Claims reserving in non-life insurance: old and new adventures

One World Actuarial Research Seminar

Katrien Antonio and Jonas Crevecoeur
LRisk - KU Leuven and ASE - University of Amsterdam

June 17, 2020
My personal website: https://katrienantonio.github.io.

Talk is based on joint work with Jonas Crevecoeur, Roel Verbelen and Gerda Claeskens.
We typically aggregate the data from the time line into a run-off triangle.

Covariates largely ignored!
This webinar’s mission statement

1. Launch a discussion on individual, granular data for loss reserving, and their features.

2. Sketch published research on the modeling of IBNR claim counts.

3. Sketch ongoing research on the development of RBNS claims.

4. Provide (data driven) guidance on the choice between aggregate and individual reserving for a given portfolio.

5. Reflect upon stream of academic research on data analytics for reserving.
Research on modelling IBNR claim counts
The insurance company is **not aware** (yet) of claims related to past exposures that are not (yet) reported!
The insurance company is not aware (yet) of claims related to past exposures that are not (yet) reported!
Research questions

IBNR claim counts

- Research questions with focus on **IBNR claim counts**:
  - *How many* claims occurred but are not yet reported, because reporting delay is subject to right truncation?
  - *When* will these IBNR claims be reported?

- **Pioneering work** by Ragnar Norberg (1993, 1999), (basic, first) implementation in Antonio & Plat (2014), new work in Verrall & Wüthrich (2016), all in continuous time!

Case study with liability claims data set

Claim occurrence process

Observation window: July 1, 1996 to August 31, 2009, MM/DD/YY format, i.e. day is natural time unit.
Case study with liability claims data set

Reporting process
Case study with liability claims data set

Reporting delay

Declining pattern in reporting delay + intra-week pattern, depending on the occurrence day of the week.

(a) All claims

(b) Monday

(c) Thursday

On (semi-)official holidays: drop in number of reported claims compared to daily average.
Case study with liability claims data set

Total IBNR counts

![Graph showing unreported claims over time from Oct 2003 to Jul 2004 and Oct 01 to Dec 01.](image-url)
The statistical model for IBNR

<table>
<thead>
<tr>
<th></th>
<th>IBNR</th>
<th>RBNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>event structure</td>
<td>single event</td>
<td>multiple, recurrent events</td>
</tr>
<tr>
<td>time horizon</td>
<td>usually quick</td>
<td>longer (in years)</td>
</tr>
<tr>
<td>time granularity</td>
<td>in days since occurrence</td>
<td>in years since reporting</td>
</tr>
<tr>
<td>covariates</td>
<td>triangle with daily occurrences and reportings</td>
<td>individual claim- and policy(holder) specific</td>
</tr>
<tr>
<td>other fields</td>
<td>nowcasting in epidemiology; occurrence of events, observed with delay</td>
<td>recurrent events with marks</td>
</tr>
</tbody>
</table>
The statistical model for IBNR

Notations

- $N_t$: the (total) number of claims that occurred on day $t$.
- $N_{t,s}$: the number of claims from day $t$ that are reported on day $s$.
- Each claim eventually gets reported, thus $N_t = \sum_{s=t}^{\infty} N_{t,s}$.
- Assumption: the $N_{t,s}$ are independent and $N_{t,s} \sim \text{POI}(\lambda_t \cdot p_{t,s})$. 
Research contributions

   - joint estimation of occurrence process and reporting delay distribution
   - use EM to optimize the likelihood in presence of missing data
   - regression at \((t,s)\) level.

   - time-change strategy, daily reporting exposures
     - focus on calendar day effects, e.g. national holidays and weekend, reporting at specific delays (e.g. 14 days, 1 year)
   - regression at \((t,s)\) level, investigate different settings via simulation study.
Time change strategy

The idea pictured!

\[
p_{t,s} = \int_{s-t}^{s-t+1} f_U(u) \, du
\]

where

\[
\phi_t(d) = \sum_{i=1}^{\alpha_{t,s}} d_i
\]

with \( \alpha_{t,s} \) reporting exposure.
Time change strategy
Structuring the reporting exposures

- Use a standard distribution for $\tilde{U}$ (e.g. exponential, lognormal).

- Explain the daily reporting exposures as a function of covariates:

$$\alpha_{t,s} = \exp(x'_{t,s} \cdot \gamma).$$

- Joint estimation of distribution $\tilde{U}$ and regression parameters to structure $\alpha_{t,s}$.

Use maximum likelihood estimation with the likelihood of the reported claims.
When are the claims that are IBNR (on August 31, 2004) reported?
Case study

IBNR Results - second evaluation
Case study

IBNR Results - third evaluation, see Hiabu (2017, SAJ) and Martínez-Miranda, Nielsen et al. (2013)
Wrap-up
IBNR reserving

- Accurate predictions under weekday and holiday effects.
- Faster detections of changes in the occurrence/reporting process (cfr. simulation study).
- Triangle at daily level, incorporate covariates.
- Longer computation time.
- More choices (i.e. distributional assumptions, selecting variables) - should be done with care!
Run-off triangles in epidemiology

---

From van de Kaaftele, Eilers & Wallinga in Epidemiology (2019)

From Bastos et al. in Statistics in Medicine (2019)
What else is there - going beyond actuarial science?

Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing

van de Kassteele, Eilers & Wallinga, in Epidemiology (2019).

- nowcasting = assessment of the current situation based on imperfect or partial information

A modelling approach for correcting reporting delays in disease surveillance data

Bastos et al., in Statistics in Medicine (2019).
Research on the development of RBNS claims
Research focus - RBNS

Settlement delay

Occurrence → Reporting → Closure

IBNR → RBNS → Closed
Development of an RBNS claim

Continuous time
Development of an RBNS claim

Continuous time

Diagram showing the development of RBNS claims over time, with axes for reporting date and payment delay (in days).
The statistical model for RBNS

<table>
<thead>
<tr>
<th>Thought</th>
<th>IBNR</th>
<th>RBNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>event structure</td>
<td>single event</td>
<td>multiple, recurrent events</td>
</tr>
<tr>
<td>time horizon</td>
<td>usually quick</td>
<td>longer (in years)</td>
</tr>
<tr>
<td>time granularity</td>
<td>in days since occurrence</td>
<td>in years since reporting</td>
</tr>
<tr>
<td>covariates</td>
<td>triangle with daily occurrences and reportings</td>
<td>individual claim- and policy(holder) specific</td>
</tr>
<tr>
<td>other fields</td>
<td>nowcasting in epidemiology, occurrence of events, observed with delay</td>
<td>recurrent events with marks</td>
</tr>
</tbody>
</table>
The statistical model
RBNS claim development in discrete time

May, 2004
Claim reported

March, 2005
Payment: 250

July, 2005
Payment: 700

March, 2006
Payment: 3200

September, 2006
Claim closes
The statistical model

RBNS claim development in discrete time

Development period 1
- Claim reported
- Payment: 0

Development period 2
- Payment: 950

Development period 3
- Payment: 3200
- Claim closed

May, 2004
Claim reported

March, 2005
Payment: 250

July, 2005
Payment: 700

March, 2006
Payment: 3200

September, 2006
Claim closes

Time
INDIVIDUAL CLAIMS, DISCRETE TIME PERIODS

CLAIM $k$

$X_k \rightarrow U_k^2 \rightarrow U_k^3 \rightarrow \cdots \rightarrow U_k^T$

REPORTING PERIOD

LAYERED STRUCTURE

$C_k \rightarrow P_k \rightarrow Y_k$
A hierarchical reserving model for RBNS claims

Layers

- Index the individual claims by $k$ and the development periods by $j$.

- Our approach is modular or layered:
  
  - $\mathbf{x}_k$ denotes the (observed, fixed) claim information available at the end of the first development period, i.e. the reporting period
    
    e.g. cause of claim, policy(holder) covariates, initial case estimate
  
  - $\mathbf{U}^j_k$ is the vector with claim $k$’s updated information in development period $j$
    
    depends on portfolio at hand, e.g. $\mathbf{U}^j_k = (C^j_k, P^j_k, Y^j_k)$ with a settlement indicator $C^j_k$, a payment indicator $P^j_k$ and payment size $Y^j_k$.

- Need more modular components ⇒ extend $\mathbf{U}^j_k$!
A hierarchical reserving model for RBNS claims

Hierarchical structure

- Update vectors $U^j_k$ for claim $k$ are observed from development period 2 to $\tau_k$:

$$\mathcal{R}^{\text{Obs}} = \{U^j_k \mid k = 1, \ldots, n, j = 2, \ldots, \tau_k\}.$$ 

- We introduce a time dynamic hierarchical structure (see Frees & Valdez, 2008, JASA):

$$\mathcal{L}(\mathcal{R}^{\text{Obs}}) = \prod_{k=1}^{n} f\left(U_k^{(2)}, \ldots, U_k^{(\tau_k)} \mid x_k\right)$$

$$= \prod_{k=1}^{n} \prod_{j=2}^{\tau_k} f\left(U^j_k \mid U_k^{(2)}, \ldots, U_k^{(j-1)}, x_k\right).$$

Thus, future development depends on the past.
A hierarchical reserving model for RBNS claims

Hierarchical structure

- As a last step, we introduce a layered hierarchical structure for $U^j_k$:

$$\mathcal{L}(R^{obs}) = \prod_{k=1}^{n} \prod_{j=2}^{\tau_k} \prod_{l=1}^{s} f\left(U^j_{k,l} \mid U^{(2)}_k, \ldots, U^{(j-1)}_k, U_{k,1}, \ldots, U^j_{k,l-1}, x_k\right),$$

with $s$ the number of layers in the update vector.

- For example, with $U^j_k = (C^j_k, P^j_k, Y^j_k)$ we focus on the three essential building blocks from Antonio & Plat (2014)!

