Frameworks for General Insurance Ratemaking: Beyond the Generalized Linear Model

Peng Shi† and James Guszcza‡

† University of Wisconsin-Madison
‡ Deloitte Consulting

CAS RPM Seminar
March 10, 2015

Outline

Background

- Some background
  - Predictive modeling book edited by Frees, Meyers and Derrig
  - This case study contributes a chapter in Volume II
  - Data and code will be available on book website
- Chapter goal: discuss pure premium ratemaking within a broader statistical modeling framework
- Unique features of insurance data require advanced statistical methods
  - Heavy tailed and skewed data
  - Multivariate nature of bundling products
- We discuss different modeling strategy, and we emphasize that model selection depends on the data format

Some Notations

- For each policy i, an analyst could observe
  - N_i - the number of claims
  - K_i - the type of claims
  - Y_{ink} - the amount of each claim by type
  - $Y_{in} = \sum_k Y_{ink}$, $n = 1, \cdots, N_i$ - amount of each claim
  - $S_{ik} = Y_{i1k} + \cdots + Y_{iN_ik}$ - aggregate claim amount by type
  - $S_i = \sum_k S_{ik}$ - aggregate claim amount for policyholder i

Personal Auto Dataset

- Massachusetts automobile claims dataset from CAR
  - Made public by Massachusetts Executive Office of Energy and Environmental Affairs
  - Contain experience in year 2006 for about 3.25 million policies
  - Two types of claims: liability and PIP
- We draw a random sample of 150,000 policyholders (two-third training and one-third validation)

Table : Claim frequency

Count	0	1	2	3	4	4+
Frequency	95,443	4,324	219	12	2	0

Table : Percentiles of claim size

	5%	10%	25%	50%	75%	90%	95%
Liability	237.00	350.00	675.50	1,464.00	3,465.00	10,596.90	19,958.75
PIP	2.00	5.00	84.00	1,371.50	3,300.00	7,548.50	8,232.00

Covariates

	Mean			Average Loss		
	Overall	No claim	≥ 1 claim	Liability	PIP	Total
Rating Group						
A - adult	0.747	0.749	0.703	155.20	18.45	173.65
B - business	0.014	0.014	0.014	199.65	16.48	216.13
I - <3 yrs exp	0.043	0.042	0.078	332.38	26.24	358.63
M - 3-6 yrs exp	0.044	0.043	0.067	283.92	22.32	306.24
S - senior	0.152	0.153	0.138	119.15	12.29	131.44
Territory Group						
1 - least risky	0.185	0.188	0.132	92.53	8.76	101.29
2	0.193	0.194	0.167	135.00	9.82	144.81
3	0.113	0.114	0.091	137.21	7.47	144.68
4	0.201	0.201	0.194	154.69	16.39	171.08
5	0.189	0.187	0.227	203.39	24.58	227.97
6 - most risky	0.120	0.117	0.189	296.94	47.58	344.52

Tweedie

- A Poisson sum of gamma random variables
  - $S_i = (Y_{i1} + \cdots + Y_{iN_i})/\omega_i$
  - $N_i \sim Poisson(\omega_i \lambda_i)$
  - $Y_{ij}\ (j = 1, \cdots, N_i) \sim gamma(\alpha, \gamma_i)$
- The Tweedie belongs to the exponential familiy with the reparameterizations:

$$\lambda_i = \frac{\mu_i^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \gamma_i = \phi(p-1)\mu_i^{p-1}$$

- Location μ, dispersion ϕ, and power p, denoted by $Tweedie(\mu, \phi, p)$

$$\mathrm{E}(Y_i) = \mu_i \quad \text{and} \quad \mathrm{Var}(Y_i) = \frac{\phi}{\omega_i}\mu_i^p$$

- Data availability
  - Both S_i and N_i are observed

$$f_i(n,s) = a(n,s;\phi/\omega_i,p)\exp\left\{\frac{\omega_i}{\phi}b(s;\mu_i,p)\right\}$$

  - Only S_i are recorded

$$f_i(y) = \exp\left[\frac{\omega_i}{\phi}b(s;\mu_i,p) + c(s;\phi/\omega_i)\right]$$

- Dispersion modeling?
  - Tweed GLM: $g_\mu(\mu_i) = \mathbf{x}_i'\beta$
  - Dispersion model: $g_\phi(\phi_i) = \mathbf{z}_i'\eta$

