Special cases of longitudinal data: discrete and multivariate

Emiliano A. Valdez, Ph.D., F.S.A.
University of Connecticut

joint work with P. Shi*, P.H. Katuwandeniyage**

* Northern Illinois University
** University of Connecticut

Universidad de Barcelona, España
16-18 July 2012

Two papers

- Shi and Valdez (2008), Longitudinal modeling of claim counts using jitters, *Scandinavian Actuarial Journal*, accepted for publication.
- Katuwandeniyage and Valdez (2012), Multivariate longitudinal data analysis for actuarial applications, work in progress.

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 for claim count

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

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

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

- 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 (2011).
  - 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

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

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

Construction of the t-plot

The null hypothesis of a t-plot is that a sample comes from an elliptical multivariate distribution. Such hypothesis could be tested according to the procedure below:

(a) Transform the claim counts to variables on $(0,1)$ by $\hat{\iota}_{it} = F^*_{it}(n_{it} - u_{it}; \hat{\boldsymbol{\beta}}, \hat{\psi})$ for $t = 1, \cdots, T_i$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ are the maximum likelihood estimates. Under the null hypothesis, $\boldsymbol{\iota}_i = (\iota_{i1}, \cdots, \iota_{iT_i})$ is a realization of the hypothesized elliptical copula.

(b) Compute the quantiles of $\hat{\iota}_{it}$ by $\hat{\zeta}_{it} = H_t^{-1}(\hat{\iota}_{it})$ for $t = 1, \cdots, T_i$, where H_t denotes the marginal distribution associated with the elliptical copula. Thus, if the copula is well-specified, $\hat{\boldsymbol{\zeta}}_i = (\hat{\zeta}_{i1}, \cdots, \hat{\zeta}_{iT_i})'$ follows the multivariate elliptical distribution of $E_{T_i}(\mathbf{0}, \boldsymbol{\Sigma}, g_{T_i})$.

- continued

(c) Calculate vector $\hat{\boldsymbol{\zeta}}_i^* = (\hat{\zeta}_{i1}^*, \cdots, \hat{\zeta}_{iT_i}^*)' = \widehat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\zeta}}_i$, and construct the t statistic for policyholder i:

$$t_i(\hat{\boldsymbol{\zeta}}_i^*) = \frac{\sqrt{T_i}\bar{\hat{\zeta}}_i^*}{\sqrt{(T_i-1)^{-1}\sum_{t=1}^{T_i}(\hat{\zeta}_{it}^* - \bar{\hat{\zeta}}_i^*)^2}},$$

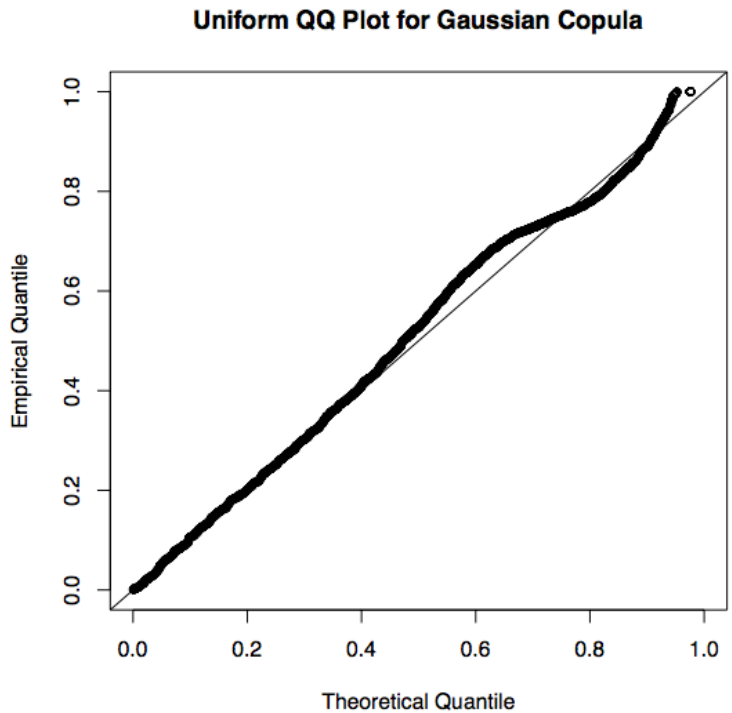
with $\widehat{\boldsymbol{\Sigma}}$ the maximum likelihood estimator of $\boldsymbol{\Sigma}$ and $\bar{\hat{\zeta}}_i^* = T_i^{-1}\sum_{t=1}^{T_i}\hat{\zeta}_{it}^*$. Thus $t_i(\hat{\boldsymbol{\zeta}}_i^*)$ should be from a standard t distribution with $T_i - 1$ degrees of freedom.

(d) Repeat steps (a) - (c) for $i = 1, \cdots, I$, and calculate the t statistics $t_i(\hat{\boldsymbol{\zeta}}_i^*)$ for all policyholders in the sample. Define the transformed variable $\varsigma_i = G_{T_i-1}(t_i(\hat{\boldsymbol{\zeta}}_i^*))$, where G_{T_i-1} denotes the cdf of a t distribution with $T_i - 1$ degrees of freedom.

If the copula captures the dependence structure properly, $\boldsymbol{\varsigma} = (\varsigma_1, \cdots, \varsigma_n)'$ should be a random sample from a uniform (0,1). This can be easily verified using standard graphical tools or goodness-of-fit tests.

Results of model validation

t-plot



Out-of-sample validation

	Standard Models		Copula Models		
	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.

Multivariate longitudinal data analysis for actuarial applications

Motivation

- Unlike univariate longitudinal studies, multivariate longitudinal analysis allows for understanding the joint evolution of multivariate responses over a period of time.
- There is increasing interest on multivariate longitudinal data analysis, especially in biostatistics, where longitudinal analysis is quite common.
- However, we found that overall, there is lack of attention devoted to multivariate longitudinal data analysis.
- We are looking into the potential of the use of this type of analysis in insurance and actuarial science.

Introduction

- In the presence of repeated observations over time, the natural approach for data analysis is univariate longitudinal model. (e.g. Shi and Frees, 2010 and Frees et al, 1999)
- Repeated observations over time for many responses require multivariate longitudinal framework.
- Model accuracy can be improved by incorporating dependency among multiple responses.
- Response variables are typically assumed to have multivariate normal distribution.
- Multivariate longitudinal data analysis is becoming a popular tool in data analysis.
- There is a developing interest on multivariate longitudinal analysis in actuarial context (e.g Shi, 2011).

Some literature

- Frees, E.W. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press, Cambridge.

