Longitudinal Modeling of Claim Counts using Jitters

joint work with Peng Shi, Northern Illinois University

UConn Actuarial Science Seminar
2 December 2011

Emiliano A. Valdez
Department of Mathematics
University of Connecticut
Storrs, Connecticut, USA

Outline

Background

- Two-part model for pure premium calculation: decompose total claims into claim frequency (number of claims) and claim severity (amount of claim, given a claim occurs).
- Several believe that the claim frequency, or claim counts, is the more important component.
- Past claims experience provide invaluable insight into some of the policyholder risk characteristics for experience rating or credibility ratemaking.
- Modeling longitudinal claim counts can assist to test economic hypothesis within the context of a multi-period contract.
- It might be insightful to explicitly measure the association of claim counts over time (intertemporal dependence).

Longitudinal data

- Assume we observe claim counts, N_{it}, for a group of policyholders i, for $i = 1, 2, \ldots, m$, in an insurance portfolio over T_i years.
- For each policyholder, the observable data is a vector of claim counts expressed as $(N_{i1}, \ldots, N_{iT_i})$.
- Data may be unbalanced: length of time T_i observed may differ among policyholders.
- Set of observable covariates $\mathbf{x}_{it}$ useful to sub-divide the portfolio into classes of risks with homogeneous characteristics.
- Here, we present an alternative approach to modeling longitudinal insurance claim counts using copulas and compare its performance with standard and traditional count regression models.

Literature

- Alternative models for longitudinal counts:
  - Random effects models: the most popular approach
  - Marginal models with serial correlation
  - Autoregressive and integer-valued autoregressive models
  - Common shock models
- Useful books on count regression
  - Cameron and Trivedi (1998): Regression Analysis of Count Data
  - Denuit et al. (2007): Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems
  - Frees (2009): Regression Modeling with Actuarial and Financial Applications
  - Winkelmann (2010): Econometric Analysis of Count Data
- The recent survey work of Boucher, Denuit and Guillén (2010) provides for a comparison of the various models.

Literature - continued

- Copula regression for multivariate discrete data:
  - Increasingly becoming popular
  - Applications found in various 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)
  - Modeling longitudinal insurance claim counts:
    - Frees and Wang (2006): model joint pdf of latent variables
    - Boucher, Denuit and Guillén (2010): model joint pmf of claim counts
- Be pre-cautious when using copulas for multivariate discrete observations: non-uniqueness of the copula, vague interpretation of the nature of dependence. See Genest and Nešlehová (2007).
- We adopt an approach close to Madsen and Fang (2010): joint regression analysis.

Random effects models

- To capture the intertemporal dependence within subjects, the most popular approach is to introduce a common random effect, say α_i, to each observation.
- The joint pmf for $(N_{i1}, \ldots, N_{iT_i})$ can be expressed as

$$\Pr(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i}) = \int_0^\infty \Pr(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i} | \alpha_i) f(\alpha_i) d\alpha_i$$

  where $f(\alpha_i)$ is the density function of the random effect.
- Typical assumption is conditional independence as follows:

$$\Pr(N_{i1} = n_{i1}, \ldots, N_{iT_i} = n_{iT_i} | \alpha_i) = \Pr(N_{i1} = n_{i1} | \alpha_i) \times \cdots \times \Pr(N_{iT_i} = n_{iT_i} | \alpha_i).$$

Some known random effects models

- Poisson $N_{it} \sim \text{Poisson}(\tilde{\lambda}_{it})$
  - $\tilde{\lambda}_{it} = \eta_i \lambda_{it} = \eta_i \omega_{it} \exp(\mathbf{x}_{it}^{'} \boldsymbol{\beta})$, and $\eta_i \sim \text{Gamma}(\psi, \psi)$
  - $\tilde{\lambda}_{it} = \omega_{it} \exp(\alpha_i + \mathbf{x}_{it}^{'} \boldsymbol{\beta})$, and $\alpha_i \sim \text{N}(0, \sigma^2)$
- Negative Binomial
  - NB1: $1 + 1/\nu_i \sim \text{Beta}(a, b)$
    $\Pr(N_{it} = n_{it} | \nu_i) = \frac{\Gamma(n_{it}+\lambda_{it})}{\Gamma(\lambda_{it})\Gamma(n_{it}+1)} \left(\frac{\nu_i}{1+\nu_i}\right)^{\lambda_{it}} \left(\frac{1}{1+\nu_i}\right)^{n_{it}}$
  - NB2: $\alpha_i \sim \text{N}(0, \sigma^2)$
    $\Pr(N_{it} = n_{it} | \alpha_i) = \frac{\Gamma(n_{it}+\psi)}{\Gamma(\psi)\Gamma(n_{it}+1)} \left(\frac{\psi}{\tilde{\lambda}_{it}+\psi}\right)^{\psi} \left(\frac{\tilde{\lambda}_{it}}{\tilde{\lambda}_{it}+\psi}\right)^{n_{it}}$
- Zero-inflated models
  - $\Pr(N_{it} = n_{it} | \delta_i, \alpha_i) = \begin{cases} \pi_{it} + (1-\pi_{it}) f(n_{it}|\alpha_i) & \text{if } n_{it} = 0 \\ (1-\pi_{it}) f(n_{it}|\alpha_i) & \text{if } n_{it} > 0 \end{cases}$.
  - $\log\left(\frac{\pi_{it}}{1-\pi_{it}} \middle| \delta_i\right) = \delta_i + \mathbf{z}_{it}^{'} \gamma$,
  - ZIP ($f \sim$ Poisson) and ZINB ($f \sim NB$)