General Insurance Ratemaking
Shi Guszcza
Introduction
Data
Univariate Modeling
Multivariate Modeling
Prediction
Concluding Remarks

	Cost and Claim Counts				Cost Only			
	Mean Model		Dispersion Model		Mean Model		Dispersion Model	
Parameter	Est	S.E.	Est	S.E.	Est	S.E.	Est	S.E.
intercept	5.634	0.087	5.647	0.083	5.634	0.088	5.646	0.084
rating group = A	0.267	0.070	0.263	0.071	0.267	0.071	0.263	0.072
rating group = B	0.499	0.206	0.504	0.211	0.500	0.209	0.506	0.213
rating group = I	1.040	0.120	1.054	0.106	1.040	0.121	1.054	0.108
rating group = M	0.811	0.122	0.835	0.113	0.811	0.123	0.834	0.114
territory group = 1	-1.209	0.086	-1.226	0.086	-1.210	0.087	-1.226	0.087
territory group = 2	-0.830	0.083	-0.850	0.080	-0.831	0.084	-0.850	0.081
territory group = 3	-0.845	0.095	-0.863	0.097	-0.845	0.097	-0.862	0.098
territory group = 4	-0.641	0.081	-0.652	0.077	-0.641	0.082	-0.652	0.078
territory group = 5	-0.359	0.080	-0.368	0.074	-0.360	0.081	-0.368	0.075
p	1.631	0.004	1.637	0.004	1.629	0.004	1.634	0.004
dispersion								
intercept	5.932	0.015	5.670	0.041	5.968	0.016	5.721	0.043
rating group = A			0.072	0.034			0.064	0.035
rating group = B			0.006	0.101			0.010	0.105
rating group = I			-0.365	0.051			-0.356	0.054
rating group = M			-0.206	0.054			-0.209	0.056
territory group = 1			0.401	0.042			0.374	0.043
territory group = 2			0.323	0.039			0.301	0.040
territory group = 3			0.377	0.047			0.365	0.048
territory group = 4			0.266	0.037			0.260	0.039
territory group = 5			0.141	0.036			0.132	0.037
loglik		-61121.090		-60988.180		-60142.140		-60030.140

Frequency-Severity Models

- Suppose one can observe data at claim level, i.e. both N_i and Y_{in} are available
- Two-part model follows

$$f(N,Y) = f(N) \times f(Y|N)$$

- Based on conditional decomposition and does not require independence between Y and N like Tweedie
- Use count regression for the frequency component $f(N)$
  - Poisson, NB, Zero-inflated, Hurdle ... (see *Volume I*)
- Use fat-tailed regression for the severity component $f(Y|N)$
  - GLM, parametric (GG,GB2 etc.), quantile regression ... (see *Volume I*)
- The above formulation allows us to estimate the two parts separately

- Suppose one can observe data only at policy level, i.e. S_i or $\{N_i, S_i\}$ are available
- Strategy:
  - Model the mass probability at zero, i.e. $Pr(S=0)$, using a binary regression, such as logit or probit.
  - Model the positive claim amount, i.e. $f_S(s|S>0)$, using a fat-tailed regression.
- Likelihood

$$f_S(s) = \begin{cases} Pr(S=0) & s=0 \\ f_S(s|S>0)\times Pr(S>0) & s>0 \end{cases}$$

- Estimation

$$\begin{aligned} loglik = & \sum_{\{i:S_i=0\}} Pr(S_i=0) + \sum_{\{i:S_i>0\}} Pr(S_i>0) & \leftarrow frequency \\ & + \sum_{\{i:S_i>0\}} \ln f_S(s_i|S_i>0) & \leftarrow severity \end{aligned}$$

