Multivariate negative binomial models for insurance claim counts

Peng Shi (Northern Illinois University) and Emiliano A. Valdez (University of Connecticut)

30 November 2012, Philadelphia, PA *Temple University Seminar*

University of Connecticut

Background

For short term insurance contracts, there are generally two major components to decompose the cost of claims:

$$\text{cost of claims} = \text{frequency} \times \text{severity}$$

Many believe that the frequency component:

- may be the more important of these two components
- reveals, to an extent, more of the riskiness of the policyholder

Model improvements can be made on the prediction on frequency:

- usual accounting for heterogeneity - risk classification variables
- using additional information available in data recorded e.g. types of claim
- additional insight on the association between the types of claim

Work motivated by our empirical data

We obtained claims data from one randomly selected automobile insurance company in Singapore:

- claims experience for a single year
- focus on "non-fleet" policy
- considered only policyholders with comprehensive coverage

Summary of response and explanatory variables

Variable	Description	Mean	StdDev
Responses			
N_1	number of claims of third-party bodily injury	0.055	0.243
N_2	number of claims of own damage	0.092	0.315
N_3	number of claims of third-party property damage	0.036	0.195
Covariates			
young	equals 1 if age is less than 35	0.402	0.490
lowncd	equals 1 if NCD is less than 20	0.358	0.479
vage	vehicle age	6.983	2.460
private	equals 1 for private car	0.904	0.295
vlux	equals 1 for luxury car	0.062	0.241
smallcap	equals 1 if vehicle capacity is small	0.105	0.307

Negative binomial distribution model

N is said to follow a negative binomial distribution if its pmf may be expressed as

$$\Pr(N=n)=\frac{\Gamma(\eta+n)}{\Gamma(\eta)\Gamma(n+1)}\left(\frac{1}{1+\psi}\right)^{\eta}\left(\frac{\psi}{1+\psi}\right)^{n},\ n=0,1,2,\cdots$$

and is denoted as $N\sim \mathrm{NB}(\psi,\eta)$ for $\psi,\ \eta>0$.

- Mean $\mathrm{E}(N)=\eta\psi$ and variance $\mathrm{Var}(N)=\eta\psi(1+\psi)$
- Unlike the Poisson distribution, the NB accommodates overdispersion via parameter ψ.
- As $\psi\to 0$, overdispersion vanishes and the NB converges to the Poisson distribution.

Negative binomial regression models

For regression, consider the mean specified in terms of covariates $\lambda = \eta\psi = \exp(\boldsymbol{x}'\boldsymbol{\beta})$, which may come in two different parameterizations:

- NB-I model: $\eta = \sigma^{-2}\exp(\boldsymbol{x}'\boldsymbol{\beta})$
- NB-II model: $\eta = \sigma^{-2}$

Both assume same mean structure but different dispersion ϕ such that

$$\mathrm{Var}(N|\boldsymbol{x}) = \phi\mathrm{E}(N|\boldsymbol{x}).$$

- The NB-I model implies a constant dispersion $\phi = 1 + \sigma^2$.
- The NB-II model allows for subject heterogeneity in the dispersion $\phi = 1 + \sigma^2\exp(\boldsymbol{x}'\boldsymbol{\beta})$.
- See Winkelmann (2008).

Goodness of fit of the marginal models

N_1	Observed	Fitted			
		Poisson	ZIP	NegBin-I	NegBin-II
0	7461	7452.91	7460.46	7460.96	7461.25
1	393	406.59	392.40	391.21	391.11
2	17	14.10	20.32	20.72	20.41
3	3	0.39	0.79	1.06	1.14
χ^2		**18.67**	**6.71**	**4.24**	**3.62**
N_2	Observed	Fitted			
		Poisson	ZIP	NegBin-I	NegBin-II
0	7206	7191.17	7205.87	7204.83	7206.51
1	619	644.36	616.84	619.13	616.45
2	44	36.78	48.35	46.65	47.16
3	4	1.62	2.80	3.18	3.58
4	1	0.06	0.13	0.20	0.28
χ^2		**20.49**	**6.71**	**3.47**	**2.14**
N_3	Observed	Fitted			
		Poisson	ZIP	NegBin-I	NegBin-II
0	7601	7595.88	7597.61	7600.51	7600.69
1	262	271.38	267.93	262.47	262.52
2	10	6.59	7.73	10.57	10.27
3	1	0.14	0.23	0.43	0.48
χ^2		**7.38**	**3.38**	**0.78**	**0.58**

The multivariate claim count data

Assume we have a portfolio of m policyholders observed over a fixed time period (cross-sectional).

