Ratemaking with a Copula-Based Multivariate Tweedie Model

Peng Shi

joint work with Xiaoping Feng and Jean-Philippe Boucher

University of Wisconsin - Madison

CAS RPM Seminar
March 10, 2015

Outline

Insurance Claims Modeling

- Two types of predictive models for insurance claims
  - Frequency severity model: Two-stage framework
  - Pure premium: Tweedie
- Both belong to GLM and straightforward to implement
- More details in “Predictive Modeling Applications in Actuarial Science”, edited by Frees, Derrig, and Meyers

Multilevel Structure

- We often assume individual risks are independent which is not always the case
- We examine a specific case where dependence among risks is often introduced through the multilevel structure
- Many examples in property-casualty insurance
  - Automobile insurance: household/fleet, multiple coverage
  - Homeowner: neighborhood, multi-peril coverage
  - Health: individual/group
  - Worker's compensation
  - ⋮
- One can argue that dependence among risks is important for claim management
- Our goal is to incorporate the correlation into the claims modeling framework

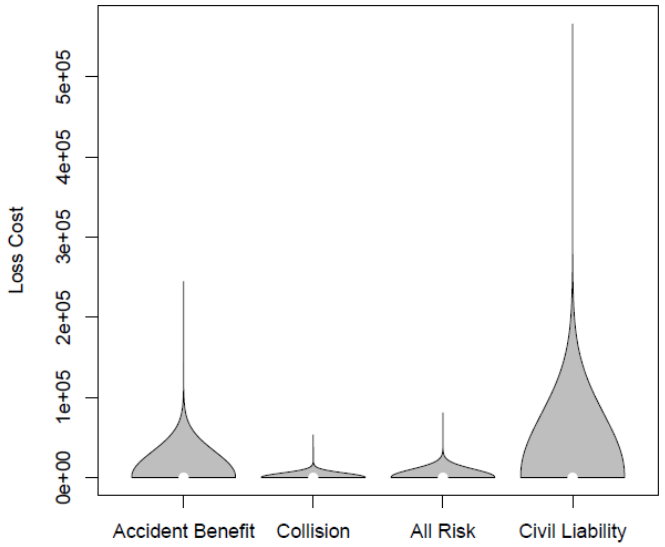
Automobile Insurance

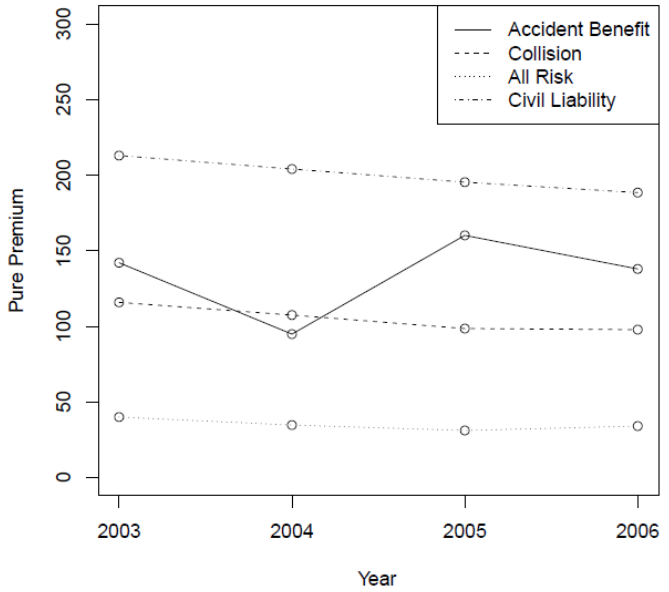
- Personal automobile insurance in Canada
- Four types of coverage
  - Accident benefit - no-fault insurance benefit to the injured insured
  - Collision - damage to the policyholder's vehicle due to collision
  - All risk - damage to the policyholder's vehicle due to risks other than collision
  - Civil liability - bodily injury and property damage of third party
- We examine a longitudinal dataset for years 2003-2006
- Each policy could cover multiple vehicles