- Seemingly unrelated regressions (SUR) approach
  - Rochon, J. (1996) Analyzing bivariate repeated measures for discrete and continuous outcome variable. *Biometrics* 52: 740-50.
- The random effects approach
  - Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association* 77: 190-195.
  - Shah, A., N.M. Laird, and D. Schoenfeld (1997). A random effects model with multiple characteristics with possibly missing data. *Journal of the American Statistical Association* 92: 775-79.
  - Fieuws, S. and G. Verbeke (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics* 62: 424-431.
- Copula approach
  - Lambert, P. and F. Vandenhende (2002). A copula based model for multivariate non normal longitudinal data: analysis of a dose titration safety study on a new antidepressant. *Statistics in Medicine* 21: 3197-3217.
  - Shi, P. (2011). Multivariate longitudinal modeling of insurance company expenses. *Insurance: Mathematics and Economics.* In Press.

Our contribution

- Methodology
  - We propose the use of a random effects model to capture dynamic dependency and heterogeneity, and a copula function to incorporate dependency among the response variables.
- Multivariate longitudinal analysis for actuarial applications
  - We intend to explore actuarial-related problems within multivariate longitudinal context, and apply our proposed methodology.
- NOTE: Our results are very preliminary at this stage.

Notation

Suppose we have a set of observations on n subjects collected over T time periods for a set of m response variables. Let $y_{it,k}$ denote the observation from i^{th} individual in t^{th} time period on k^{th} response. Hence, for a given subject, the matrix

$$\mathbf{Y_i} = \begin{bmatrix} y_{i1,1} & y_{i2,1} & . & . & . & y_{iT,1} \\ y_{i1,2} & y_{i2,2} & . & . & . & y_{iT,2} \\ . & . & . & . & . & . \\ y_{i1,m} & y_{i2,m} & . & . & . & y_{iT,m} \end{bmatrix} \text{ where } i = 1, 2, ...n$$

represents observations over T time periods corresponding to m number of response variables.

By letting $\mathbf{y_{it}} = (y_{it,1}, y_{it,2}, \ldots, y_{it,m})'$ for $t = 1, 2, \ldots, T$, we can express $\mathbf{Y_i} = (\mathbf{y_{i1}}, \mathbf{y_{i2}}, \ldots, \mathbf{y_{iT}})$.

Notation - continued

- Collected q set of covariates associated with each observed subject i can be represented as

$$\mathbf{X_{it}} = \begin{bmatrix} x_{it1,1} & x_{it2,1} & . & . & . & x_{itq,1} \\ x_{it1,2} & x_{it2,2} & . & . & . & x_{itq,2} \\ . & . & . & . & . & . \\ x_{it1,m} & x_{it2,m} & . & . & . & x_{itq,m} \end{bmatrix}$$

where $i = 1, 2, ...n$ and $t = 1, 2, \ldots, T$

If $\mathbf{x_{it,k}} = (x_{it1,k}, x_{it2,k}, \ldots, x_{itp,k})$ for $k = 1, 2, ...m$, we can similarly express $\mathbf{x_{it}} = (\mathbf{x_{it,1}}, \mathbf{x_{it,2}}, \ldots, \mathbf{x_{it,m}})$.

- We use $\alpha_{\mathbf{ik}}$ to represent the random effects component corresponding to the i^{th} subject from the k^{th} response variable.
- $G(\alpha_{ik})$ represents the pre-specified distribution function of random effect α_{ik}.

Key features of our approach

- Obviously, the extension from univariate to multivariate longitudinal analysis.
- Types of dependencies captured:
  - the dependence structure of the response using copulas - provides flexibility
  - the intertemporal dependence within subjects and unobservable subject-specific heterogeneity captured through the random effects component - provides tractability
- The marginal distribution models:
  - any family of flexible enough distributions can be used
  - choose family so that covariate information can be easily incorporated
- Other key features worth noting:
  - the parametric model specification provides flexibility for inference e.g. MLE for estimation
  - model construction can accommodate both balanced and unbalanced data - an important feature for longitudinal data

Some model assumptions

While several of these model assumptions are simplified during the initial stages of our investigation, many can be modified to make the model more reasonable, practicable and flexible for several applications:

- The observations $\{\mathbf{Y_i}\}$ are independent for a given time t and response k.
- Each response variable over time is assumed to belong to the same class of distributions.
- The covariates $\{\mathbf{x_{it}}\}$ are non-stochastic variables.
- The random effects components $\{\alpha_{ik}\}$ are independent and identically distributed.
- Random effects and covariates are independent.
- The same family of copula functions is applicable over time.

Copula function

For arbitrary m uniform random variables on the unit interval, copula function, C, can be uniquely defined as

$$C(u_1, u_2, \ldots, u_m) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_m \leq u_m).$$

- Joint distribution:

$$F(y_1, y_2, \ldots, y_m) = C(F_1(y_1), \ldots, F_m(y_m)),$$

where $F_k(y_k)$ are marginal distribution functions.

- Joint density:

$$f(y_1, y_2, \ldots, y_m) = c(F_1(y_1), ..., F_m(y_m)) \prod_{k=1}^{m} f_k(y_k),$$

where $f_k(y_k)$ are marginal density functions and c is the density associated with copula C.

Multivariate joint distribution

Suppose we observe m number of response variables over T time periods for n subjects. Observed data for subject i is

$$\{(y_{i1,1}, y_{i1,2}, \ldots, y_{i1,m}), \ldots, (y_{iT,1}, y_{iT,2}, \ldots, y_{iT,m})\}$$

so that

$$\mathbf{Y_{it}} = (y_{it,1}, y_{it,2}, \ldots, y_{it,m}) \text{ for } i = 1, 2, \ldots, n \text{ and } t = 1, 2, \ldots, T$$

is the i^{th} observation in the t^{th} time period corresponding to m responses. The joint distribution of m response variables over time can be expressed as

$$H(\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}}) = \mathbf{P}(\mathbf{Y_{i1}} \leq \mathbf{y_{i1}}, \ldots, \mathbf{Y_{iT}} \leq \mathbf{y_{iT}}).$$

If $\{\alpha_{ik}\}$ represent random effects with respect to the k^{th} response variable, conditional joint distribution at time t is

$$H(\mathbf{y_{it}}|\alpha_{i1}, \ldots, \alpha_{im}) = C(F(y_{it,1}|\alpha_{i1}), \ldots, F(y_{it,m}|\alpha_{im})).$$

- continued

Conditional joint density at time t:

$$h(\mathbf{y_{it}}|\alpha_{i1},\ldots,\alpha_{im}) = c(F(y_{it,1}|\alpha_{i1}),\ldots,F(y_{it,m}|\alpha_{im}))\prod_{k=1}^{m} f(y_{it,k}|\alpha_{ik})$$

where $F(y_{it,k}|\alpha_{ik})$ denotes the distribution function of k^{th} response variable at time t. If ω represents the set of parameters in the model, the likelihood of the i^{th} subject is given by

$$L(\boldsymbol{\omega}|(\mathbf{y_{i1}},\mathbf{y_{i2}},\ldots,\mathbf{y_{iT}})) = h(\mathbf{y_{i1}},\mathbf{y_{i2}},\ldots,\mathbf{y_{iT}}|\boldsymbol{\omega}).$$

