Longitudinal Modeling of Claim Counts using Jitters

joint work with Dr. Peng Shi, Northern Illinois University

The Ronald H. and Mary E. Simon Actuarial Science Lecture
Michigan State University, East Lansing

13 April 2012
Outline

1 Introduction
   Background
   Literature

2 Modeling
   Random effects models
   Copula models
   Continuous extension with jitters
   Some properties

3 Empirical analysis
   Model specification
   Singapore data

4 Inference
   Variable selection
   Estimation results
   Model validation

5 Concluding remarks

6 Selected reference
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 $x_{it}$ 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.
Copula regression for multivariate discrete data:
- Increasingly becoming popular
- Applications found in various disciplines:
  - Biostatistics: Song et al. (2008), Madsen and Fang (2010)
  - Actuarial science: Purcaru and Denuit (2003), Shi and Valdez (2011)
- Modeling longitudinal insurance claim counts:
  - Boucher, Denuit and Guillén (2010): model joint pmf of claim counts


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 $\alpha_i$, 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** $N_{it} \sim \text{Poisson}(\tilde{\lambda}_{it})$
  
  \[ \tilde{\lambda}_{it} = \eta_i \lambda_{it} = \eta_i \omega_{it} \exp(x_{it}' \beta), \text{ and } \eta_i \sim \text{Gamma}(\psi, \psi) \]
  
  \[ \tilde{\lambda}_{it} = \omega_{it} \exp(\alpha_i + x_{it}' \beta), \text{ and } \alpha_i \sim \text{N}(0, \sigma^2) \]

- **Negative Binomial**
  
  \[ \text{NB1: } 1 + 1/\nu_i \sim \text{Beta}(a, b) \]
  
  \[
  \Pr(N_{it} = n_{it}|\nu_i) = \frac{\Gamma(n_{it}+\nu_i)}{\Gamma(\nu_i)\Gamma(n_{it}+1)} \left( \frac{\nu_i}{\nu_i + 1 + \nu_i} \right)^{\lambda_{it}} \left( \frac{1}{1 + \nu_i} \right)^{n_{it}}
  \]
  
  \[ \text{NB2: } \alpha_i \sim \text{N}(0, \sigma^2) \]
  
  \[
  \Pr(N_{it} = n_{it}|\alpha_i) = \frac{\Gamma(n_{it}+\psi)}{\Gamma(\psi)\Gamma(n_{it}+1)} \left( \frac{\psi}{\tilde{\lambda}_{it} + \psi} \right)^{\lambda_{it}} \left( \frac{\tilde{\lambda}_{it}}{\tilde{\lambda}_{it} + \psi} \right)^{n_{it}}
  \]

- **Zero-inflated models**

  \[ \Pr(N_{it} = n_{it}|\delta_i, \alpha_i) = \begin{cases} 
  \pi_{it} + (1 - \pi_{it})f(n_{it}|\alpha_i) & \text{if } n_{it} = 0 \\
  (1 - \pi_{it})f(n_{it}|\alpha_i) & \text{if } n_{it} > 0
  \end{cases} \]

  \[ \log \left( \frac{\pi_{it}}{1 - \pi_{it}} \right) | \delta_i = \delta_i + z_{it}' \gamma, \]

  \[ \text{ZIP (} f \sim \text{Poisson) and ZINB (} f \sim \text{NB) } \]
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 $i$th policyholder $(N_{i1}^*, N_{i2}^*, \ldots, N_{iT}^*)$ may be expressed as:

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

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

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

- 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:**
  
  \[
  (N_{a1}, N_{b1}) \prec_c (N_{a2}, N_{b2}), \text{ then } (N^*_{a1}, N^*_{b1}) \prec_c (N^*_{a2}, N^*_{b2})
  \]

- **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 $\lambda_{it} = \exp(x_{it}'\beta)$.

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

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>88.10</td>
<td>85.86</td>
<td>85.21</td>
<td>83.88</td>
<td>90.41</td>
<td>85.62</td>
<td>86.89</td>
<td>87.18</td>
<td>89.71</td>
<td>3480</td>
<td>86.9</td>
</tr>
<tr>
<td>2</td>
<td>1.47</td>
<td>2.00</td>
<td>1.25</td>
<td>1.83</td>
<td>0.00</td>
<td>0.65</td>
<td>1.37</td>
<td>0.92</td>
<td>0.57</td>
<td>50</td>
<td>1.25</td>
</tr>
<tr>
<td>3</td>
<td>0.37</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>1.37</td>
<td>0.00</td>
<td>0.15</td>
<td>0.18</td>
<td>0.00</td>
<td>6</td>
<td>0.15</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.18</td>
<td>0.00</td>
<td>0.00</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>Number</td>
<td>546</td>
<td>601</td>
<td>480</td>
<td>273</td>
<td>73</td>
<td>306</td>
<td>656</td>
<td>546</td>
<td>525</td>
<td>4006</td>
<td>100</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Person Age (in years)</th>
<th>Percentage by Count</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>25 and younger</td>
<td>73.33</td>
<td>23.33</td>
</tr>
<tr>
<td>26-35</td>
<td>87.49</td>
<td>11.12</td>
</tr>
<tr>
<td>36-45</td>
<td>86.63</td>
<td>11.80</td>
</tr>
<tr>
<td>46-60</td>
<td>86.85</td>
<td>11.92</td>
</tr>
<tr>
<td>60 and over</td>
<td>91.67</td>
<td>6.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>91.49</td>
<td>7.98</td>
<td>0.53</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Male</td>
<td>86.64</td>
<td>11.86</td>
<td>1.28</td>
<td>0.16</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No Claims Discount (NCD)</th>
<th>Percentage by Count</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>84.83</td>
<td>13.17</td>
</tr>
<tr>
<td>10</td>
<td>86.21</td>
<td>12.58</td>
</tr>
<tr>
<td>20</td>
<td>89.21</td>
<td>9.25</td>
</tr>
<tr>
<td>30</td>
<td>89.16</td>
<td>9.49</td>
</tr>
<tr>
<td>40</td>
<td>88.60</td>
<td>11.40</td>
</tr>
<tr>
<td>50</td>
<td>88.83</td>
<td>10.46</td>
</tr>
</tbody>
</table>