For each policyholder, we observe a vector of claim counts expressed as

$$(N_{i1}, N_{i2}, \ldots, N_{ik}),$$

where N_{ij} is the claim count associated with type j, with $j = 1, 2, \ldots, k$, and $i = 1, 2, \ldots, m$.

Each policyholder also reveals a set of observable covariates $\boldsymbol{x}_i$ useful to sub-divide the portfolio into classes of risks with homogeneous characteristics.

We present and compare various alternatives for modeling multivariate claim counts using classical methods based on common shocks and the modern methods of using copulas.

Traditional count regression models are used in the modeling of the marginal component.

Alternative multivariate models

Some literature on incorporating dependence structure in a multivariate count outcomes:

- use of a common additive error:
  - Kocherlakota and Kocherlakota (1992), Johnson et al. (1997), Winkelmann (2000), Karlis and Meligkotsidou (2005), and Bermúdez and Karlis (2011)
  - Poisson, zero-inflated Poisson, and negative binomial to capture overdispersion
- mixture model with a multiplicative error:
  - Hausmann et al. (1984), Day and Chung (1992)
  - may be restricted to non-negative covariance structure - alternatives here are Aitchison and Ho (1989) and van Ophem (1999)

Multivariate models using copulas

Construct general discrete multivariate distribution to support more complex correlation structures through the use of copulas:

- although believed still in its infancy, increasingly been popular
- applications extend to several disciplines:
  - Economics: Prieger (2002), Cameron et al. (2004), Zimmer and Trivedi (2006)
  - Biostatistics: Song et al. (2008), Madsen and Fang (2010)
  - Actuarial science: Purcaru and Denuit (2003), Shi and Valdez (2011), Shi and Valdez (2012)
- Boucher, Denuit and Guillén (2008) provides a survey of models for insurance claim counts with time dependence.
- Genest and Nešlehová (2007) says be pre-cautious when using copulas: non-uniqueness of the copula, vague interpretation of the nature of dependence.

Multivariate models considered

Multivariate models constructed based on common shocks as follows:

$$\begin{cases} N_{i1} = U_{i1} + U_{i12} + U_{i13} \\ N_{i2} = U_{i2} + U_{i12} + U_{i23} \\ N_{i3} = U_{i3} + U_{i13} + U_{i23} \end{cases}$$

- accommodates pair-wise, but no global, dependence.

Multivariate models based on copulas:

$$F_i(n_1, n_2, n_3|\boldsymbol{x}_i) = C(F_1(n_1|\boldsymbol{x}_{i1}), F_2(n_2|\boldsymbol{x}_{i2}), F_3(n_3|\boldsymbol{x}_{i3})|\boldsymbol{x}_i; \boldsymbol{\Theta})$$

with

$$f_i(n_1, n_2, n_3|\boldsymbol{x}_i) = \sum_{l_1=0}^{1}\sum_{l_2=0}^{1}\sum_{l_3=0}^{1}(-1)^{l_1+l_2+l_3}C(u_{i1,l_1}, u_{i2,l_2}, u_{i3,l_3}|\boldsymbol{x}_i; \boldsymbol{\Theta})$$

Families of copulas examined

The class of mixtures of max-id copulas:

- extends construction of the Archimedean copula by mixing bivariate max-id copulas
- a multivariate distribution, say H, is max-id if H^γ is a CDF for all $\gamma > 0$, Joe (1993)
- allows to capture a global together with pair-wise dependence
- the copula appears complex but in manageable explicit form:

$$C(u_1, u_2, u_3) = \varphi\left(- \sum_{1 \le j < k \le 3} \log R_{jk} \left(e^{-\nu_j \varphi^{-1}(u_j)}, e^{-\nu_k \varphi^{-1}(u_k)} \right) - \sum_{j=1}^{3} \omega_j \nu_j \varphi^{-1}(u_j) \right)$$