We can write

$$h(\mathbf{y_{i1}},\mathbf{y_{i2}},\ldots,\mathbf{y_{iT}}|\boldsymbol{\omega}) = \int_{\alpha_{i1}}\cdots\int_{\alpha_{im}} h(\mathbf{y_{i1}},\ldots,\mathbf{y_{iT}}|\alpha_{i1},\ldots,\alpha_{im})\, dG(\alpha_{i1})\cdots dG(\alpha_{im})$$

Under independence over time for a given random effect:

$$h(\mathbf{y_{i1}},\ldots,\mathbf{y_{iT}}|\alpha_{i1},\ldots,\alpha_{im}) = \prod_{t=1}^{T} h(\mathbf{y_{it}}|\alpha_{i1},\ldots,\alpha_{im})$$

- continued

$$= \int_{\alpha_{i1}} \cdots \int_{\alpha_{im}} \prod_{t=1}^{T} h(\mathbf{y_{it}}|\alpha_{i1}, \ldots, \alpha_{im}) dG(\alpha_{i1}) \cdots dG(\alpha_{im})$$

and from the previous slides, we have

$$= \int_{\alpha_{i1}} \cdots \int_{\alpha_{im}} \prod_{t=1}^{T} c(F(y_{it,1}|\alpha_{i1}), \ldots, F(y_{it,m}|\alpha_{im})) \prod_{k=1}^{m} f(y_{it,k}|\alpha_{ik}) dG(\alpha_{i1}) \cdots dG(\alpha_{im})$$

Then, we can write the log likelihood function as

$$\sum_i \log \Big\{ \int_{\alpha_{i1}} \cdots \int_{\alpha_{im}} \prod_{t=1}^{T} \prod_{k=1}^{m} c(F(y_{it,1}|\alpha_1), \ldots, F(y_{it,m}|\alpha_m)) \times f(y_{it,k}|\alpha_{ik}) dG(\alpha_{i1}) \cdots dG(\alpha_{im}) \Big\}$$

Choice for the marginals: the class of GB2

The model specification is flexible enough to accommodate any marginals; however, for our purposes, we chose the class of GB2 distributions. For $Y \sim \text{GB2}(a, b, p, q)$ with $a \neq 0, b, p, q > 0$:

- Density function:

$$f_y(y) = \frac{|a|\, y^{ap-1} b^{aq}}{B(p,q)(b^a + y^a)^{(p+q)}}$$

  where $B(\cdot, \cdot)$ is the usual Beta function.

- Distribution function:

$$F_y(y) = B\left(\frac{(y/b)^a}{1 + (y/b)^a}; p, q\right)$$

  where $B(\cdot; \cdot, \cdot)$ is the incomplete Beta function.

- Mean:

$$\text{E}(Y) = b\, \frac{B(p + 1/a, q - 1/a)}{B(p,q)}.$$

GB2 regression through the scale parameter

Suppose $\mathbf{x}$ is a vector of known covariates:

- We have: $Y|\mathbf{x} \sim \text{GB2}(a, b(\mathbf{x}), p, q)$, where

$$b(\mathbf{x}) = \alpha + \beta' \mathbf{x}$$

- Define residuals $\varepsilon_i = Y_i e^{-(\alpha_i + \beta' \mathbf{x}_i)}$ so that

$$\log Y_i = \log \alpha_i + \beta' \mathbf{x}_i + \log \varepsilon_i$$

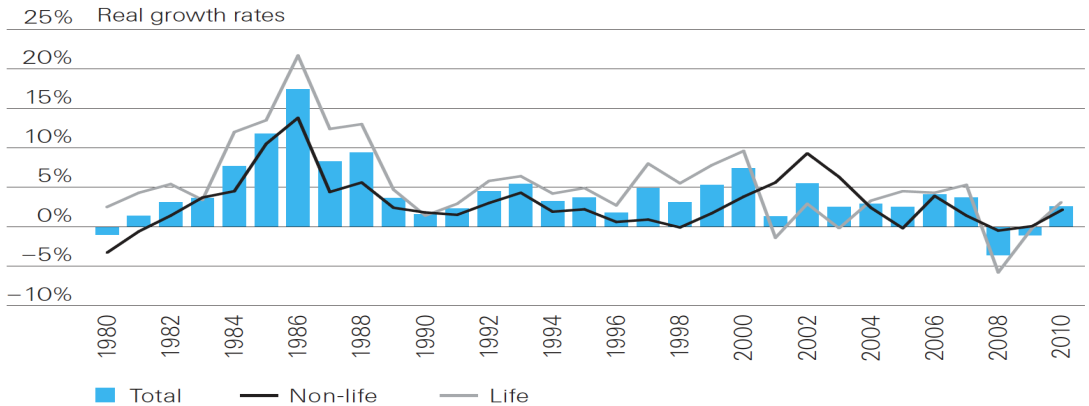
  where $\varepsilon_i \sim \text{GB2}(a, 1, p, q))$.

- PP plots can then be used for diagnostics.

Some empirical work on GB2

- Income or wealth distributions
  - McDonald (1984)
  - Butler and McDonald (1989)
  - McDonald and Mantrala (1993, 1995)
  - Bordley and McDonald (1993)
  - McDonald and Xu (1995)
- Unemployment duration
  - McDonald and Butler (1987)
- Insurance loss
  - Cummins, Dionne, McDonald and Pritchett (1990) - fire losses
  - appeared in Insurance: Mathematics and Economics

Case study - global insurance demand



Source: Swiss Re Economic Research & Consulting

Response variables that can be used for insurance demand:

- Insurance density: Premiums per capita
- Insurance penetration: Ratio of insurance premiums to GDP
- Insurance in force: Outstanding face amount plus dividend

Some common covariates that have appeared in the literature:

- Education; Income / GDP growth; Inflation
- Urbanization
- Dependency ratio
- Death ratio / Life expectancy