| 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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>RE-Poisson</th>
<th>RE-NegBin</th>
<th>RE-ZIP</th>
<th>RE-ZINB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>p-value</td>
<td>Estimate</td>
<td>p-value</td>
</tr>
<tr>
<td>intercept</td>
<td>-1.7173</td>
<td>&lt;.0001</td>
<td>1.6404</td>
<td>0.1030</td>
</tr>
<tr>
<td>young</td>
<td>0.6408</td>
<td>0.0790</td>
<td>0.6543</td>
<td>0.0690</td>
</tr>
<tr>
<td>midfemale</td>
<td>-0.7868</td>
<td>0.0310</td>
<td>-0.7692</td>
<td>0.0340</td>
</tr>
<tr>
<td>zeroncd</td>
<td>0.2573</td>
<td>0.0050</td>
<td>0.2547</td>
<td>0.0060</td>
</tr>
<tr>
<td>vage</td>
<td>-0.0438</td>
<td>0.0210</td>
<td>-0.0442</td>
<td>0.0210</td>
</tr>
<tr>
<td>vbrand1</td>
<td>0.5493</td>
<td>&lt;.0001</td>
<td>0.5473</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>vbrand2</td>
<td>0.1831</td>
<td>0.0740</td>
<td>0.1854</td>
<td>0.0710</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>AR(1)</th>
<th>Exchangeable</th>
<th>Toeplitz(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>StdErr</td>
<td>Estimate</td>
</tr>
<tr>
<td>intercept</td>
<td>-1.8028</td>
<td>0.0307</td>
<td>-1.8422</td>
</tr>
<tr>
<td>young</td>
<td>0.6529</td>
<td>0.0557</td>
<td>0.7130</td>
</tr>
<tr>
<td>midfemale</td>
<td>-0.6956</td>
<td>0.0588</td>
<td>-0.6786</td>
</tr>
<tr>
<td>zeroncd</td>
<td>0.2584</td>
<td>0.0198</td>
<td>0.2214</td>
</tr>
<tr>
<td>vage</td>
<td>-0.0411</td>
<td>0.0051</td>
<td>-0.0422</td>
</tr>
<tr>
<td>vbrand1</td>
<td>0.5286</td>
<td>0.0239</td>
<td>0.5407</td>
</tr>
<tr>
<td>vbrand2</td>
<td>0.1603</td>
<td>0.0166</td>
<td>0.1752</td>
</tr>
<tr>
<td>ϕ</td>
<td>2.9465</td>
<td>0.1024</td>
<td>2.9395</td>
</tr>
<tr>
<td>ρ₁</td>
<td>0.1216</td>
<td>0.0028</td>
<td>0.1152</td>
</tr>
<tr>
<td>ρ₂</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

\[
\begin{align*}
f_{iT+1}(n_{iT+1}|n_{i1}, \ldots, n_{iT}) &= \Pr(N_{iT+1} = n_{iT+1}|N_{i1} = n_{i1}, \ldots, N_{iT} = n_{iT}) \\
&= \mathbb{E}_{u_i} \left[ c(F_{iT}^*(n_{i1} - U_{i1}), \ldots, F_{iT}^*(n_{iT} - U_{iT}); \theta) \prod_{t=1}^{T+1} f_{it}^*(n_{it} - U_{it}) \right] \\
&= \mathbb{E}_{u_i} \left[ c(F_{iT}^*(n_{i1} - U_{i1}), \ldots, F_{iT}^*(n_{iT} - U_{iT}); \theta) \prod_{t=1}^{T} f_{it}^*(n_{it} - U_{it}) \right].
\end{align*}
\]

- **Performance measures used:**
  - LogLik \(= \sum_{i=1}^{M} \log (f_{iT+1}(n_{iT+1}|n_{i1}, \ldots, n_{iT}))\)
  - MSPE \(= \sum_{i=1}^{M} [n_{iT+1} - \mathbb{E}(N_{iT+1}|N_{i1} = n_{i1}, \ldots, N_{iT} = n_{iT})]^2\)
  - MAPE \(= \sum_{i=1}^{M} |n_{iT+1} - \mathbb{E}(N_{iT+1}|N_{i1} = n_{i1}, \ldots, N_{iT} = n_{iT})|\)
Results of model validation

**t-plot**

Uniform QQ Plot for Gaussian Copula

Out-of-sample validation

<table>
<thead>
<tr>
<th></th>
<th>Standard Model</th>
<th>Copula Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RE-Poisson</td>
<td>RE-NegBin</td>
</tr>
<tr>
<td>LogLik</td>
<td>-177.786</td>
<td>-177.782</td>
</tr>
<tr>
<td>MSPE</td>
<td>0.107</td>
<td>0.107</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.213</td>
<td>0.213</td>
</tr>
</tbody>
</table>
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


