A revisit of the hierarchical insurance claims modeling

Emiliano A. Valdez
Michigan State University

joint work with E.W. Frees*

* University of Wisconsin – Madison

Statistical Society of Canada (SSC) 2014 Annual Meeting
Toronto, ON Canada
25-28 May 2014

A collection of work

- Frees and Valdez (2008), Hierarchical Insurance Claims Modeling, *Journal of the American Statistical Association*, Vol. 103, No. 484, pp. 1457-1469.
- Frees, Shi and Valdez (2009), Actuarial Applications of a Hierarchical Insurance Claims Model, *ASTIN Bulletin*, Vol. 39, No. 1, pp. 165-197.
- Young, Valdez and Kohn (2009), Multivariate Probit Models for Conditional Claim Types, *Insurance: Mathematics and Economics*, Vol. 44, No. 2, pp. 214-228.

Basic data set-up

- "Policyholder" i is followed over time $t = 1, \ldots, 9$ years
- Unit of analysis "it" – a registered vehicle insured i over time t (year)
- Have available: exposure e_{it} and covariates (explanatory variables) $\mathbf{x}_{it}$
    - covariates often include age, gender, vehicle type, driving history and so forth
- Goal: understand how time t and covariates impact claims C_{it}.
- Statistical methods viewpoint
    - basic regression set-up - almost every analyst is familiar with:
        - part of the basic actuarial education curriculum
    - incorporating cross-sectional and time patterns is the subject of longitudinal data analysis - a widely available statistical methodology

More complex data set-up

- Some variations that might be encountered when examining insurance company records
- For each "it", could have multiple claims, $j = 0, 1, \ldots, 5$
- For each claim C_{itj}, possible to have one or a combination of three (3) types of losses:
    1. losses for injury to a party other than the insured $C_{itj,1}$ - "injury";
    2. losses for damages to the insured, including injury, property damage, fire and theft $C_{itj,2}$ - "own damage"; and
    3. losses for property damage to a party other than the insured $C_{itj,3}$ - "third party property".
- Distribution for each claim is typically medium to long-tail
- The full multivariate claim may not be observed. For example:

Distribution of claims, by claim type observed

Value of M	1	2	3	4	5	6	7	Total
Claim by Combination	(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

The hierarchical insurance claims model

- Traditional to predict/estimate insurance claims distributions:

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

- Joint density of the aggregate loss can be decomposed as:

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

- This natural decomposition allows us to investigate/model each component separately.

Model features

- Allows for risk rating factors to be used as explanatory variables that predict both the frequency and the multivariate severity components.
- Helps capture the long-tail nature of the claims distribution through the GB2 distribution model.
- Provides for a "two-part" distribution of losses - when a claim occurs, not necessary that all possible types of losses are realized.
- Allows to capture possible dependencies of claims among the various types through a t-copula specification.

Data

- Model is calibrated with detailed, micro-level automobile insurance records over nine years [1993 to 2001] of a randomly selected Singapore insurer.
- Information was extracted from the policy and claims files.
- Unit of analysis - a registered vehicle insured i over time t (year).
- The observable data consist of
  - number of claims within a year: N_{it}, for $t = 1, \ldots, T_i,\ i = 1, \ldots, n$
  - type of claim: M_{itj} for claim $j = 1, \ldots, N_{it}$
  - the loss amount: C_{itjk} for type $k = 1, 2, 3$
  - known deductible: d_{it} - applicable only for "own damages"
  - exposure: e_{it}
  - vehicle characteristics: described by the vector $\mathbf{x}_{it}$
- The data available therefore consist of

$$\{d_{it}, e_{it}, \mathbf{x}_{it}, N_{it}, M_{itj}, C_{itjk}\}.$$

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.
- These risk factors/characteristics help explain the heterogeneity among the individual policyholders.