Copula models

- Joint pmf using copula:

$$\Pr(N_{i1} = n_{i1}, \ldots, N_{iT} = n_{iT}) = \sum_{j_1=1}^{2} \cdots \sum_{j_T=1}^{2} (-1)^{j_1+\cdots+j_T} C(u_{1j_1}, \ldots, u_{Tj_T})$$

  Here, $u_{t1} = F_{it}(n_{it})$, $u_{t2} = F_{it}(n_{it} - 1)$, and F_{it} denotes the distribution of N_{it}

- Downside of the above specification:
  - contains 2^T terms and becomes unmanageable for large T
  - involves high-dimensional integration
  - other critiques for the case of multivariate discrete data: see Genest and Něslehová (2007)

Continuous extension with jitters

- Define $N_{it}^* = N_{it} - U_{it}$ where $U_{it} \sim \text{Uniform}(0,1)$
- The joint pdf of jittered counts for the ith policyholder $(N_{i1}^*, N_{i2}^*, \ldots, N_{iT}^*)$ may be expressed as:

$$f_i^*(n_{i1}^*, \ldots, n_{iT}^*) = c(F_{i1}^*(n_{i1}^*), \ldots, F_{iT}^*(n_{iT}^*); \boldsymbol{\theta}) \prod_{t=1}^{T} f_{it}^*(n_{it}^*)$$

- Retrieve the joint pmf of $(N_{i1}, \ldots, N_{iT})$ by averaging over the jitters:

$$f_i(n_{i1}, \ldots, n_{iT}) = \mathrm{E}_{\boldsymbol{U}_i}\left[c(F_{i1}^*(n_{i1} - U_{i1}), \ldots, F_{iT}^*(n_{iT} - U_{iT}); \boldsymbol{\theta}) \prod_{t=1}^{T} f_{it}^*(n_{it} - U_{it}) \right]$$

- Based on relations:
  - $F_{it}^*(n) = F_{it}([n]) + (n - [n]) f_{it}([n+1])$
  - $f_{it}^*(n) = f_{it}([n+1])$

Some properties with jittering

It is interesting to note that with continuous extension with jitters, we preserve:

- concordance ordering:

$$\text{If } (N_{a1}, N_{b1}) \prec_c (N_{a2}, N_{b2}), \text{ then } (N_{a1}^*, N_{b1}^*) \prec_c (N_{a2}^*, N_{b2}^*)$$

- Kendall's tau coefficient:

$$\tau(N_{a1}, N_{b1}) = \tau(N_{a1}^*, N_{b1}^*)$$

Proof can be found in Denuit and Lambert (2005).

Model specification

- Assume f_{it} follows NB2 distribution:

$$f_{it}(n) = \Pr(N_{it} = n) = \frac{\Gamma(n+\psi)}{\Gamma(\psi)\Gamma(n+1)} \left(\frac{\psi}{\lambda_{it}+\psi}\right)^{\psi} \left(\frac{\lambda_{it}}{\lambda_{it}+\psi}\right)^{n},$$

  with $\lambda_{it} = \exp(\mathbf{x}_{it}'\beta)$.

- Consider elliptical copulas for the jittered counts and examine three dependence structure (e.g. $T = 4$):

$$\text{autoregressive:} \quad \Sigma_{AR} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

$$\text{exchangeable:} \quad \Sigma_{EX} = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

$$\text{Toeplitz:} \quad \Sigma_{TOEP} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & \rho_2 & \rho_1 & 1 \end{pmatrix}$$

- Likelihood based method is used to estimate the model.
- A large number of simulations are used to approximate the likelihood.

