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
   Some literature

The model specification
   Notation
   Key features of our approach
   Multivariate joint distribution
   Choice for the marginals: the class of GB2

Case study
   Global insurance demand

Additional work intended

Selected reference
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


- The random effects approach

- Seemingly unrelated regressions (SUR) approach

- Copula approach
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 \( y_{it} = (y_{it,1}, y_{it,2}, \ldots, y_{it,m})' \) for \( t = 1, 2, \ldots, T \), we can express \( Y_i = (y_{i1}, y_{i2}, \ldots, y_{iT}) \).

- Covariates associated with the \( i^{th} \) subject in \( t^{th} \) time period on the \( k^{th} \) response can be expressed as \( x_{it} = (x_{it,1}, x_{it,2}, \ldots, x_{it,m}) \) where \( x_{it,k} = (x_{it1,k}, x_{it2,k}, \ldots, x_{itp,k}) \) for \( k = 1, 2, \ldots, m \).

- We use \( \alpha_{ik} \) 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 \( \alpha_{ik} \).
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

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

- 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:

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

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

P. Kumara and E.A. Valdez, U of Connecticut Multivariate longitudinal data analysis 8/28
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

$$Y_{it} = (y_{it,1}, y_{it,2}, \ldots, y_{it,m})$$

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

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

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

$$H(y_{it} | \alpha_{i1}, \ldots, \alpha_{im}) = C(F(y_{it,1} | \alpha_{i1}), \ldots, F(y_{it,m} | \alpha_{im})).$$
Conditional joint density at time $t$:

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

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

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

We can write

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

Under independence over time for a given random effect:

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

and from the previous slides, we have

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

Then, we can write the log likelihood function as

\[
\sum_{i} \log \left\{ \int_{\alpha_{i1}} \cdots \int_{\alpha_{im}} \prod_{t=1}^{T} \prod_{k=1}^{m} c(F(y_{it,1} | \alpha_{i1}), \ldots, F(y_{it,m} | \alpha_{im})) \\
\quad \times f(y_{it,k} | \alpha_{ik}) dG(\alpha_{i1}) \cdots dG(\alpha_{im}) \right\}
\]
The model specification is flexible enough to accommodate any marginals; however, for our purposes, we chose the class of GB2 distributions. For $Y \sim GB2(a, b, p, q)$ 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:**

$$E(Y) = b \frac{B(p + 1/a, q - 1/a)}{B(p, q)}.$$
GB2 regression through the scale parameter

Suppose $x$ is a vector of known covariates:

- We have: $Y|x \sim GB2(a, b(x), p, q)$, where
  
  $$b(x) = \alpha + \beta'x$$

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

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

- PP plots can then be used for diagnostics.

- See also McDonald (1984), McDonald and Butler (1987)
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

<table>
<thead>
<tr>
<th>Dependent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-life density</td>
</tr>
<tr>
<td>Life density</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Independent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP per capita</td>
</tr>
<tr>
<td>Religious</td>
</tr>
<tr>
<td>Urbanization</td>
</tr>
<tr>
<td>Death rate</td>
</tr>
<tr>
<td>Dependency ratio</td>
</tr>
</tbody>
</table>

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

P. Kumara and E.A. Valdez, U of Connecticut
Multiple time series plot

Non-life insurance

Life insurance

P. Kumara and E.A. Valdez, U of Connecticut Multivariate longitudinal data analysis
Multiple time series plot: removed 3 countries

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

Non-life insurance

Life insurance

premiums per capita

year

2004 2005 2006 2007 2008 2009

0 500 1000 1500 2000 2500

0 1000 2000 3000

P. Kumara and E.A. Valdez, U of Connecticut

Multivariate longitudinal data analysis
Some summary statistics

Summary statistics of variables in year 2004 to 2009:

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

Correlation matrix of covariates in year 2004 to 2009:

<table>
<thead>
<tr>
<th></th>
<th>GDP per capita</th>
<th>Death rate</th>
<th>Urbanization</th>
<th>Religious</th>
<th>Dependency ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDP per capita</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death rate</td>
<td>(0.01, 0.03)</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urbanization</td>
<td>(0.49, 0.52)</td>
<td>(-0.16, -0.15)</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religious</td>
<td>(-0.29, -0.25)</td>
<td>(-0.38, -0.34)</td>
<td>(-0.14, -0.13)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Dependency ratio</td>
<td>(0.58, 0.62)</td>
<td>(0.53, 0.54)</td>
<td>(0.30, 0.32)</td>
<td>(-0.53, -0.52)</td>
<td>-</td>
</tr>
</tbody>
</table>
Scatter plots of the two response variables

Year 2004
Pearson correlation: 0.80

Year 2005
Pearson correlation: 0.78

Year 2006
Pearson correlation: 0.77

Year 2007
Pearson correlation: 0.75

Year 2008
Pearson correlation: 0.78

Year 2009
Pearson correlation: 0.74

x-axis: non-life insurance and y-axis: life insurance
Scatter plots of the ranked response variables

Year 2004

Year 2005

Year 2006

Year 2007

Year 2008

Year 2009

\(x\)-axis: non-life insurance and \(y\)-axis: life insurance

P. Kumara and E.A. Valdez, U of Connecticut

Multivariate longitudinal data analysis
Histograms of two responses from year 2004 to 2009
Model calibration

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

\[ C(u_1, u_2; \rho) = \Phi_{\rho}(\Phi^{-1}(u_1), \Phi^{-1}(u_2)) \]

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

\[ \alpha_{ik} \sim N(0, \sigma_k^2) \]
Model estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Non-life insurance density</th>
<th>Life insurance density</th>
<th>p-val</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std Error</td>
<td>p-val</td>
<td>Estimate</td>
</tr>
<tr>
<td>Covariates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDP per capita</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0001</td>
</tr>
<tr>
<td>Religious</td>
<td>-0.0085</td>
<td>0.0023</td>
<td>0.0000</td>
<td>-0.0231</td>
</tr>
<tr>
<td>Urbanization</td>
<td>0.0567</td>
<td>0.0022</td>
<td>0.0000</td>
<td>0.0279</td>
</tr>
<tr>
<td>Death rate</td>
<td></td>
<td></td>
<td></td>
<td>0.0035</td>
</tr>
<tr>
<td>Dependency ratio (old)</td>
<td></td>
<td></td>
<td></td>
<td>-0.0440</td>
</tr>
<tr>
<td>GB2 Marginals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>2.5636</td>
<td>0.1397</td>
<td>0.0000</td>
<td>1.0427</td>
</tr>
<tr>
<td>p</td>
<td>1.3957</td>
<td>0.1356</td>
<td>0.0000</td>
<td>3.7321</td>
</tr>
<tr>
<td>q</td>
<td>0.5369</td>
<td>0.0364</td>
<td>0.0000</td>
<td>0.5081</td>
</tr>
<tr>
<td>Random effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Sigma_\alpha$</td>
<td>0.6471</td>
<td>0.0535</td>
<td>0.0000</td>
<td>0.8507</td>
</tr>
</tbody>
</table>

Gaussian copula:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Error</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.5174</td>
<td>0.0315</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
PP plots of the residuals for marginal diagnostics

Non-life Insurance

Year 2004

Year 2005

Year 2006

Year 2007

Year 2008

Year 2009

P. Kumara and E.A. Valdez, U of Connecticut

Multivariate longitudinal data analysis
PP plots of the residuals for marginal diagnostics

Life Insurance

Year 2004

Year 2005

Year 2006

Year 2007

Year 2008

Year 2009
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


- Thank you -