Table : Distribution of number of vehicles per policy

	1	2	3	4	Total
Number	77352	10058	253	7	87670
Percentage	88.23	11.47	0.289	0.01	100

Claim Costs


Loss Cost
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Accident Benefit
Collision
All Risk
Civil Liability


Pure Premium
0
50
100
150
200
250
300
2003
2004
2005
2006
Year
Accident Benefit
Collision
All Risk
Civil Liability

Predictors

- The data set contains basic rating variables
  - policyholder characteristics, driving history, vehicle characteristics
  - use binary predictors in the analysis

Table : Summary of rating variables

Variable	Description	Mean
young	=1 if age between 16 and 25	0.021
seinor	=1 if age more than 60	0.175
marital	=1 if married	0.731
homeowner	=1 if homeowner	0.658
experience	=1 if more than ten years of experience	0.921
conviction	=1 if positive number of convictions	0.050
newcar	=1 if new car	0.896
leasecar	=1 if lease car	0.153
business	=1 if business use	0.037
highmilage	=1 if drive more than 10,000 miles	0.704
multidriver	=1 if more than two drivers	0.063

Tweedie

- A Poisson sum of gamma random variables
  - $Y = (X_1 + \cdots + X_N)/\omega$
  - $N \sim Poisson(\omega\lambda)$
  - $Y_j \ (j = 1, \cdots, N) \sim gamma(\alpha, \gamma)$
- The Tweedie belongs to the exponential familiy with the reparameterizations:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \gamma = \phi(p-1)\mu^{p-1}$$

- The density function is shown as

$$f(y) = \exp\left[\frac{\omega}{\phi}\left(\frac{-y}{(p-1)\mu^{p-1}} - \frac{\mu^{2-p}}{2-p}\right) + S(y;\phi/\omega)\right]$$

  with $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \frac{\phi}{\omega}\mu^p$
- Dispersion modeling?
  - Tweed GLM: $g_\mu(\mu) = \mathbf{x}'\boldsymbol{\beta}$
  - Dispersion model: $g_\phi(\phi) = \mathbf{z}'\boldsymbol{\eta}$

Bivariate Tweedie

- A *copula* is a multivariate distribution function with uniform marginals. Let $U_1, \ldots, U_J$ be J uniform random variables on (0,1). Their distribution function
$$H(u_1, \ldots, u_J) = Pr(U_1 \leq u_1, \ldots, U_J \leq u_J)$$
- Consider two tweedie marginals Y_1 and Y_2 with cdf F_1 and F_2. Define the joint distribution using the copula H such that
$$F(y_1, y_2) = H(F_1(y_1), F_2(y_2))$$
- Use Gaussian copula $H(u1, u2) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-2}(u_2))$
$$f(y_1, y_2) = \begin{cases} H(F_1(0), F_2(0)) & \text{if } y_1 = 0 \text{ and } y_2 = 0 \\ f_1(y_1) h_1(F_1(y_1), F_2(0)) & \text{if } y_1 > 0 \text{ and } y_2 = 0 \\ f_2(y_2) h_2(F_1(0), F_2(y_2)) & \text{if } y_1 = 0 \text{ and } y_2 > 0 \\ f_1(y_1) f_2(y_2) h(F_1(y_1), F_2(y_2) & \text{if } y_1 > 0 \text{ and } y_2 > 0 \end{cases}$$
- Here
$$h(u_1, u_2) = \partial H(u_1, u_2) / \partial u_1 \partial u_2$$
$$h_j(u_1, u_2) = \partial H(u_1, u_2) / \partial u_j \text{ for } j = 1, 2$$
- They have close-form expressions for Gaussian copula