	Frequency				Severity			
	NegBin		ZINB		Gamma		GG	
Parameter	Est	S.E.	Est	S.E.	Est	S.E.	Est	S.E.
intercept	-2.559	0.051	-2.185	0.865	8.179	0.066	7.601	0.079
rating group = A	0.039	0.044	-0.133	0.678	0.235	0.056	0.207	0.064
rating group = B	0.186	0.130	-0.025	0.835	0.382	0.167	0.306	0.190
rating group = I	0.793	0.067	0.551	0.873	0.257	0.084	0.259	0.096
rating group = M	0.550	0.070	0.398	0.683	0.284	0.089	0.208	0.102
territory group = 1	-0.866	0.053	-1.068	0.121	-0.376	0.068	-0.245	0.079
territory group = 2	-0.647	0.050	-0.867	0.128	-0.223	0.064	-0.166	0.073
territory group = 3	-0.703	0.060	-0.777	0.111	-0.168	0.077	-0.115	0.088
territory group = 4	-0.517	0.048	-0.655	0.091	-0.175	0.061	-0.119	0.070
territory group = 5	-0.283	0.046	-0.451	0.112	-0.117	0.059	-0.053	0.067
zero model								
intercept			-0.104	1.709				
rating group = A			-1.507	0.818				
rating group = B			-2.916	3.411				
rating group = I			-5.079	5.649				
rating group = M			-1.260	1.388				
territory group = 1			-2.577	11.455				
territory group = 2			-3.894	52.505				
territory group = 3			-0.509	0.965				
territory group = 4			-1.145	2.264				
territory group = 5			-1.583	4.123				
loglik	-19147.500		-19139.000		-43748.500		-43504.510	

- Two types of coverage: S_1-Liability, S_2-PIP
- Use Tweedie for S_1 and another Tweedie for S_2
- Use a parametric copula H to construct the joint distribution of S_1 and S_2

$$f(s_1,s_2) = \begin{cases} H(F_1(0),F_2(0)) & \text{if } s_1 = 0 \text{ and } s_2 = 0 \\ f_1(s_1)h_1(F_1(s_1),F_2(0)) & \text{if } s_1 > 0 \text{ and } s_2 = 0 \\ f_2(s_2)h_2(F_1(0),F_2(s_2)) & \text{if } s_1 = 0 \text{ and } s_2 > 0 \\ f_1(s_1)f_2(s_2)h(F_1(s_1),F_2(s_2)) & \text{if } s_1 > 0 \text{ and } s_2 > 0 \end{cases}$$

Tweedie		
	Marginal	Frank Copula
θ		4.659
		(0.332)
Loglik	65930.30	65520.92
$\chi^2(1)$		818.76
Double GLM		
	Marginal	Frank Copula
θ		5.580
		(0.384)
Loglik	65771.47	65308.59
$\chi^2(1)$		925.76
$\chi^2(18)$	317.66	424.66

- Two semi-continuous claim outcomes
- Consider four scenarios: $\{S_1=0,S_2=0\}$, $\{S_1>0,S_2=0\}$, $\{S_1=0,S_2>0\}$, $\{S_1>0,S_2>0\}$
- The joint distribution can be expressed as

$$
\begin{aligned}
& f(s_1,s_2) \\
= & \begin{cases}
\Pr(S_1=0,S_2=0) & \text{if } s_1=0,s_2=0 \\
\Pr(S_1>0,S_2=0)\times f_1(s_1|s_1>0) & \text{if } s_1>0,s_2=0 \\
\Pr(S_1=0,S_2>0)\times f_2(s_2|s_2>0) & \text{if } s_1=0,s_2>0 \\
\Pr(S_1>0,S_2>0)\times f(s_1,s_2|s_1>0,s_2>0) & \text{if } s_1>0,s_2>0
\end{cases}
\end{aligned}
$$

- Define $R_1=I(S_1>0)$ and $R_2=I(S_2>0)$

- Bivariate frequency (R_1, R_2)
  - Copula

$$\begin{cases} \Pr(R_1 = 1, R_2 = 1) = 1 - F_1(0) - F_2(0) - H(F_1(0), F_2(0)) \\ \Pr(R_1 = 1, R_2 = 0) = F_2(0) - H(F_1(0), F_2(0)) \\ \Pr(R_1 = 0, R_2 = 1) = F_1(0) - H(F_1(0), F_2(0)) \\ \Pr(R_1 = 0, R_2 = 0) = H(F_1(0), F_2(0)) \end{cases}$$

  - Dependence ratio (see Chapter)
  - Odds ratio (see Chapter)
- Bivariate severity (S_1, S_2)
  - Use another copula for the joint distribution of (S_1, S_2)

$$\begin{aligned} & f(s_1, s_2 | s_1 > 0, s_2 > 0) \\ =& h(F_1(s_1|s_1 > 0), F_2(s_{i2}|s_2 > 0)) \prod_{j=1}^{2} f_j(s_j|y_j > 0) \end{aligned}$$