About the data set

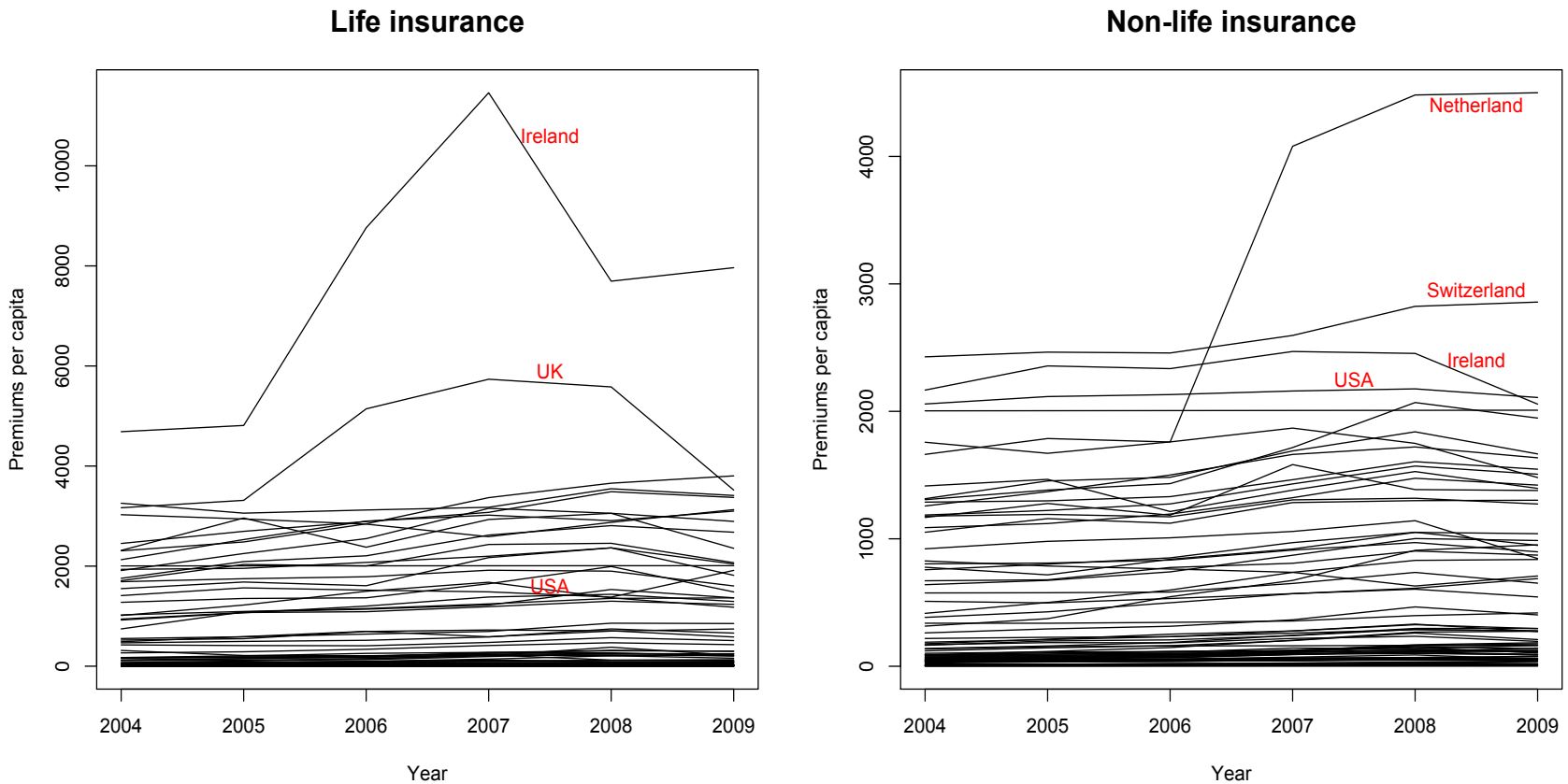
Data set

- 2 responses: life and non-life insurance
- 5 predictor variables
- 75 countries
- 6 years data (from year 2004 to year 2009)

Variables in the model

Dependent variables	
Life density	Premiums per capita in life insurance
Non-life density	Premiums per capita in non-life insurance
Independent variables	
GDP per capita	Ratio of gross domestic product (current US dollars) to total population
Religious	Percentage of Muslim population
Urbanization	Percentage of urban population to total population
Death rate	Percentage of death
Dependency ratio	Ratio of population over 65 to working population

Multiple time series plot



Some summary statistics

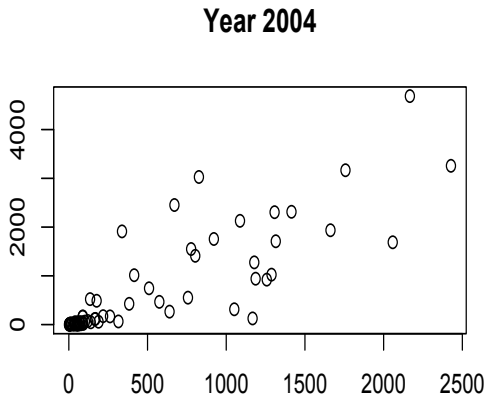
Summary statistics of variables in year 2004 to 2009:

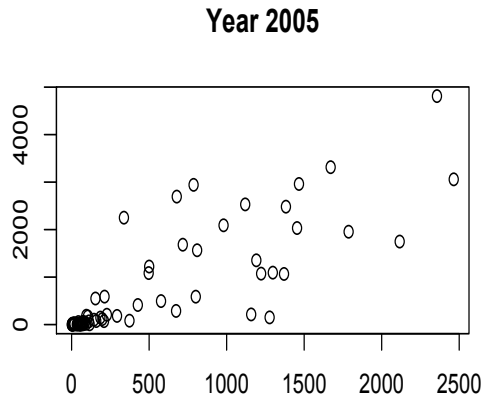
Variable	Minimum	Maximum	Mean	Correlation with Life insurance	Correlation with Non-life insurance
Life insurance	(0.49, 1.28)	(4686.8, 11460)	(614.23, 886)	1.00	(0.66, 0.84)
Non-life insurance	(0.74, 1.26)	(2427.6, 4499.6)	(445.3, 612.1)	(0.66, 0.84)	1.00
GDP per capita	(375.2, 550.9)	(56311.5, 94567.9)	(14937.3, 21791.7)	(0.67, 0.82)	(0.84, 0.91)
Death rate	(1.5, 1.52)	(16.17, 17.11)	(7.8, 8)	(0.03, 0.06)	(0.04, 0.07)
Urbanization	(11.92, 13.56)	(100,100)	(65.37, 66.76)	(0.28, 0.36)	(0.41, 0.44)
Religious	(0.01,0.01)	(99.61, 99.61)	(21.35, 21.35)	(-0.30, -0.27)	(-0.32, -0.27)
Dependency ratio	(1.25, 1.39)	(29.31, 33.92)	(15.11, 15.78)	(0.39, 0.53)	(0.52, 0.57)

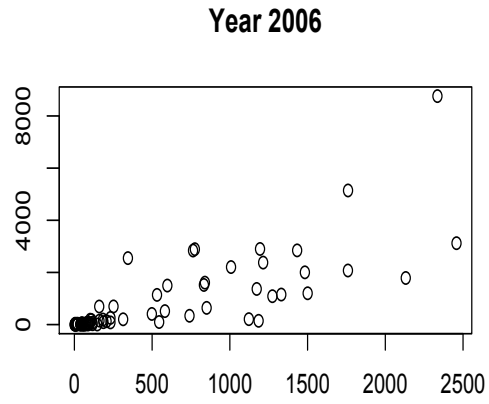
Correlation matrix of covariates in year 2004 to 2009:

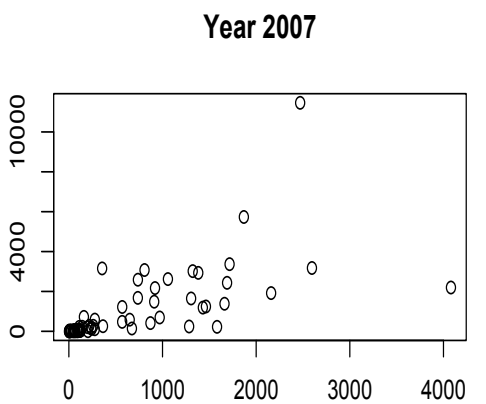
	GDP per capita	Death rate	Urbanization	Religious	Dependency ratio
GDP per capita	1.00				
Death rate	(0.01, 0.03)	1.00			
Urbanization	(0.49, 0.52)	(-0.15, -0.14)	1.00		
Religious	(-0.30, -0.27)	(-0.38, -0.35)	(-0.15, -0.14)	1.00	
Dependency ratio	(0.58, 0.62)	(0.53, 0.54)	(0.32, 0.34)	(-0.53, -0.53)	1.00

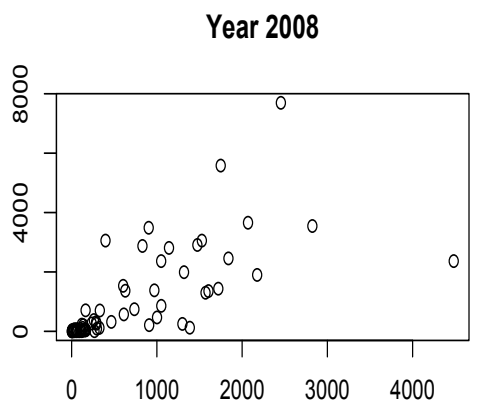
Scatter plots of the two response variables

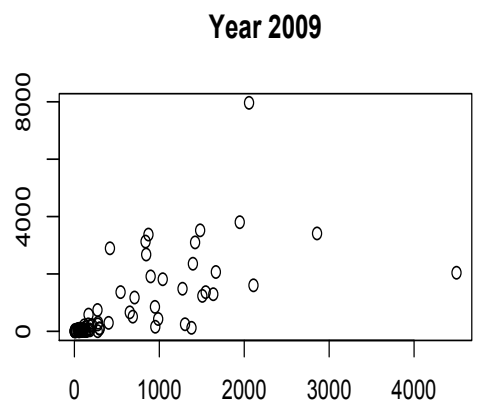




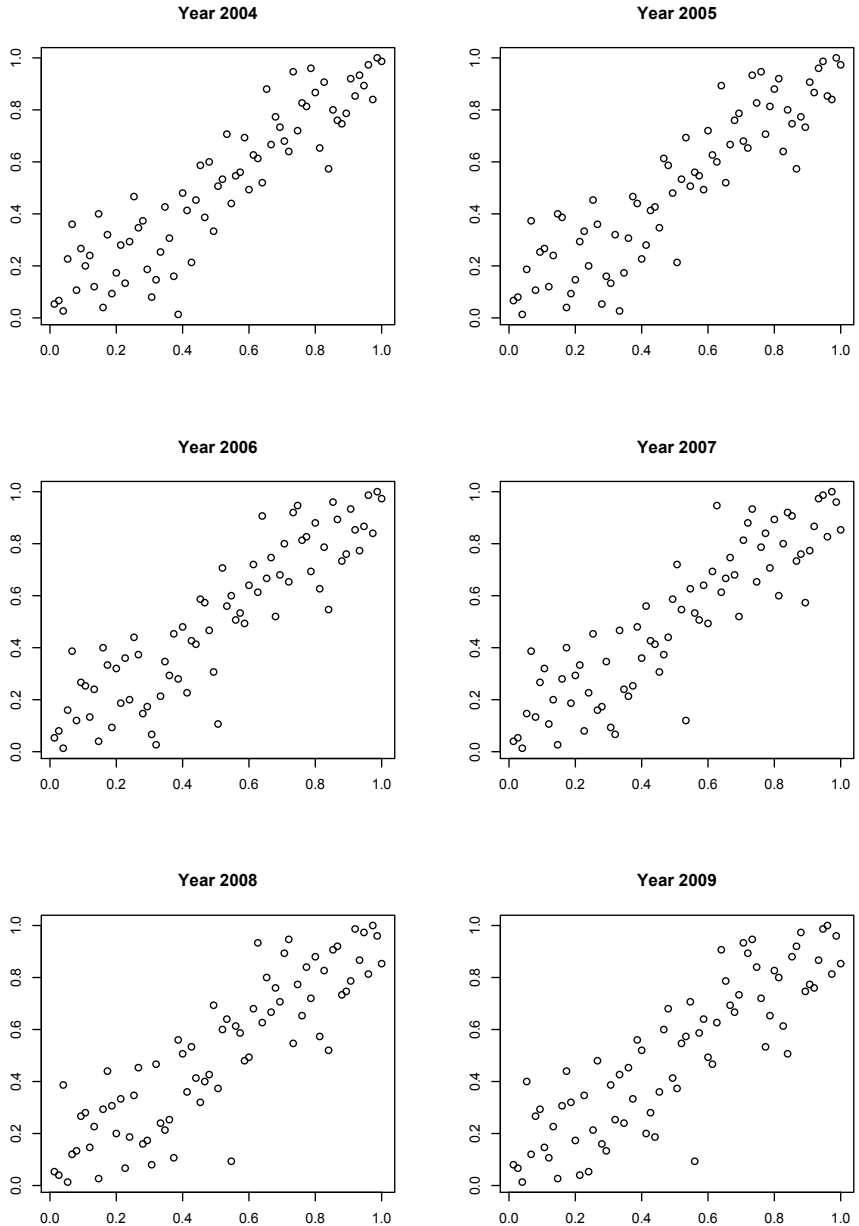




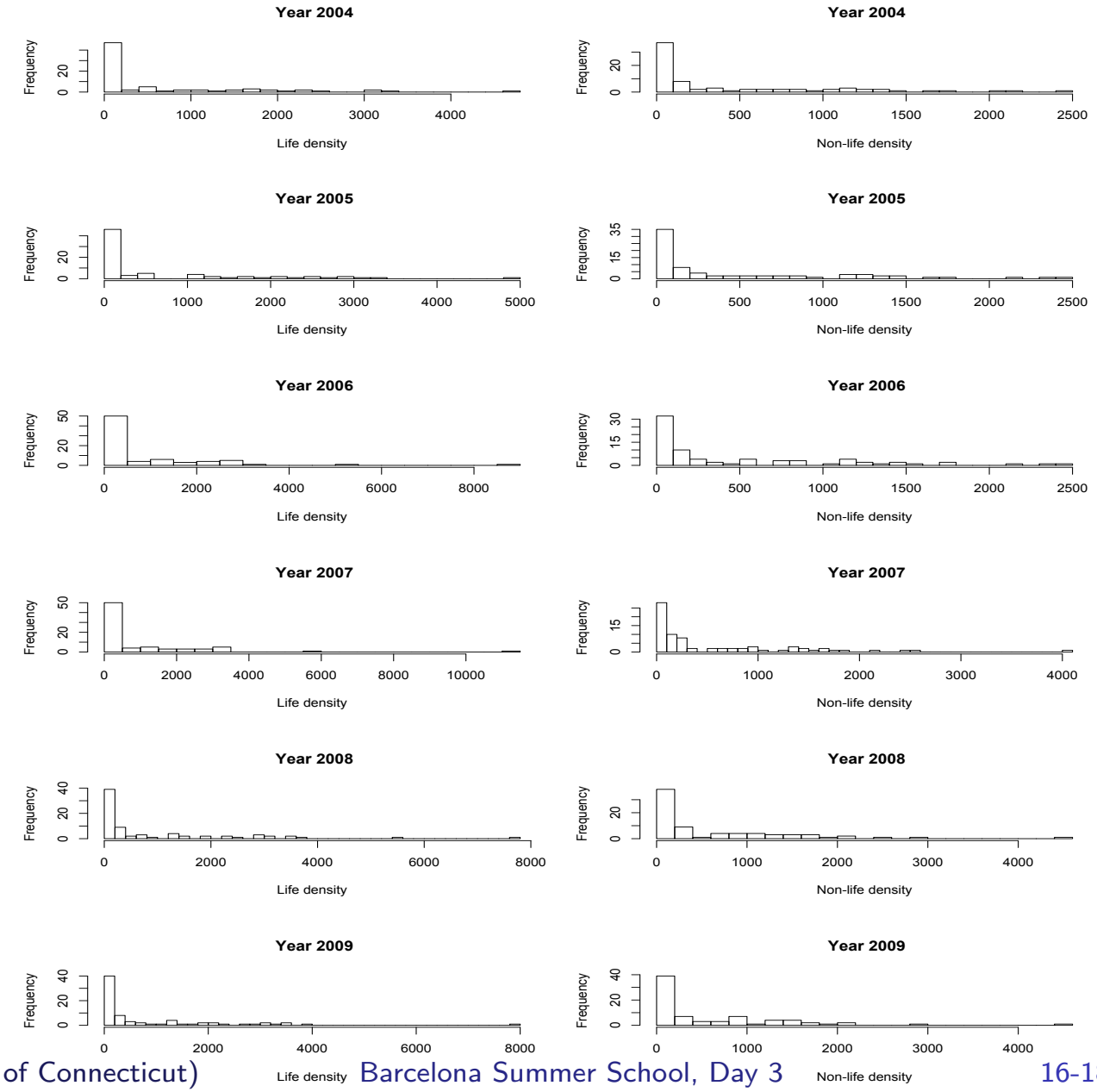




Scatter plots of the ranked response variables



Histograms of two responses from year 2004 to 2009



Model calibration

- Marginals: GB2 with regression on the scale parameter
- Gaussian copula:

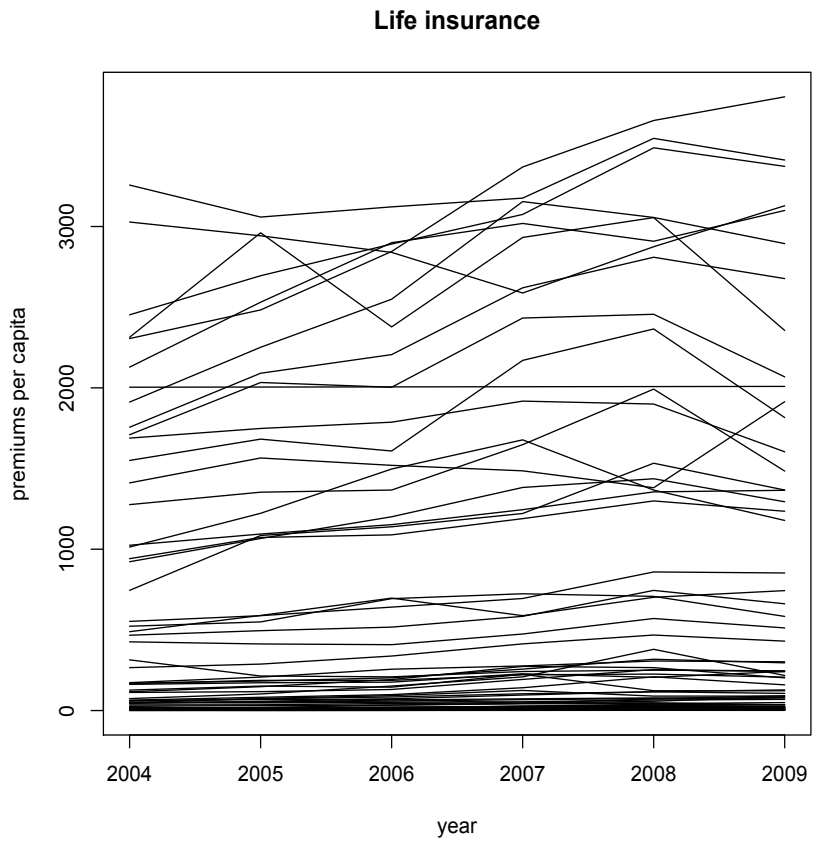
$$C(u_1, u_2; \rho) = \mathbf{\Phi}_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$$

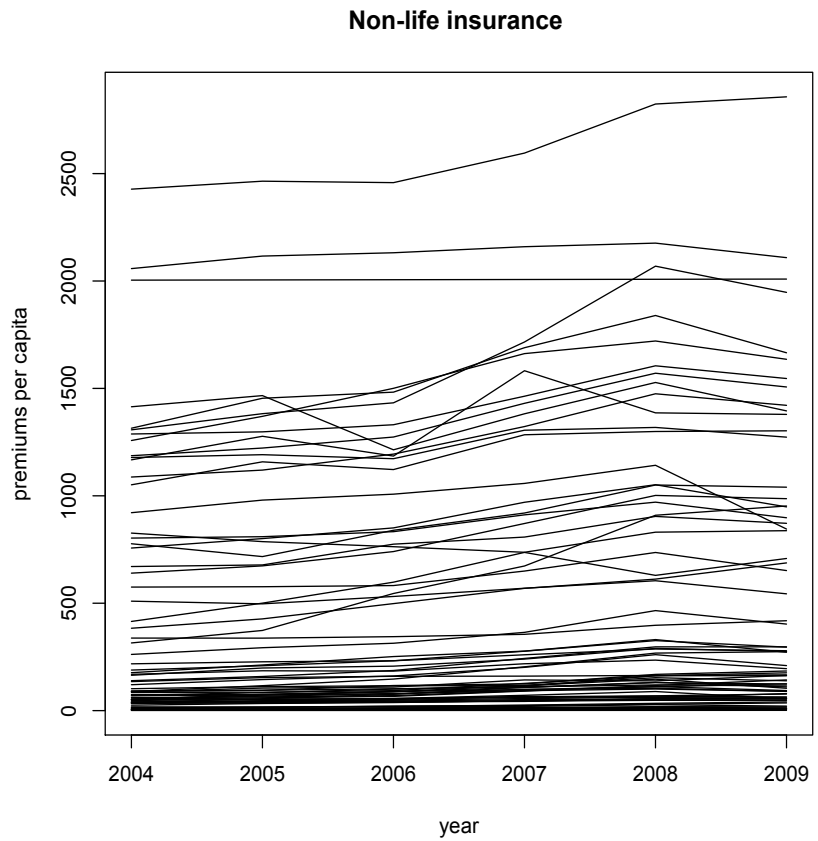
- Natural assumption for random effect for the k^{th} response:

$$\alpha_{ik} \sim N\left(0, \sigma_k^2\right)$$

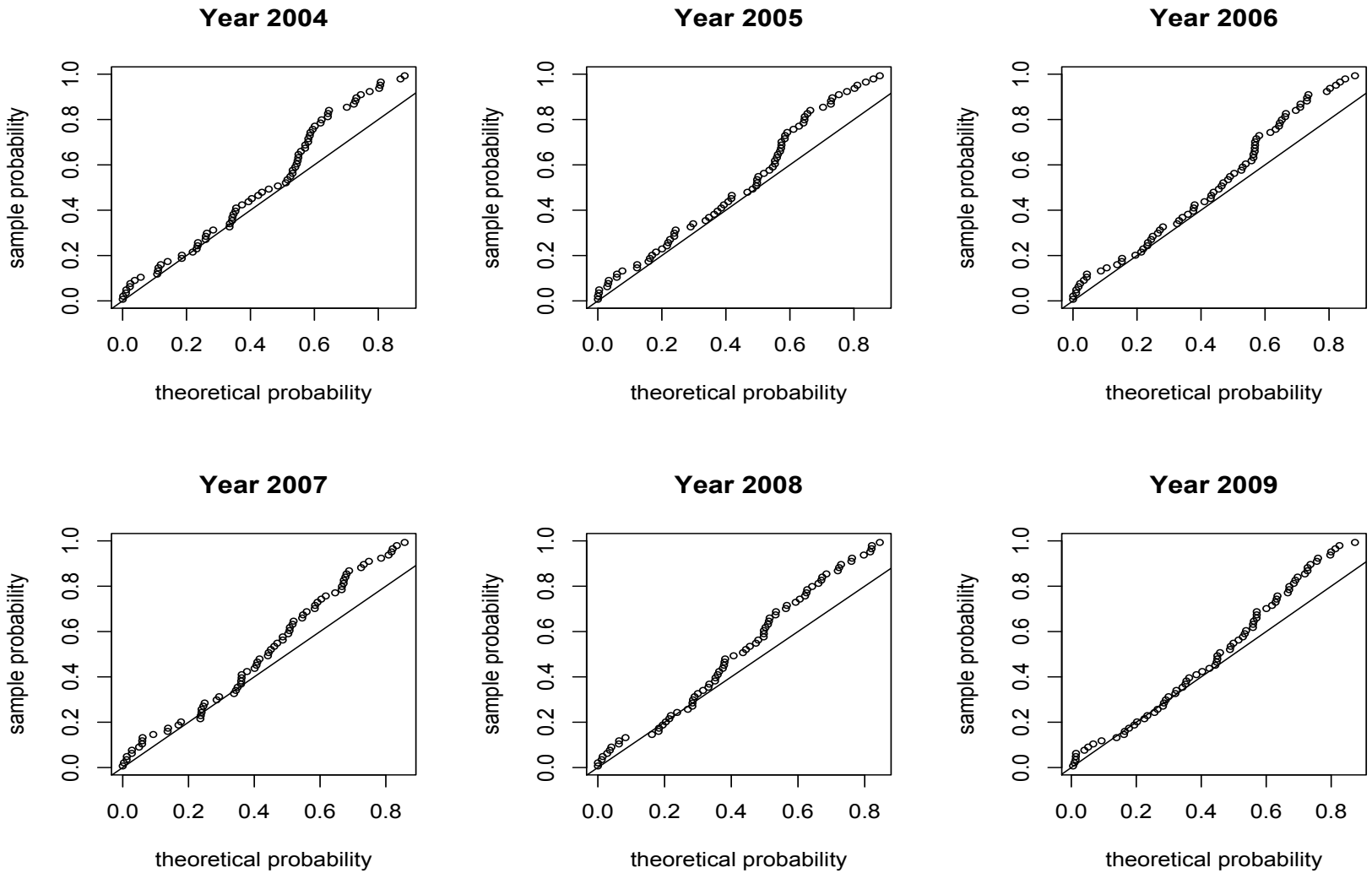
Multiple time series plot

After removing Ireland, Netherlands and the UK in the dataset:

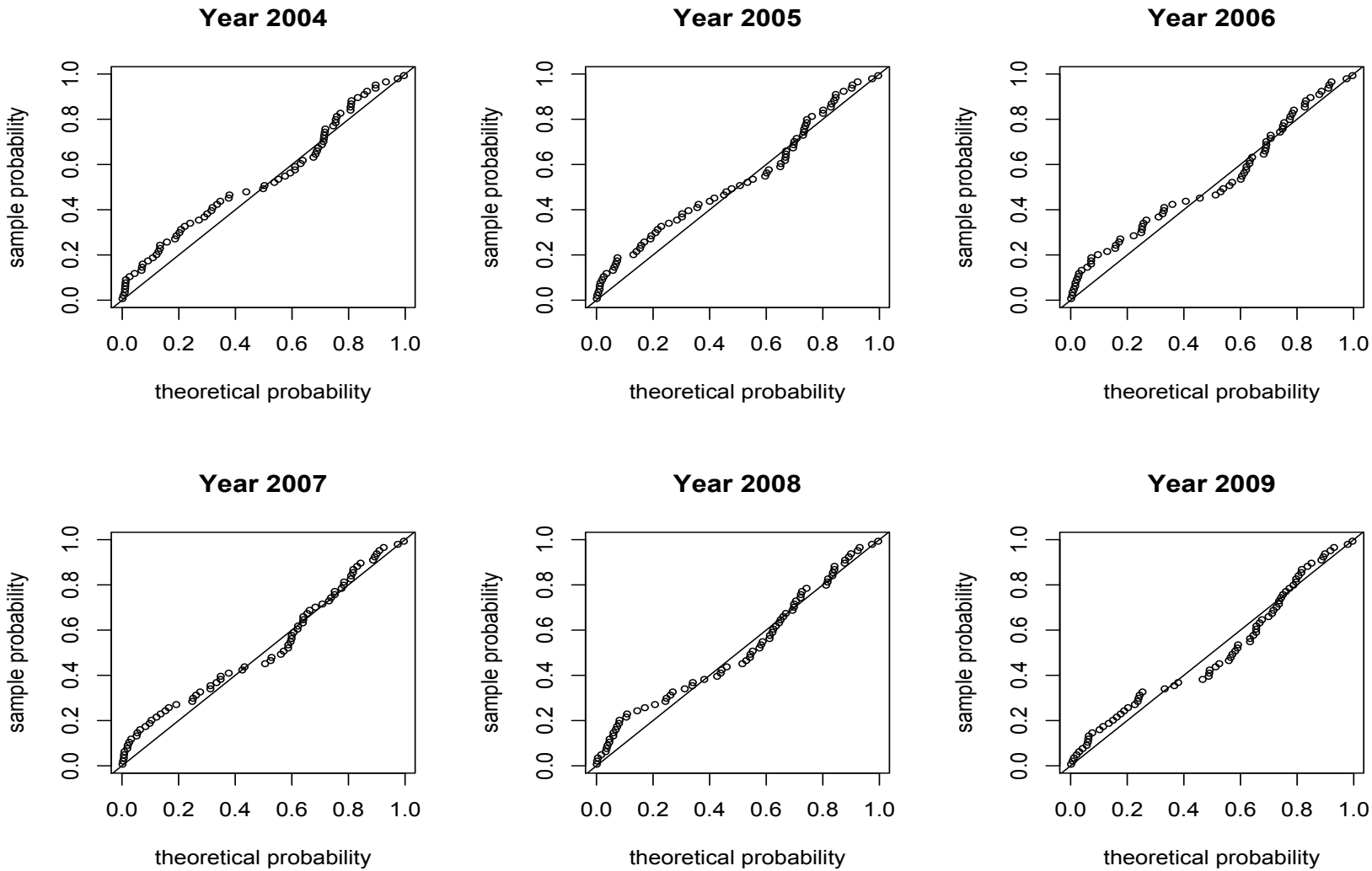




PP plots of the residuals for marginal diagnostics: Life Insurance



PP plots of the residuals for marginal diagnostics: Non-life Insurance



Model estimates

	Univariate fitted model for insurance demand					
	Life insurance density			Non-life insurance density		
Parameter	**Estimate**	**Std Error**	**p-val**	**Estimate**	**Std Error**	**p-val**
Covariates						
GDP per capita	0.0001	0.0000	0.0000	0.0001	0.0000	0.0000
Religious	-0.0231	0.0040	0.0000	-0.0085	0.0023	0.0000
Urbanization	0.0279	0.0061	0.0000	0.0567	0.0022	0.0000
Death rate	0.0035	0.0333	0.9164			
Dependency ratio (old)	-0.0440	0.0297	0.1390			
GB2 Marginals						
a	1.0427	0.0611	0.0000	2.5636	0.1397	0.0000
p	3.7321	0.5371	0.0000	1.3957	0.1356	0.0000
q	0.5081	0.0330	0.0000	0.5369	0.0364	0.0000
Random effect						
$Sigma_\alpha$	0.8507	0.1088	0.0000	0.6471	0.0535	0.0000

Gaussian copula:

Parameter	Estimate	Std Error	p-val
ρ	0.7375	0.0376	0.0000

Additional work intended

- Implementing diagnostic tests for model validation
- Handling unbalanced and missing data.
- Identifying more actuarial related problems within a multivariate longitudinal framework.
  - e.g. there is a rapid development in loss reserving using multiple loss triangle.
- Alternative approach:

  Use Multivariate generalized liner models for response in each time period and use copula to capture the inter-temporal dependence.
- (Possible) handling discrete response variables incorporating jitters.

Dependent loss triangles

In property and casualty, different lines of business and their risks are associated with each other frequently.

Hence, the aggregate risk of the portfolio depends on the association between different lines of business.

Understanding these associations is important for pricing, risk management, capital allocation and loss reserving, to name a few.

The classical approach, commonly used together with several variations, is to use univariate chain ladder method.

Ajne (1994) showed that simple additivity of loss triangles between lines of business does not provide similar results as aggregated loss triangle under the chain ladder approach. This indicates the importance of modeling loss triangles that capture their dependencies.

- continued

According to Holmberg (1994) and Schmidt (2006), there are are possible dependencies that can be observed from a portfolio of loss triangles:

a the dependency within accident years,

b the dependency between accident years, and

c the dependency between different line of business

Some work that have appeared in the literature:

- Mack (1993): extended distribution free methods like chain ladder to multivariate stochastic reserving
- Schmidt (2006): proposed multivariate chain ladder method where the dependence structure was incorporated into parameter estimates
- Zhang (2010): employed seemingly unrelated regression models
- Shi (2011) and de Jong (2011): explored the flexibility of copula functions, accommodating various correlation structures

- continued

It is well known that in the presence of m portfolios of risks with the same number of development years and accident years, n, incremental losses may be represented by a loss triangle.

We intend to rearrange the classical loss triangle to facilitate a longitudinal framework for claims from each line of business.

	Development year						
Calendar year	0	1	. . .	k	. . .	$n-1$	n
0	$\mathbf{S_{0,0}}$						
1	$\mathbf{S_{1,0}}$	$\mathbf{S_{0,1}}$					
$\vdots$	$\vdots$	$\vdots$					
k	$\mathbf{S_{k,0}}$	$\mathbf{S_{k-1,1}}$	. . .	$\mathbf{S_{0,k}}$			
$\vdots$	$\vdots$	$\vdots$		$\vdots$			
$n-1$	$\mathbf{S_{n-1,0}}$	$\mathbf{S_{n-2,1}}$	. . .	$\mathbf{S_{n-1-k,k}}$	. . .	$\mathbf{S_{0,n-1}}$	
n	$\mathbf{S_{n,0}}$	$\mathbf{S_{n-1,1}}$	. . .	$\mathbf{S_{n-k,k}}$	. . .	$\mathbf{S_{1,n-1}}$	$\mathbf{S_{0,n}}$

Selected reference

- Beck, T. and Webb, I. (2003). Economic, Demographic and institutional determinants of life insurance consumption across countries. *World Bank Economic Review* 17: 51-99
- Browne, M. and Kim, K. (1993). An International analysis of life insurance demand. *The Journal of Risk and Insurance* 60: 616-634
- Browne, M., Chung, J., and Frees, E.W. (2000). International property-liability insurance consumption. *The Journal of Risk and Insurance* 67: 73-90
- Kettani, H. (2010). World muslim population: 1950-2020. *International Journal of Environmental Science and Development*.
- Outreville, J. (1996). Life insurance market in developing countries. *The Journal of Risk and Insurance* 63: 263-278
- Shi, P. and Frees, E.W. (2010). Long-tail Longitudinal Modeling of Insurance Company Expenses. *Insurance: Mathematics and Economics* 47: 303-314