Multivariate Tweedie

- Let Y_{ikjt} denote the insurance cost for
  - Household (Cluster) i $(=1,\cdots,M)$
  - Vehicle k $(=1,\cdots,K_i)$
  - Coverage type j $(=1,\cdots,J_i)$
  - Period t $(=1,\cdots,T_i)$
- For example $K_2=2$, $J_i=3$, $T_i=4$, let $\mathbf{Y}_i=(Y_{i111},\cdots,Y_{i114},Y_{i121},\cdots,Y_{i124},\cdots,Y_{i231},\cdots,Y_{i234})'$
- Each marginal follows Tweedie distribution
- Using Gaussian copula to build the joint distribution $F_i(\mathbf{y}_i)=H(F(y_{i111}),\cdots,F(Y_{i114}),F(Y_{i121}),\cdots,F(Y_{i124}),\cdots,F(Y_{i231}),\cdots,F(Y_{i234}))$
- Need to specify the association matrix R in the gaussian copula

Multivariate Tweedie

- Define $R = B \otimes P$ where

$$B = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix} \text{ and } P = \begin{pmatrix} P_{11} & \sigma_{12}P_{12} & \sigma_{13}P_{13} \\ \sigma_{21}P_{21} & P_{22} & \sigma_{23}P_{23} \\ \sigma_{31}P_{31} & \sigma_{32}P_{32} & P_{33} \end{pmatrix}$$

- Further

$$\sigma_{jj'} = \frac{\tau_{jj'}\sqrt{1-\rho_j^2}\sqrt{1-\rho_{j'}^2}}{1-\rho_j\rho_{j'}} \text{ and } P_{jj'}^{AR} = \begin{pmatrix} 1 & \rho_{j'} & \rho_{j'}^2 & \rho_{j'}^3 \\ \rho_j & 1 & \rho_{j'} & \rho_{j'}^2 \\ \rho_j^2 & \rho_j & 1 & \rho_{j'} \\ \rho_j^3 & \rho_j^2 & \rho_j & 1 \end{pmatrix}$$

- Summarize dependence parameters
  - Between vehicles δ
  - Between coverage types τ_{12}, τ_{13}, τ_{23}
  - Temporal ρ_1, ρ_2, ρ_3

Composite Likelihood

- Use composite likelihood to estimate model parameters

$$cl_i(\boldsymbol{\theta};\mathbf{y}) = \frac{1}{m_i - 1}(sum\ of\ bivariate\ loglik)$$

- Use the inverse of the Godambe information matrix to get standard error

$$G_N^{-1}(\boldsymbol{\theta}) = H_N^{-1}(\boldsymbol{\theta})J_N(\boldsymbol{\theta})H_N^{-1}(\boldsymbol{\theta})$$

  where $H_N(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 cl_i(\boldsymbol{\theta};y_i)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}$ and $J_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial cl_i(\boldsymbol{\theta};y_i)}{\partial\boldsymbol{\theta}}\frac{\partial cl_i(\boldsymbol{\theta};y_i)}{\partial\boldsymbol{\theta}'}$

- Model Comparison
  - CLAIC = $-2cl(\boldsymbol{\theta};\mathbf{y}) + 2tr(J(\boldsymbol{\theta})H(\boldsymbol{\theta})^{-1})$
  - CLBIC = $-2cl(\boldsymbol{\theta};\mathbf{y}) + \log(\dim(\boldsymbol{\theta}))tr(J(\boldsymbol{\theta})H(\boldsymbol{\theta})^{-1})$

Marginal Models

- Model dispersion as well as mean
- Different predictors for each type

Table : Parameter estimates for accident benefit

Mean			Dispersion		
	Est.	S.E.		Est.	S.E.
intercept	5.732	0.13	intercept	7.107	0.05
conviction	0.807	0.14	homeowner	-0.099	0.03
homeowner	-0.225	0.07	experience	0.489	0.05
experience	-0.324	0.12	multidriver	-0.237	0.06
young	-0.911	0.22	marrital	0.174	0.04
senior	-0.466	0.11	senior	0.151	0.05
highmilage	-0.578	0.08			
p	1.703	0.00			

Dependence

- Strong association among coverage types
- Small serial correlation and cluster effect