- Joe and Hu (1997)

Families of copulas examined - continued

The class of elliptical copulas:

- has been popular because of straightforward calculation of the copula while at the same time, allowing for flexible correlation structures
- considered both Gaussian and t copulas, but ended up choosing Gaussian as a better model of the two (based on our empirical data)
- Gaussian copula:

$$C(u_1, u_2, u_3) = \mathbf{\Phi_\Sigma}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3)\right)$$

- to circumvent the problem often encountered with using copulas for multivariate discrete data, we continuitize the discrete form using jitters

Continuous extension with jitters

Define $\widetilde{N}_{ij} = N_{ij} - U_{ij}$ for $j \in \{1, 2, 3\}$ where $U_{ij} \sim \text{Uniform}(0, 1)$. The joint pdf of jittered counts for the ith policyholder $\widetilde{\boldsymbol{N}}_i = (\widetilde{N}_{i1}, \widetilde{N}_{i2}, \widetilde{N}_{i3})$ may be expressed as:

$$\widetilde{f}_i(\widetilde{n}_{i1}, \widetilde{n}_{i2}, \widetilde{n}_{i3}|\boldsymbol{x}_i) = c(\widetilde{F}_1(\widetilde{n}_{i1}|\boldsymbol{x}_{i1}), \widetilde{F}_2(\widetilde{n}_{i2}|\boldsymbol{x}_{i2}), \widetilde{F}_3(\widetilde{n}_{i3}|\boldsymbol{x}_{i3})|\boldsymbol{x}_i; \boldsymbol{\Theta}) \prod_{j=1}^{3} \widetilde{f}_{ij}(\widetilde{n}_{ij}|\boldsymbol{x}_{ij})$$

Retrieve the joint pmf of (N_{i1}, N_{i2}, N_{i3}) by averaging over the jitters:

$$\begin{aligned} & f_i(n_{i1}, n_{i2}, n_{i3}|\boldsymbol{x}_i) \\ =& \mathrm{E}_{\boldsymbol{U}_i}\left(c(\widetilde{F}_1(n_{i1} - U_{i1}|\boldsymbol{x}_{i1}), \widetilde{F}_2(n_{i2} - U_{i2}|\boldsymbol{x}_{i2}), \widetilde{F}_3(n_{i3} - U_{i3}|\boldsymbol{x}_{i3})|\boldsymbol{x}_i; \boldsymbol{\Theta}) \prod_{j=1}^{3} \widetilde{f}_j(n_{ij} - U_{ij}|\boldsymbol{x}_{ij}) \right) \end{aligned}$$

Based on relations:

- $\widetilde{F}_j(\widetilde{n}_{ij}|\boldsymbol{x}_{ij}) = F_j([\widetilde{n}_{ij}]|\boldsymbol{x}_{ij}) + (\widetilde{n}_{ij} - [\widetilde{n}_{ij}]) f_j([\widetilde{n}_{ij} + 1]|\boldsymbol{x}_{ij})$
- $\widetilde{f}_j(\widetilde{n}_{ij}|\boldsymbol{x}_{ij}) = f_j([\widetilde{n}_{ij} + 1]|\boldsymbol{x}_{ij})$

Multivariate NB-I with common shocks

	Bodily Injury		Own Damage		Property Damage	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
intercept	-3.671	-8.91	-2.551	-8.80	-3.264	-8.83
young	0.317	1.29	0.136	1.33	0.287	1.27
lowncd	0.478	1.97	0.247	2.40	0.606	2.74
vage	-0.034	-0.44	0.024	0.26	-0.024	-0.35
vage×vage	0.002	0.61	-0.011	-1.40	0.003	1.45
private	-0.930	-2.87	0.403	2.17	-1.091	-4.00
vlux	0.635	1.55	0.099	0.55	-1.493	-1.49
smallcap	-0.220	-0.58	-0.775	-3.54	-0.605	-1.59
	Estimate	95% CI				
$\ln \lambda_{12}$	-3.379	(-3.535, -3.223)				
$\ln \lambda_{13}$	-3.709	(-3.899, -3.518)				
$\ln \lambda_{23}$	-5.762	(-6.417, -5.108)				
$\ln \sigma^2$	-4.917	(-7.827, -2.007)				
	Goodness-of-Fit					
Loglikelihood	-4496.69					
AIC	9049.37					
BIC	9244.57					