Singapore data

- For our empirical analysis, claims data are obtained from an automobile insurance company in Singapore
- Data was over a period of nine years 1993-2001.
- Data for years 1993-2000 was used for model calibration; year 2001 was our hold-out sample for model validation.
- Focus on “non-fleet” policy
- Limit to policyholders with comprehensive coverage

Number and Percentage of Claims by Count and Year

	Percentage by Year									Overall	
Count	1993	1994	1995	1996	1997	1998	1999	2000	2001	Number	Percent
0	88.10	85.86	85.21	83.88	90.41	85.62	86.89	87.18	89.71	3480	86.9
1	10.07	12.15	13.13	14.29	8.22	13.73	11.59	11.54	9.71	468	11.7
2	1.47	2.00	1.25	1.83	0.00	0.65	1.37	0.92	0.57	50	1.25
3	0.37	0.00	0.21	0.00	1.37	0.00	0.15	0.18	0.00	6	0.15
4	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.18	0.00	2	0.05
Number	546	601	480	273	73	306	656	546	525	4006	100

Summary statistics

- Data contain rating variables including:
  - vehicle characteristics: age, brand, model, make
  - policyholder characteristics: age, gender, marital status
  - experience rating scheme: no claim discount (NCD)

Number and Percentage of Claims by Age, Gender and NCD

	Percentage by Count					Overall	
	0	1	2	3	4	Number	Percent
Person Age (in years)							
25 and younger	73.33	23.33	3.33	0.00	0.00	30	0.75
26-35	87.49	11.12	1.19	0.10	0.10	1007	25.14
36-45	86.63	11.80	1.35	0.17	0.06	1780	44.43
46-60	86.85	11.92	1.05	0.18	0.00	1141	28.48
60 and over	91.67	6.25	2.08	0.00	0.00	48	1.20
Gender							
Female	91.49	7.98	0.53	0.00	0.00	188	4.69
Male	86.64	11.86	1.28	0.16	0.05	3818	95.31
No Claims Discount (NCD)							
0	84.83	13.17	1.61	0.26	0.13	1549	38.67
10	86.21	12.58	1.20	0.00	0.00	747	18.65
20	89.21	9.25	1.54	0.00	0.00	584	14.58
30	89.16	9.49	1.08	0.27	0.00	369	9.21
40	88.60	11.40	0.00	0.00	0.00	193	4.82
50	88.83	10.46	0.53	0.18	0.00	564	14.08
Number by Count	3480	468	50	6	2	4006	100

Variable selection

- Preliminary analysis chose:
  - *young*: 1 if below 25, 0 otherwise
  - *midfemale*: 1 if mid-aged (between 30-50) female drivers, 0 otherwise
  - *zeroncd*: 1 if zero ncd, 0 otherwise
  - *vage*: vehicle age
  - *vbrand1*: 1 for vehicle brand 1
  - *vbrand2*: 1 for vehicle brand 2
- Variable selection procedure used is beyond scope of our work.

Estimation Results

Estimates of standard longitudinal count regression models

	RE-Poisson		RE-NegBin		RE-ZIP		RE-ZINB	
Parameter	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
intercept	-1.7173	<.0001	1.6404	0.1030	-1.6780	<.0001	-1.7906	<.0001
young	0.6408	0.0790	0.6543	0.0690	0.6232	0.0902	0.6371	0.0853
midfemale	-0.7868	0.0310	-0.7692	0.0340	-0.7866	0.0316	-0.7844	0.0319
zeroncd	0.2573	0.0050	0.2547	0.0060	0.2617	0.0051	0.2630	0.0050
vage	-0.0438	0.0210	-0.0442	0.0210	-0.0436	0.0227	-0.0438	0.0224
vbrand1	0.5493	<.0001	0.5473	<.0001	0.5481	<.0001	0.5478	<.0001
vbrand2	0.1831	0.0740	0.1854	0.0710	0.1813	0.0777	0.1827	0.0755
LogLik	-1498.40		-1497.78		-1498.00		-1497.50	
AIC	3012.81		3013.57		3016.00		3017.00	
BIC	3056.41		3062.62		3070.50		3077.00	