Table : Estimates of dependence parameters

	Est.	S.E.
ρ_1	0.163	0.022
ρ_2	0.051	0.013
ρ_3	0.099	0.015
ρ_4	0.101	0.011
τ_{12}	0.436	0.010
τ_{13}	0.049	0.018
τ_{14}	0.641	0.009
τ_{23}	0.013	0.012
τ_{24}	0.351	0.007
τ_{34}	0.021	0.010
δ	0.082	0.020

Model Comparison

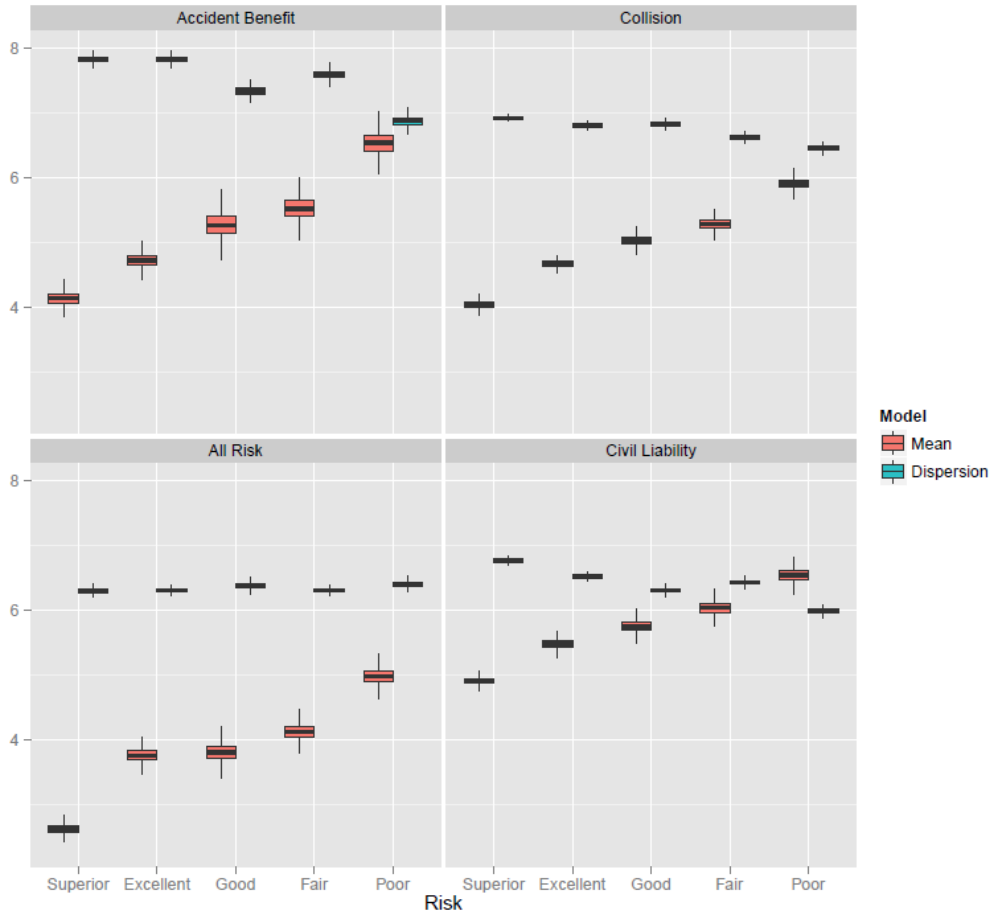
- Table below summarizes the goodness-of-fit statistics of alternative specifications
- Smaller statistics indicate better fit

Table : Goodness-of-fit statistics

Model	Description	CLAIC	CLBIC
$M0$	independence	958,513	959,142
$M1$	no temporal	957,747	958,375
$M2$	no cross-sectional	958,501	959,130
$M3$	no cluster	957,734	958,363
$M4$	no dispersion	958,491	959,120
$M5$	full model	957,730	958,358