Multivariate NB-I using mixtures of max-id copulas

	Bodily Injury		Own Damage		Property Damage	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
NB-I						
intercept	-2.293	-14.53	-1.857	-13.36	-2.411	-12.70
young	0.281	2.69	0.172	2.12	0.260	2.17
lowncd	0.451	4.68	0.318	4.08	0.435	3.78
vage	-0.071	-2.14	-0.106	-3.20	-0.040	-1.01
vage×vage	0.002	1.45	0.002	0.79	0.002	1.09
private	-0.234	-1.58	0.241	1.91	-0.710	-3.70
vlux	0.017	0.08	0.026	0.17	-0.571	-1.86
smallcap	-0.442	-2.54	-0.628	-4.06	-0.767	-3.33
	Estimate	95% CI				
$\ln \sigma_1^2$	-2.239	(-3.013, -1.464)				
$\ln \sigma_2^2$	-2.415	(-3.165, -1.665)				
$\ln \sigma_3^2$	-2.391	(-3.275, -1.508)				
θ_{12}	1.429	(1.277, 1.582)				
θ_{13}	1.472	(1.289, 1.655)				
θ	1.427	(1.332, 1.521)				
	Goodness-of-Fit					
Loglikelihood	-4287.56					
AIC	8635.12					
BIC	8844.26					

Multivariate NB-II using mixtures of max-id copulas

	Bodily Injury		Own Damage		Property Damage	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
NB-II						
intercept	-2.248	-11.87	-1.851	-11.02	-2.393	-11.93
young	0.301	3.14	0.166	2.16	0.258	2.24
lowncd	0.410	4.20	0.300	3.85	0.409	3.51
vage	-0.071	-1.98	-0.101	-2.71	-0.035	-0.86
vage×vage	0.002	1.14	0.001	0.56	0.002	0.80
private	-0.261	-1.58	0.231	1.80	-0.735	-4.48
vlux	-0.034	-0.17	-0.006	-0.04	-0.634	-2.00
smallcap	-0.494	-2.74	-0.665	-4.23	-0.819	-3.46
	Estimate	95% CI				
$\ln \sigma_1^2$	0.391	(-0.072, 0.854)				
$\ln \sigma_2^2$	-0.229	(-0.791, 0.333)				
$\ln \sigma_3^2$	0.551	(-0.094, 1.195)				
θ_{12}	1.431	(1.279, 1.584)				
θ_{13}	1.480	(1.299, 1.661)				
θ	1.425	(1.343, 1.507)				
	Goodness-of-Fit					
Loglikelihood	-4286.36					
AIC	8632.72					
BIC	8841.86					

Multivariate NB-I using Gaussian copula

	Bodily Injury		Own Damage		Property Damage	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
NB-I						
intercept	-2.404	-12.71	-1.967	-11.24	-2.505	-11.55
young	0.291	2.83	0.185	2.27	0.278	2.20
lowncd	0.478	4.70	0.343	4.22	0.481	3.84
vage	-0.072	-2.04	-0.112	-3.00	-0.049	-1.22
vage×vage	0.003	1.44	0.002	0.75	0.002	1.50
private	-0.190	-1.16	0.372	2.52	-0.668	-3.79
vlux	0.017	0.08	0.007	0.05	-0.663	-1.88
smallcap	-0.465	-2.49	-0.629	-3.88	-0.842	-3.28
	Estimate	95% CI				
$\ln \sigma_1^2$	-2.940	(-3.767, -2.112)				
$\ln \sigma_2^2$	-2.998	(-3.830, -2.167)				
$\ln \sigma_3^2$	-3.199	(-4.394, -2.005)				
ρ_{12}	0.798	(0.770, 0.826)				
ρ_{13}	0.802	(0.769, 0.836)				
ρ_{23}	0.583	(0.532, 0.635)				
	Goodness-of-Fit					
Loglikelihood	-4293.58					
AIC	8647.16					
BIC	8856.30					

Multivariate NB-II using Gaussian copula

	Bodily Injury		Own Damage		Property Damage	
	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
NB-II						
intercept	-2.359	-12.32	-1.946	-11.05	-2.481	-11.32
young	0.306	2.99	0.184	2.26	0.283	2.23
lowncd	0.448	4.41	0.330	4.05	0.462	3.68
vage	-0.073	-2.00	-0.112	-2.89	-0.048	-1.14
vage×vage	0.002	1.29	0.002	0.69	0.002	1.30
private	-0.214	-1.31	0.361	2.46	-0.691	-3.87
vlux	-0.023	-0.11	-0.015	-0.10	-0.706	-1.98
smallcap	-0.506	-2.67	-0.661	-4.02	-0.879	-3.37
	Estimate	95% CI				
$\ln \sigma_1^2$	-0.175	(-0.943, 0.592)				
$\ln \sigma_2^2$	-0.716	(-1.466, 0.033)				
$\ln \sigma_3^2$	-0.061	(-1.143, 1.020)				
ρ_{12}	0.799	(0.771, 0.827)				
ρ_{13}	0.804	(0.770, 0.838)				
ρ_{23}	0.585	(0.534, 0.637)				
	Goodness-of-Fit					
Loglikelihood	-4291.81					
AIC	8643.61					
BIC	8855.28					

Predictive applications

This application examines the financial losses of third-party bodily injury, own damage, and third-party property damage, in a comprehensive coverage.

For demonstration purposes, we look into the variable $L_i = N_{i1} + N_{i2} + N_{i3}$ that is the basis of the premium for risk class i.

Five hypothetical risk classes: Ranking from high to low risk levels, they are Excellent, Very Good, Good, Fair, and Poor.

For example, the policyholders in class Excellent are older than 35, have an NCD score above 20, and drive a 10-year old private small luxury car.

Hypothetical risk classification profile

	young	lowncd	vage	private	vlux	smallcap
Excellent	0	0	10	1	1	1
Very Good	0	1	5	0	1	1
Good	1	0	5	1	0	0
Fair	0	1	0	0	1	0
Poor	1	1	0	0	0	0

Prediction estimates of claim frequency

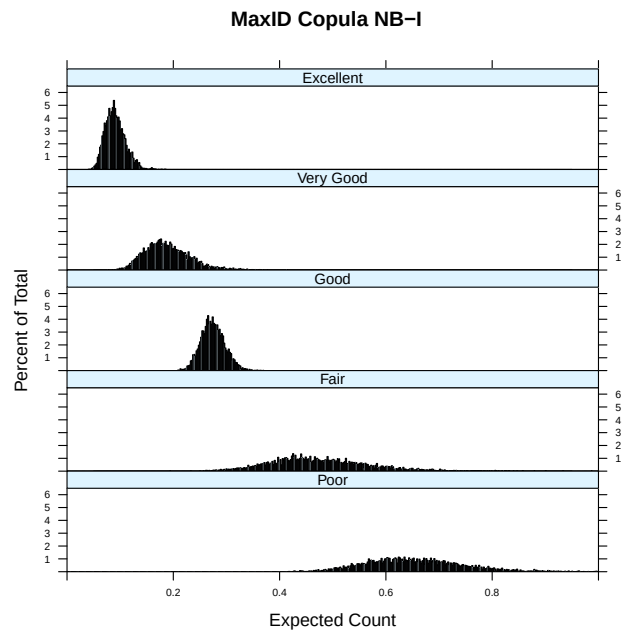
	Excellent				Very Good			
	MVNB	MaxID	Gaussian	Product	MVNB	MaxID	Gaussian	Product
ooo	0.9031	0.9469	0.9472	0.9247	0.8452	0.9017	0.8991	0.8829
oo+	0.0015	0.0010	0.0018	0.0062	0.0069	0.0049	0.0058	0.0163
o+o	0.0234	0.0232	0.0237	0.0406	0.0376	0.0282	0.0258	0.0498
+oo	0.0121	0.0092	0.0105	0.0268	0.0478	0.0225	0.0236	0.0465
o++	0.0029	0.0003	0.0002	0.0003	0.0031	0.0010	0.0006	0.0009
+o+	0.0227	0.0018	0.0019	0.0002	0.0226	0.0070	0.0072	0.0009
++o	0.0327	0.0123	0.0123	0.0012	0.0343	0.0205	0.0222	0.0026
+++	0.0017	0.0052	0.0025	0.0000	0.0024	0.0142	0.0157	0.0000
	Good				**Fair**			
	MVNB	MaxID	Gaussian	Product	MVNB	MaxID	Gaussian	Product
ooo	0.8151	0.8372	0.8283	0.7612	0.7678	0.7656	0.7699	0.7089
oo+	0.0134	0.0098	0.0125	0.0380	0.0122	0.0116	0.0124	0.0410
o+o	0.0989	0.0742	0.0745	0.1198	0.0891	0.0885	0.0761	0.1344
+oo	0.0100	0.0160	0.0125	0.0618	0.0616	0.0330	0.0266	0.0858
o++	0.0045	0.0046	0.0039	0.0060	0.0041	0.0048	0.0032	0.0078
+o+	0.0209	0.0079	0.0074	0.0031	0.0218	0.0128	0.0129	0.0050
++o	0.0332	0.0262	0.0318	0.0097	0.0391	0.0465	0.0560	0.0163
+++	0.0041	0.0241	0.0290	0.0005	0.0042	0.0372	0.0430	0.0009

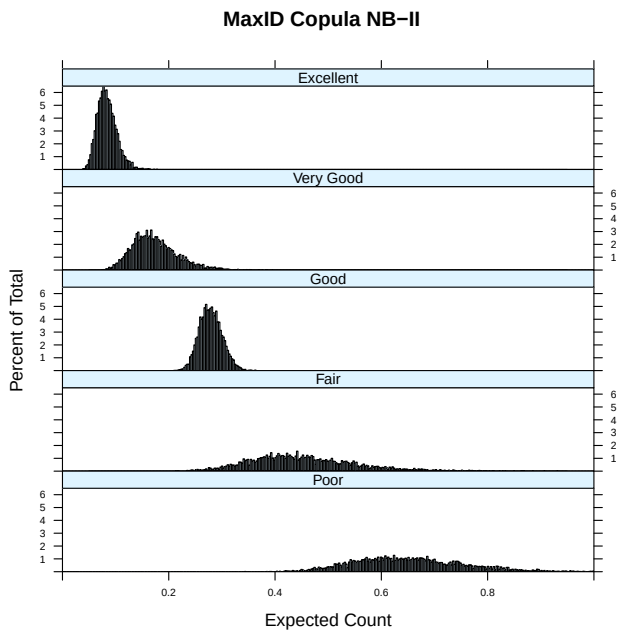
	Poor			
	MVNB	MaxID	Elliptical	Product
ooo	0.7228	0.6936	0.6962	0.5920
oo+	0.0705	0.0388	0.0413	0.0943
o+o	0.0872	0.0834	0.0727	0.1423
+oo	0.0417	0.0367	0.0236	0.1034
o++	0.0113	0.0138	0.0104	0.0227
+o+	0.0248	0.0267	0.0295	0.0165
++o	0.0346	0.0426	0.0478	0.0249
+++	0.0070	0.0644	0.0785	0.0040