- The framework incorporates static (via $x_k$) as well as dynamic features.
A hierarchical reserving model for RBNS claims

Three layers $U^j_k = (C^j_k, P^j_k, Y^j_k)$

- $C^j_k$ is one if claim $k$ settles in development period $j$ and zero otherwise

  $$C^j_k \mid U^{(2)}_k, \ldots, U^{(j-1)}_k, x_k \sim \text{Bernoulli}(p(U^{(2)}_k, \ldots, U^{(j-1)}_k, x_k)).$$

- $P^j_k$ is one if there is a payment for claim $k$ in development period $j$ and zero otherwise

  $$P^j_k \mid U^{(2)}_k, \ldots, U^{(j-1)}_k, C^j_k, x_k \sim \text{Bernoulli}(q(U^{(2)}_k, \ldots, U^{(j-1)}_k, C^j_k, x_k)).$$

- $Y^j_k$ is the payment size, given $P^j_k = 1$. The payment size is gamma distributed with mean

  $$E(Y^j_k \mid U^{(2)}_k, \ldots, U^{(j-1)}_k, C^j_k, P^j_k, x_k) = \mu(U^{(2)}_k, \ldots, U^{(j-1)}_k, C^j_k, P^j_k, x_k).$$
A hierarchical reserving model for RBNS claims

Model calibration

Guidelines for the model calibration:

- you can use your preferred predictive model (e.g. GLM or Gradient Boosting Machine)
- apply $k$-fold cross validation (to prevent overfitting) and a weighted likelihood

$$\prod_{k=1}^{n} \prod_{j=2}^{\tau_k} w_j \cdot f \left( U_{k,1}^{j}, \ldots, U_{k}^{(j-1)}, U_{k,1}, \ldots, U_{k,l-1}, x_k \right),$$

per layer $l$ in the model.

The weights tackle the covariate shift (or imbalance), e.g. more claims with later development years in lower vs. upper ‘triangle’.
Bridging aggregate and individual reserving

The choice between an individual and an aggregate reserving model then depends (in a data driven way) on:

- the covariates included in the predictive model, e.g. for layer $l$

$$E(U^j_{k,l}) = \tilde{\alpha}_{i_k+r_k,l} \cdot \beta_{j,l} \cdot \exp \left\{ \phi \left( U^2_k, \ldots, U^{j-1}_k, U^j_{k,1}, \ldots, U^j_{k,l-1}, x_k \right) \right\},$$

with $i_k + r_k$ the reporting year of claim $k$.

For layer $l$, if $\phi(.) = 0$ (i.e. no relevant covariates) summing the individual, claim-specific updates reduces to a triangle with the multiplicative chain ladder structure (since reporting).
Connections with the literature (a selection)

Denuit & Trufin (2017, 2018), Wahl et al. (2019), RBNS in Double CL

Reserving by combining multiple runoff triangles for closure, payment and size.

Wüthrich (2018), special issue Risks edited by Taylor (2020)

Machine learning methods (e.g. regression tree, GBM) for the payment and closure indicator.

Larsen (2007)

GLM fitted for each of the components ($C$, $P$ and $Y$) in the development process.
Data exploration

Home insurance claims - global events
Data exploration

Home insurance claims - treemap
Multiple evaluation dates

(a) non–fire claims

(b) non–fire claims, exclude extreme weather events
Multiple evaluation dates

\[ \text{Percentage Error} = 100 \cdot \frac{\text{predicted} - \text{actual}}{\text{actual}}. \]
## Overall performance

<table>
<thead>
<tr>
<th>Portfolio</th>
<th>hierarchical GLM</th>
<th>hierarchical GBM</th>
<th>chain ladder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu(PE)$</td>
<td>$\mu(\mid PE \mid)$</td>
<td>$\mu(PE)$</td>
</tr>
<tr>
<td>non-fire claims</td>
<td>0.92</td>
<td><strong>7.32</strong></td>
<td>-1.80</td>
</tr>
<tr>
<td>non-fire claims, exclude</td>
<td>-9.76</td>
<td><strong>14.90</strong></td>
<td>-14.28</td>
</tr>
<tr>
<td>extreme weather</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fire-claims</td>
<td>-20.82</td>
<td><strong>26.44</strong></td>
<td>-16.42</td>
</tr>
</tbody>
</table>

Average performance is expressed as the mean percentage error and the mean absolute percentage error.
Take home insights

✓ Structure the (highly) scattered literature on analytics for loss reserving.

✓ Hybrid strategy, take data-driven position between individual and aggregate.

✓ Less is more, unify pricing and reserving methodology (e.g. GLMs, GBMs).

✓ Lessons to learn from the machine learning literature.

✓ Use multiple evaluation dates, instead of single out-of-time.

✓ Use multiple portfolios, no free lunch.
More information

For more information, please visit:

LRisk website, www.lrisk.be

https://katrienantonio.github.io

Thanks to
For an overview of the literature, please see the references in:

