Multivariate longitudinal data analysis for actuarial applications

Priyantha Kumara and Emiliano A. Valdez

astin/afir/iaals Mexico Colloquia 2012

Mexico City, Mexico, 1-4 October 2012

Outline

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 and is increasing in popularity in data analysis, e.g. biometrics.
- There is a developing interest on multivariate longitudinal analysis in actuarial context (e.g Shi, 2011).
- Model accuracy, and further understanding, can be improved by incorporating dependency among multiple responses.
- Very often because of simplicity, response variables are typically assumed to have multivariate normal distribution.

Some literature

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

- 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.
- Seemingly unrelated regressions (SUR) approach
  - Rochon, J. (1996) Analyzing bivariate repeated measures for discrete and continuous outcome variable. *Biometrics* 52: 740-50.
- 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 q covariates associated with n subjects collected over T time periods for a set of m response variables.

- Let $y_{it,k}$ denote the responses from i^{th} individual in t^{th} time period on the k^{th} response. By letting $\mathbf{y_{it}} = (y_{it,1}, y_{it,2}, \ldots, y_{it,m})'$ for $t = 1, 2, \ldots, T$, we can express $\mathbf{Y_i} = (\mathbf{y_{i1}}, \mathbf{y_{i2}}, \ldots, \mathbf{y_{iT}})$.
- Covariates associated with the i^{th} subject in t^{th} time period on the k^{th} response can be expressed as $\mathbf{x_{it}} = (\mathbf{x_{it,1}}, \mathbf{x_{it,2}}, \ldots, \mathbf{x_{it,m}})$ where $\mathbf{x_{it,k}} = (x_{it1,k}, x_{it2,k}, \ldots, x_{itp,k})$ for $k = 1, 2, ...m$.
- We use α_{ik} to represent the random effects component corresponding to the i^{th} subject from the k^{th} response variable.
- $G\left(\alpha_{ik}\right)$ 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

Copula function

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

$$C(u_1, \ldots, u_m) = P(U_1 \le u_1, \ldots, U_m \le u_m).$$

- Joint distribution:

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

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

- Joint density:

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

  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

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

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

We can write

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

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

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

and from the previous slides, we have

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

Then, we can write the log likelihood function as

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

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 $Y \sim \mathsf{GB2}(a,b,p,q)$ 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:

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

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

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

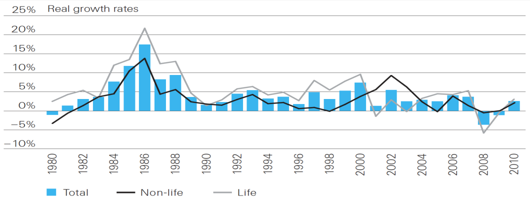
- Define residuals $\varepsilon_i = Y_i e^{-(\alpha_i + \beta'\mathbf{x}_i)}$ so that

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

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

- PP plots can then be used for diagnostics.
- See also McDonald (1984), McDonald and Butler (1987)

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:

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

About the data set

Data set

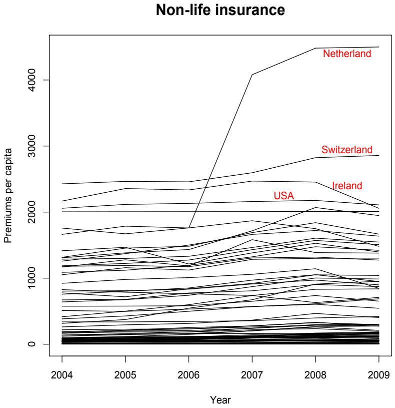
- 2 responses: life and non-life insurance
- 5 predictor variables
- 75 countries (originally, later removed 3 countries)
- 6 years data (from year 2004 to year 2009)

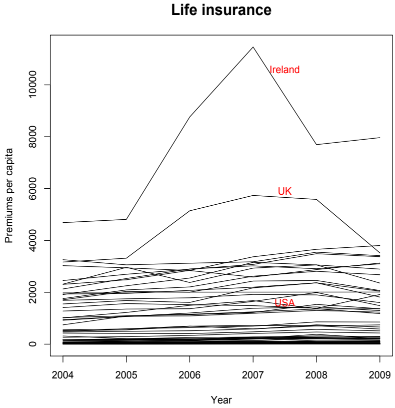
Variables in the model

Dependent variables	
Non-life density	Premiums per capita in non-life insurance
Life density	Premiums per capita in 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

Sources: Swiss Re sigma reports through the Insurance Information Institute (III); World Bank

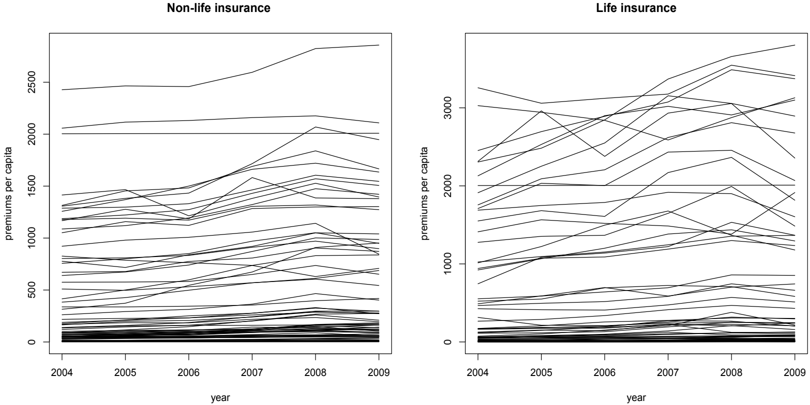
Multiple time series plot





Multiple time series plot: removed 3 countries

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



Some summary statistics

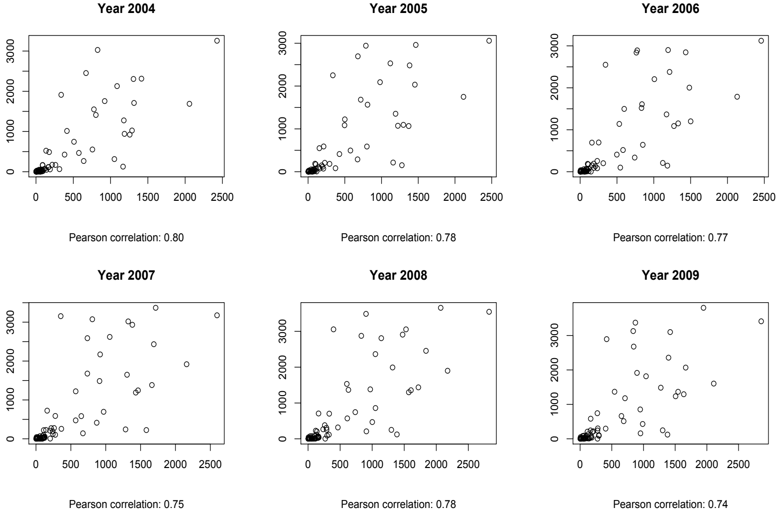
Summary statistics of variables in year 2004 to 2009:

Variable	Minimum	Maximum	Mean	Correlation with Life insurance	Correlation with Non-life insurance
Non-life insurance	(0.74, 1.26)	(2427.61, 2857.40)	(386.28, 516.99)	(0.75, 0.80)	-
Life insurance	(0.49, 1.28)	(3058.58, 3803.76)	(503.87, 697.39)	-	(0.75, 0.80)
GDP per capita	(375.20, 550.90)	(56311.50, 94567.90)	(13896.60, 20524.50)	(0.77, 0.82)	(0.90, 0.91)
Death rate	(1.50, 1.52)	(16.17, 17.11)	(7.87, 8.00)	(0.09, 0.11)	(0.06, 0.07)
Urbanization	(11.92, 13.56)	(100,100)	(64.90, 66.29)	(0.37, 0.42)	(0.45, 0.46)
Religious	(0.01,0.01)	(99.61, 99.61)	(22.12, 22.12)	(-0.30, -0.29)	(-0.30, -0.28)
Dependency ratio	(1.25, 1.39)	(29.31, 33.92)	(14.89, 15.55)	(0.57, 0.61)	(0.57, 0.60)

Correlation matrix of covariates in year 2004 to 2009:

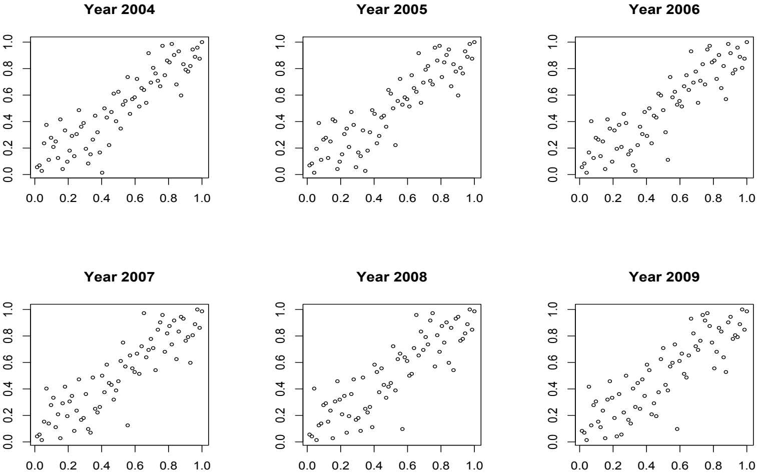
	GDP per capita	Death rate	Urbanization	Religious	Dependency ratio
GDP per capita	-				
Death rate	(0.01, 0.03)	-			
Urbanization	(0.49, 0.52)	(-0.16, -0.15)	-		
Religious	(-0.29, -0.25)	(-0.38, -0.34)	(-0.14, -0.13)	-	
Dependency ratio	(0.58, 0.62)	(0.53, 0.54)	(0.30, 0.32)	(-0.53, -0.52)	-

Scatter plots of the two response variables



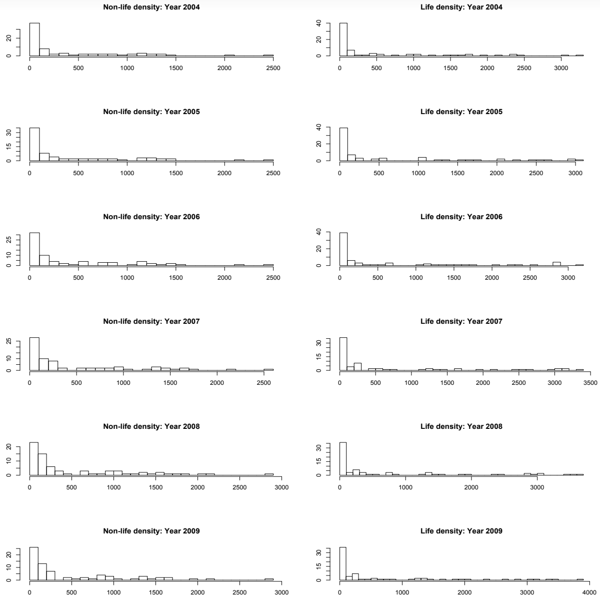
x-axis: non-life insurance and y-axis: life insurance

Scatter plots of the ranked response variables



x-axis: non-life insurance and y-axis: life insurance

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:

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

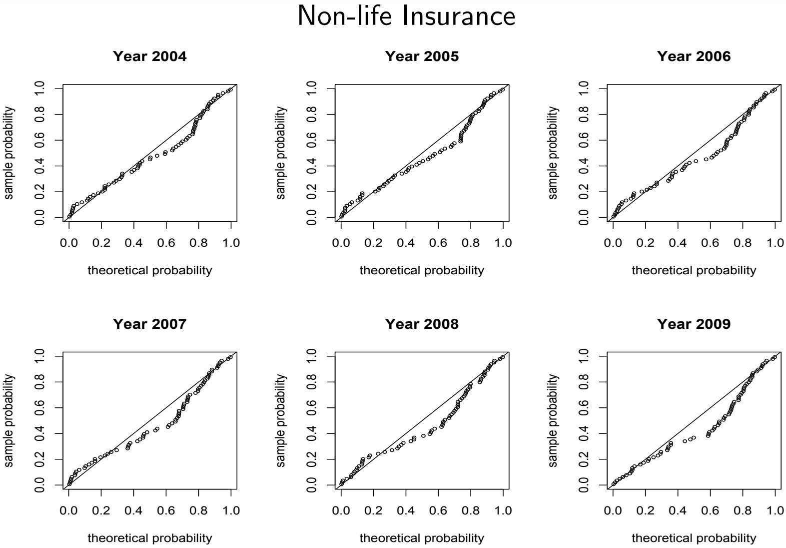
Model estimates

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

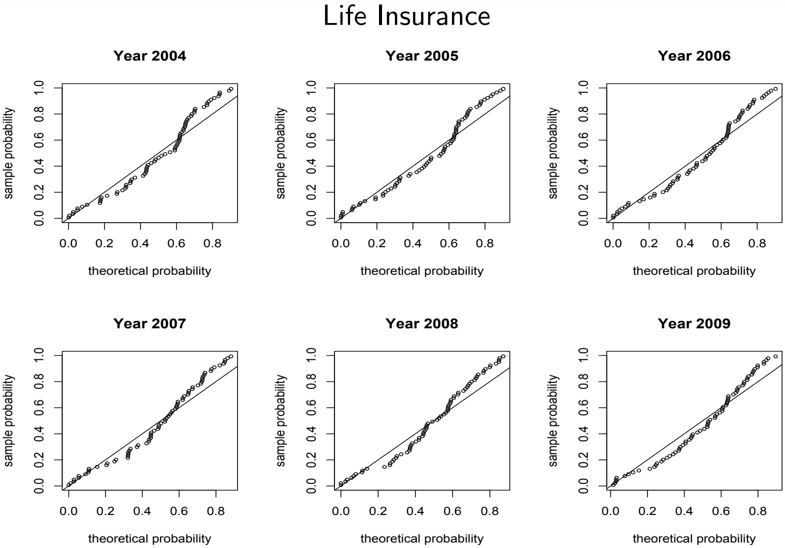
Gaussian copula:

Parameter	**Estimate**	**Std Error**	**p-val**
ρ	0.5174	0.0315	0.0000

PP plots of the residuals for marginal diagnostics



PP plots of the residuals for marginal diagnostics



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 an ongoing interest in loss reserving using multiple loss triangle.
- Alternative approach:

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

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
- 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

- Thank you -