Estimates of copula model with various dependence structures

	AR(1)		Exchangeable		Toeplitz(2)	
Parameter	Estimate	StdErr	Estimate	StdErr	Estimate	StdErr
intercept	-1.8028	0.0307	-1.8422	0.0353	-1.7630	0.0284
young	0.6529	0.0557	0.7130	0.0667	0.6526	0.0631
midfemale	-0.6956	0.0588	-0.6786	0.0670	-0.7132	0.0596
zeroncd	0.2584	0.0198	0.2214	0.0172	0.2358	0.0176
vage	-0.0411	0.0051	-0.0422	0.0056	-0.0453	0.0042
vbrand1	0.5286	0.0239	0.5407	0.0275	0.4962	0.0250
vbrand2	0.1603	0.0166	0.1752	0.0229	0.1318	0.0198
ϕ	2.9465	0.1024	2.9395	0.1130	2.9097	0.1346
ρ_1	0.1216	0.0028	0.1152	0.0027	0.1175	0.0025
ρ_2					0.0914	0.0052
LogLik	-1473.25		-1454.04		-1468.74	
AIC	2964.49		2926.08		2957.49	
BIC	3013.55		2975.13		3011.99	

Model validation

- Copula validation
  - The specification of the copula is validated using *t*-plot method as suggested in Sun et al. (2008) and Shi (2010).
  - In a good fit, we would expect to see a linear relationship in the *t*-plot.

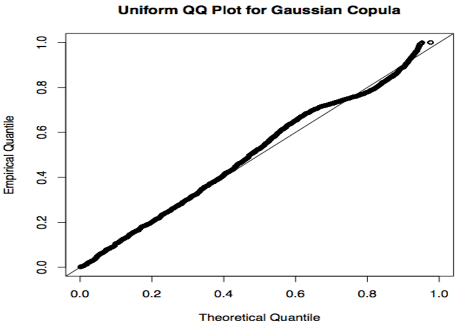
- Out-of-sample validation: based on predictive distribution calculated using

$$
\begin{aligned}
& f_{iT+1}(n_{iT+1}|n_{i1},\ldots,n_{iT}) \\
& \quad = \Pr(N_{iT+1}=n_{iT+1}|N_{i1}=n_{i1},\ldots,N_{iT}=n_{iT}) \\
& \quad = \frac{\mathrm{E}_{\boldsymbol{U}_i}\left[c(F_{i1}^*(n_{i1}-U_{i1}),\ldots,F_{iT}^*(n_{iT}-U_{iT}),F_{iT+1}^*(n_{iT+1}-U_{iT+1});\boldsymbol{\theta})\prod_{t=1}^{T+1}f_{it}^*(n_{it}-U_{it})\right]}{\mathrm{E}_{\boldsymbol{U}_i}\left[c(F_{i1}^*(n_{i1}-U_{i1}),\ldots,F_{iT}^*(n_{iT}-U_{iT});\boldsymbol{\theta})\prod_{t=1}^{T}f_{it}^*(n_{it}-U_{it})\right]}.
\end{aligned}
$$

- Performance measures used:
  - $\text{LogLik} = \sum_{i=1}^{M} \log\left(f_{iT+1}(n_{iT+1}|n_{i1},\cdots,n_{iT})\right)$
  - $\text{MSPE} = \sum_{i=1}^{M}\left[n_{iT+1}-\mathrm{E}(N_{iT+1}|N_{i1}=n_{i1},\cdots,N_{iT}=n_{iT})\right]^2$
  - $\text{MAPE} = \sum_{i=1}^{M}\left|n_{iT+1}-\mathrm{E}(N_{iT+1}|N_{i1}=n_{i1},\cdots,N_{iT}=n_{iT})\right|$

Results of model validation

t-plot



Out-of-sample validation

	Standard Model		Copula Model		
	RE-Poisson	RE-NegBin	AR(1)	Exchangeable	Toeplitz(2)
LogLik	-177.786	-177.782	-168.037	-162.717	-165.932
MSPE	0.107	0.107	0.108	0.105	0.110
MAPE	0.213	0.213	0.197	0.186	0.192

Concluding remarks

- We examined an alternative way to model longitudinal count based on copulas:
  - employed a continuous extension with jitters
  - method preserves the concordance-based association measures
- The approach avoids the criticisms often made with using copulas directly on multivariate discrete observations.
- For empirical demonstration, we applied the approach to a dataset from a Singapore auto insurer. Our findings show:
  - better fit when compared with random-effect specifications
  - validated the copula specification based on *t*-plot and its performance based on hold-out observations
- Our contributions to the literature: (1) application to insurance data, and (2) application to longitudinal count data.

Selected reference

- Denuit, M. and P. Lambert (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1), 40-57.
- Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *ASTIN Bulletin*, 37(2), 475-515.
- Hausman, J., B. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-r&d relationship. *Econometrica*, 52(4), 909-938.
- Madsen, L. and Y. Fang (2010). Joint regression analysis for discrete longitudinal data. *Biometrics*. Early view.
- Song, P., M. Li, and Y. Yuan (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, 65(1), 60-68.
- Sun, J., E. W. Frees, and M. A. Rosenberg (2008). Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics*, 42(2),817-830.