Parameter	Dependence Ratio Estimate	Dependence Ratio StdErr	Odds Ratio Estimate	Odds Ratio StdErr	Frank Copula Estimate	Frank Copula StdErr
Liability						
rating group = A	-0.008	0.046	-0.003	0.046	-0.006	0.095
rating group = B	0.210	0.137	0.202	0.137	0.206	0.094
rating group = I	0.680	0.068	0.795	0.072	0.781	0.022
rating group = M	0.415	0.075	0.471	0.077	0.455	0.019
territory group = 1	-0.739	0.057	-0.795	0.058	-0.788	0.023
territory group = 2	-0.502	0.054	-0.565	0.054	-0.555	0.043
territory group = 3	-0.585	0.064	-0.643	0.065	-0.635	0.054
territory group = 4	-0.397	0.052	-0.458	0.053	-0.448	0.037
territory group = 5	-0.184	0.050	-0.231	0.051	-0.226	0.038
PIP						
rating group = A	0.356	0.124	0.363	0.124	0.362	0.099
rating group = B	0.223	0.373	0.217	0.372	0.224	0.598
rating group = I	0.872	0.179	0.968	0.180	0.961	0.137
rating group = M	1.039	0.170	1.094	0.170	1.083	0.130
territory group = 1	-1.466	0.137	-1.502	0.137	-1.498	0.124
territory group = 2	-1.182	0.123	-1.224	0.123	-1.218	0.118
territory group = 3	-1.298	0.156	-1.336	0.156	-1.331	0.144
territory group = 4	-0.874	0.110	-0.915	0.110	-0.909	0.110
territory group = 5	-0.650	0.105	-0.679	0.105	-0.677	0.080
dependence	6.893	0.309	13.847	1.094	10.182	1.084
loglik		-20698.810		-20669.230		-20676.890
Chi-square		799.420		858.580		843.260

	Liability		PIP	
Parameter	Estimate	StdErr	Estimate	StdErr
intercept	7.437	0.081	7.955	0.220
rating group = A	0.269	0.065	0.121	0.185
rating group = B	0.272	0.190	-0.156	0.523
rating group = I	0.417	0.098	-0.033	0.275
rating group = M	0.428	0.106	-0.448	0.263
territory group = 1	-0.233	0.081	-0.049	0.226
territory group = 2	-0.196	0.075	-0.519	0.190
territory group = 3	-0.090	0.090	-0.427	0.249
territory group = 4	-0.105	0.073	-0.178	0.171
territory group = 5	-0.073	0.070	-0.100	0.164
σ	1.428	0.016	1.673	0.062
κ	0.210	0.029	1.655	0.105
θ	0.326	0.047		
df	11.258	4.633		
loglik	-44041.970			
$\chi^2(1)$	7.480			
$\chi^2(2)$	48.200			

Hierarchical Model

- Examine data at claim level
- Three-part model follows

$$f(N,T,Y) = f(N) \times f(T|N) \times f(Y|N,T)$$

  - N - number of claims
  - T - the type of claim: liability, PIP, or both
  - Y - amount of claims: (Y_1), (Y_2), or (Y_1, Y_2)
- Strategy:
  - Use a count regression for $f(N)$
  - Given an accident, use a multinomial logit regression for claim type $f(T|N)$
  - Given the type of an accident, use a copula regression for the amount $f(Y|N,T)$

Hierarchical Model

- Part I: Poisson/NB2 ...
- Part II:

$$\Pr(T = Liability) = \frac{\exp(\mathbf{x}_{i1}^{'}\beta_1)}{1+\exp(\mathbf{x}_{i1}^{'}\beta_1)+\exp(\mathbf{x}_{i2}^{'}\beta_2)}$$

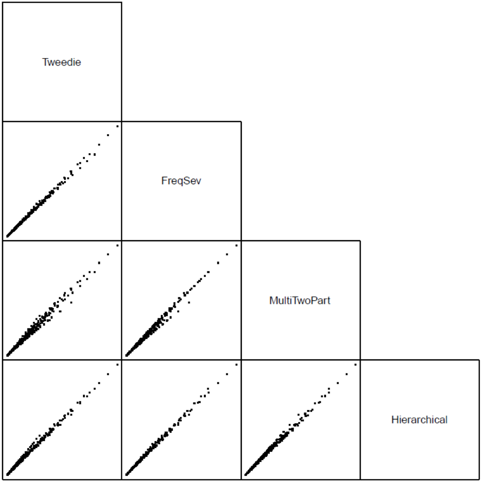
$$\Pr(T = PIP) = \frac{\exp(\mathbf{x}_{i2}^{'}\beta_2)}{1+\exp(\mathbf{x}_{i1}^{'}\beta_1)+\exp(\mathbf{x}_{i2}^{'}\beta_2)}$$