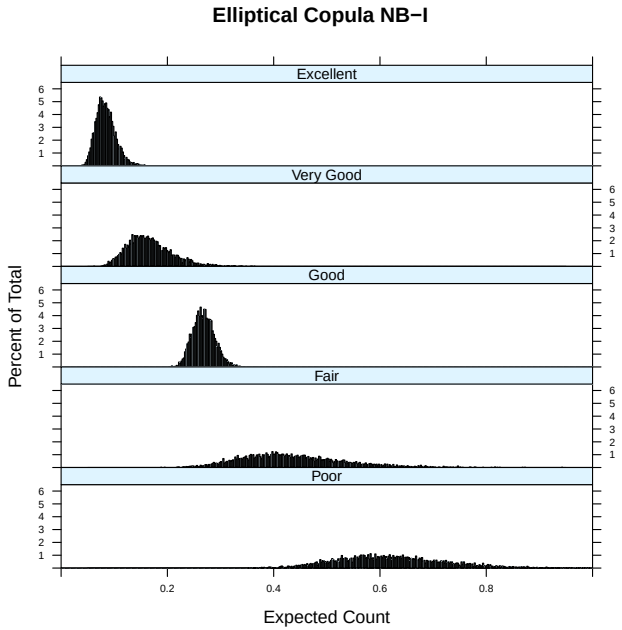
Mean and variance of total claims by risk class

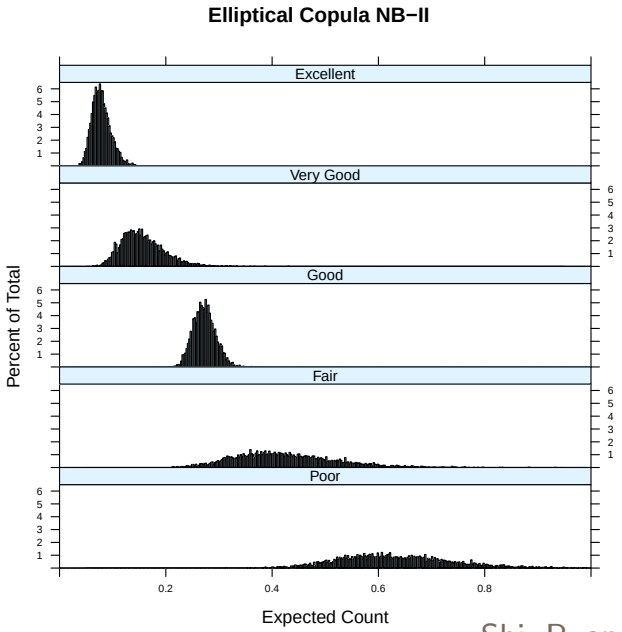
	Excellent		Very Good		Good		Fair		Poor	
	$\mathsf{E}(L_i)$	$\mathsf{Var}(L_i)$	$\mathsf{E}(L_i)$	$\mathsf{Var}(L_i)$	$\mathsf{E}(L_i)$	$\mathsf{Var}(L_i)$	$\mathsf{E}(L_i)$	$\mathsf{Var}(L_i)$	$\mathsf{E}(L_i)$	$\mathsf{Var}(L_i)$
MVNB	0.102	0.165	0.169	0.232	0.205	0.269	0.265	0.329	0.326	0.390
MaxID-I	0.087	0.178	0.181	0.394	0.273	0.582	0.460	1.003	0.645	1.462
MaxID-II	0.081	0.151	0.166	0.349	0.277	0.602	0.438	1.042	0.644	1.740
Gauss-I	0.078	0.136	0.154	0.294	0.263	0.508	0.410	0.837	0.603	1.310
Gauss-II	0.073	0.122	0.145	0.271	0.267	0.526	0.401	0.861	0.612	1.474
Indep-I	0.081	0.085	0.131	0.137	0.280	0.291	0.362	0.377	0.552	0.574
Indep-II	0.079	0.080	0.126	0.129	0.280	0.295	0.356	0.380	0.550	0.605

Predictive distributions for various models









Concluding remarks

- We considered alternative approaches to construct multivariate count regression models based on the negative binomial (NB).
- The trivariate NB models was proposed to accommodate the dependency among three types of claims: the third-party bodily injury, own damage, and third-party property damage.
- Copulas provide flexibility to model various dependence structures, allowing to separate the effects of peculiar characteristics of the margins such as thickness of tails.
- In contrast, the class of multivariate NB models based on common shocks rely on the additivity of the NB-I distribution and require a common dispersion for all marginals.
- We found that the superiority of the copula approach was supported by the better fit in our empirical analysis. We demonstrated, and made comparisons for, the usefulness of the models for predictive purposes.

- Thank you -