Covariates

- Year: the calendar year - 1993-2000; treated as continuous variable.
- Vehicle Type: automobile (A) or others (O).
- Vehicle Age: in years, grouped into 6 categories -
  - 0, 1-2, 3-5, 6-10, 11-15, $\geq$16.
- Vehicle Capacity: in cubic capacity.
- Gender: male (M) or female (F).
- Age: in years, grouped into 7 categories -
  - ages $\geq$21, 22-25, 26-35, 36-45, 46-55, 56-65, $\leq$66.
- The NCD applicable for the calendar year - 0%, 10%, 20%, 30%, 40%, and 50%.

Random effects negative binomial 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

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

- continued

Negative binomial distribution model with parameters p and r:

- $\Pr(N = k|r, p) = \binom{k+r-1}{r-1} p^r (1-p)^k.$
- Here, $\sigma = \dfrac{1}{r}$ is the dispersion parameter and
- $p = p_{it}$ is related to the mean through

$$\frac{1-p_{it}}{p_{it}} = \lambda_{it}\sigma = e_{it}\exp(\mathbf{x}'_{\lambda,it}\beta_\lambda)\sigma.$$

Multinomial claim type

- Certain characteristics help describe the claims type.
- To explain this feature, we use the multinomial logit of the form

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

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

- For our purposes, the covariates in $\mathbf{x}_{M,it}$ do not depend on the accident number j nor on the claim type m, but we do allow the parameters to depend on type m.
- Such has been proposed in Terza and Wilson (1990).
- An alternative model to claim type, multivariate probit, was considered in:
    - Young, Valdez and Kohn (2009)

Severity - 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

$$f(y) = \frac{\exp(\alpha_1 z)}{y|\sigma|B(\alpha_1, \alpha_2)\left[1 + \exp(z)\right]^{\alpha_1+\alpha_2}},$$

  where $z = (\ln y - \mu)/\sigma$.
- μ is a location, σ is a scale and α_1 and α_2 are shape parameters.
- With four parameters, distribution has great flexibility for fitting heavy tailed data.
- Introduced by McDonald (1984), used in insurance loss modeling by Cummins et al. (1990).
- 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 Distribution

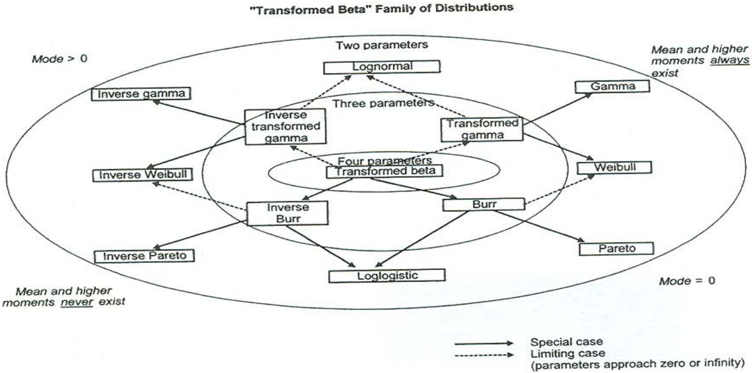


Fig. 4.7 Distributional relationships and characteristics.

Source: Klugman, Panjer and Willmot (2004), p. 72

GB2 regression

- We 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 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. (2008) used the GB2 in a longitudinal data context to forecast nursing home utilization.
- We parameterize the location parameter as $\mu_{ik} = \mathbf{x}_{ik}' \beta_k$:
  - Thus, $\beta_{k,j} = \partial \ln \mathrm{E}\left(Y \mid \mathbf{x}\right) / \partial x_j$
  - Interpret the regression coefficients as proportional changes.

