB2 in a regression context:
  - McDonald and Butler (1990) 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.
  - Sun et al. (2006) used the GB2 in a longitudinal data context to forecast nursing home utilization.
- We parameterize the location parameter as $\mu_k = \mathbf{x}'\beta_{C,k}$:
  - Interpretability of parameters.
  - Here then $\beta_{C,k,j} = \partial \ln \mathrm{E}\left(C|\ \mathbf{x}\right)/\partial x_j$, meaning that we may interpret the regression coefficients as proportional changes.

Figure 2: QQ plots for fitting the GB2 distributions



Severity likelihood

- The severity likelihood clearly depends on the combination of the types of claims observed.
- We also note the additional complication of observing claims for "own damages" type for only above the applicable excess.
  - We need to account for this in the likelihood construction.
  - Every time we observe an "own damages" claim, this would have to be conditional on observing only above the excess.

Severity likelihood, continued

- Suppose that all three types of claims are observed ($M = 7$) and that each are uncensored. In this case, the joint density would be

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

- Specifically, we can define the density for the trivariate t -distribution to be

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

and the corresponding copula as

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

Fitted copula models

Table 3.8. Fitted Copula Models

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.

What can we use the results for?

- Improve prediction because now able to predict the entire claim distribution.
- Knowledge of the entire distribution allows us to:
  - get better point estimates;
  - derive confidence interval of estimates;
  - examine the tails or extremes of the distribution; and/or
  - examine sensitivity of the parameters.
- To illustrate (in our paper), we consider the following two procedures:
  - prediction based on an individual observation, and
  - determination of expected functions of claims over different policy scenarios.

Concluding remarks

- Our paper presents a comprehensive process of hierarchical modeling of motor insurance using claims data provided to us by the General Insurance Association (GIA) of Singapore.
- The additional feature in our modeling process is the ability to account and model for the different combination of claims arising from different claim types: injury, damage to own property, and damage to third party property.
- The same process/procedure can be applied to any portfolios of insurance policies which provide a similar micro-level details of policy and claims information.
- The results can be used for better prediction of future claims experience that can be used, for instance, in experience rating.