- Part III:
  - If T=Liability, $f_1(y_1) \sim Gamma/GG/GB2...$
  - If T=PIP, $f_2(y_2) \sim Gamma/GG/GB2...$
  - If T=Both, $f(y_1,y_2) = h(F_1(y_1),F_2(y_2))f_1(y_1)f_2(y_2)$

Three-Part Model

Parameter	Liability Estimate	Liability StdErr		PIP Estimate	PIP StdErr
intercept	2.799	0.126		0.390	0.178
rating group = A	0.091	0.135		0.403	0.188
rating group = B	-0.225	0.381		-0.851	0.592
rating group = I	-0.021	0.204		-0.170	0.276
rating group = M	0.027	0.229		0.731	0.278
territory group = 1	0.429	0.200		0.287	0.232
territory group = 2	0.028	0.155		-0.210	0.191
territory group = 3	0.299	0.221		0.088	0.261
territory group = 4	-0.226	0.135		-0.254	0.166
territory group = 5	0.003	0.138		0.070	0.163
Maximum Likelihood Analysis of Variance					
Source	DF	Chi-Square	p-value		
Intercept	2	766.340	$<$0.0001		
Rating group	8	26.480	0.001		
Territory group	10	54.750	$<$0.0001		
Likelihood Ratio	40	55.890	0.049		

Out-of-Sample Comparison

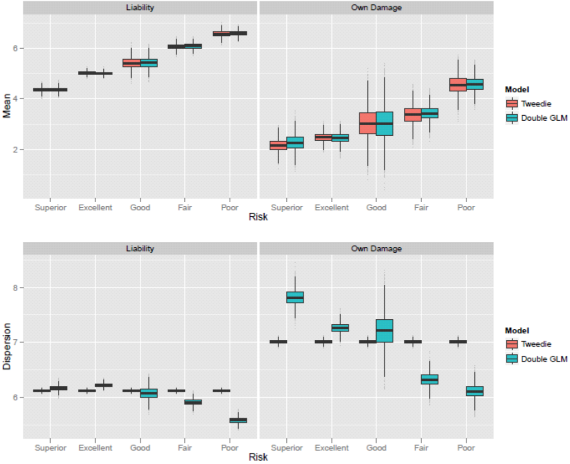


- Risk class profile

Risk Class	Rating Group					Territory Group					
	=1	=2	=3	=4	=5	=1	=2	=3	=4	=5	=6
Superior	0	0	0	0	1	1	0	0	0	0	0
Excellent	1	0	0	0	0	0	1	0	0	0	0
Good	0	1	0	0	0	0	0	0	1	0	0
Fair	0	0	0	1	0	0	0	0	0	1	0
Poor	0	0	1	0	0	0	0	0	0	0	1

- We calculate expected cost of claims for each risk class
- We quantify the variability of prediction

Prediction - Mean and Dispersion



Prediction - Frequency

- Joint distribution for high risk

	Poor			
	Tweedie		Double GLM	
	Product	Frank	Product	Frank
$\Pr(Y_1 = 0, Y_2 = 0)$	0.9215	0.9238	0.8634	0.8727
$\Pr(Y_1 > 0, Y_2 = 0)$	0.0671	0.0649	0.1088	0.0994
$\Pr(Y_1 = 0, Y_2 > 0)$	0.0106	0.0083	0.0247	0.0154
$\Pr(Y_1 > 0, Y_2 > 0)$	0.0008	0.0030	0.0031	0.0124

- For intermediate risk, predictions from the two models are similar
- For low risk, predictions are opposite of high risk

Conclusion

- We focused on the statistical problem of pure premium ratemaking
- Important but not the sole input
- Market-based pricing considerations, such as price elasticity, consumer lifetime value, and competitors rates etc, are also important

Thank you for your kind attention.

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