Dependencies among claim types

- We use a parametric copula (in particular, the t copula).
- Suppressing the $\{i\}$ subscript, we can express the joint distribution of claims (c_1, c_2, c_3) as

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

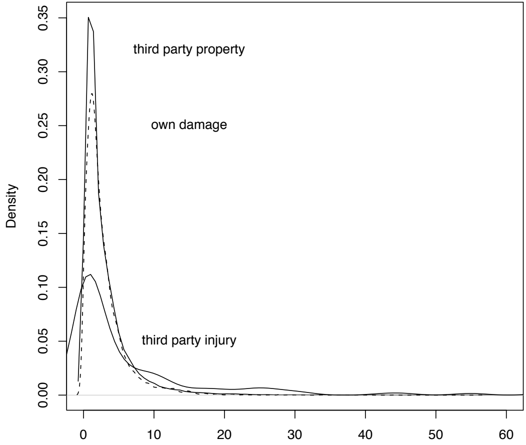
- Here, the marginal distribution of C_k is given by $\mathrm{F}_k(\cdot)$ and $\mathrm{H}(\cdot)$ is the copula.
- Modeling the joint distribution of the simultaneous occurrence of the claim types, when an accident occurs, provides the unique feature of our work.
- Some references are: Frees and Valdez (1998), Nelsen (1999).

Claim losses, by type of claim

Table 3. Summary Statistics of Claim Losses, by Type of Claim

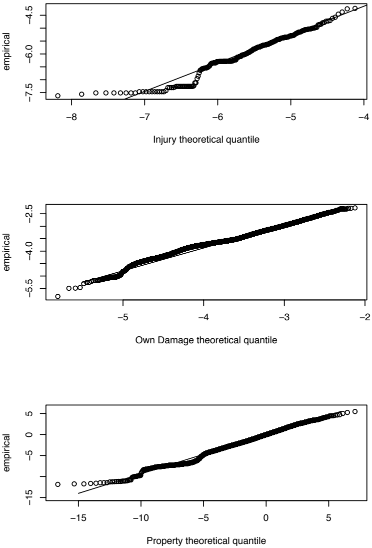
Statistic	**Third Party Injury** (C_1)	**Own Damage** (C_2) *non-censored*	**Own Damage** (C_2) *all*	**Third Party Property** (C_3)
Number	231	17,974	20,503	6,136
Mean	12,781.89	2,865.39	2,511.95	2,917.79
Standard Deviation	39,649.14	4,536.18	4,350.46	3,262.06
Median	1,700	1,637.40	1,303.20	1,972.08
Minimum	10	2	0	3
Maximum	336,596	367,183	367,183	56,156.51

Density of losses by claim type



Amount shown are in thousands.

Quantile-quantile plots for fitting GB2



Constructing the severity likelihood

The likelihood 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

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

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 copula.

- continued

For claim types $M = 1, 3, 5$, no censoring is involved and we simply integrate out the effects of the types not observed.

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.

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 copula.

- continued

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

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

- continued

For the $M = 6$ case, we have

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

where

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

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

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

The $M = 4$ case follows in the same fashion, reversing the roles of types 1 and 3.

- continued

Finally, the more complex $M = 7$ case is given by

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

and

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

With these definitions, the total severity log-likelihood for each observational unit is

$$\log\left(\mathrm{L}_S\right) = \sum_{j=1}^{7} I\left(M = j\right)\log\left(\mathrm{L}_j\right).$$

The fitted conditional severity models

Table 11. Fitted Copula Model

Parameter	Independence	Normal copula	t-copula
	Type of 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.

Some recent follow-up work

"Multivariate aggregate loss model" by Ren (IME, 2012) and "Recursions and fast Fourier transforms for a new bivariate aggregate claims model" by Jin and Ren (SAJ, 2013)

- claims arrive according to Marked Markovian arrival process (MMAP)
- also allows for dependencies between claim frequency and severity
- can get explicit forms of the aggregate loss distribution, under certain assumptions
- numerical methods to solve e.g. use of fast Fourier/Laplace transforms

Concluding remarks

- Model features
  - Allows for covariates for the frequency, type and severity components
  - Captures the long-tail nature of severity through the GB2.
  - Provides for a "two-part" distribution of losses - when a claim occurs, not necessary that all possible types of losses are realized.
  - Allows for possible dependencies among claims through a copula
  - Allows for heterogeneity from the longitudinal nature of policyholders (not claims)
- Other applications
  - Types of accidents, traffic violations, claims at-fault and no-fault
  - Could examine health care expenditure
  - Compare companies' performance using multilevel, intercompany experience