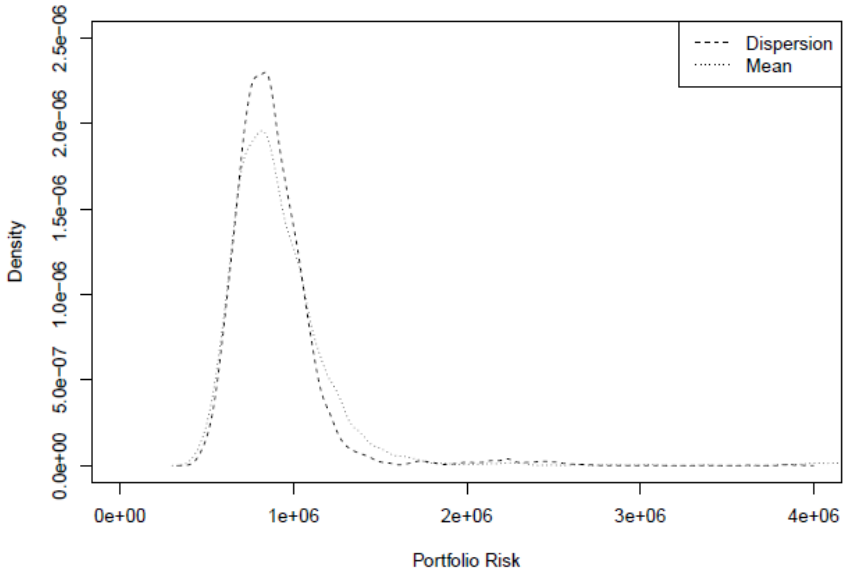
Individual Risk

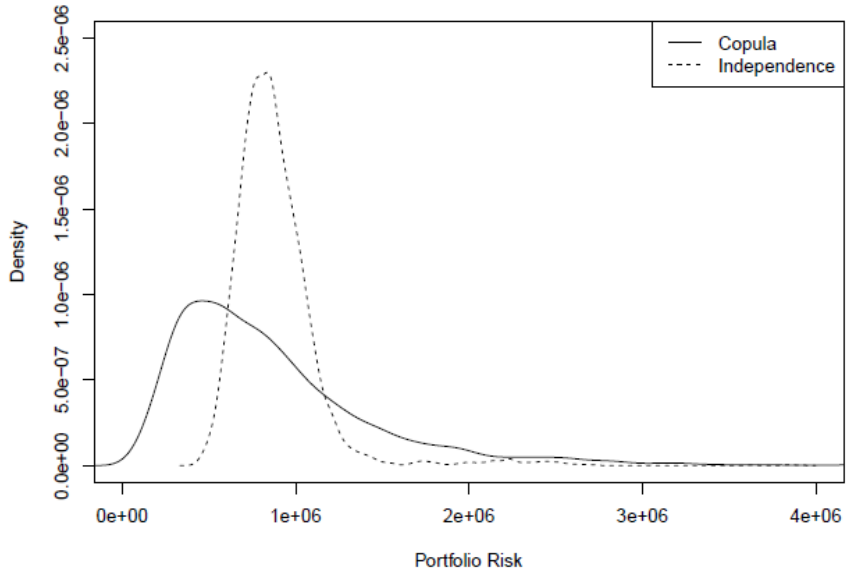
	young	senior	marital	homeowner	experience	conviction	newcar	leasecar	business	highmilage	multidriver
Excellent	0	1	1	1	1	0	0	0	0	1	0
Very Good	0	1	1	1	1	0	1	1	1	0	0
Good	0	1	1	1	0	1	1	0	1	1	0
Fire	0	1	1	1	1	1	1	1	1	0	1
Poor	0	0	0	0	0	1	1	1	1	0	1



Portfolio Risk

- Left: mean v.s. dispersion
- Right: independence v.s. copula





Conclusion

- We focused on the multilevel structure of claims data
- Tweedie model was considered as an example

Thank you for your kind attention.

Learn more about my research at:
`https://sites.google.com/a/wisc.edu/peng-shi/`