Segmented individual claim reserve estimation using BART refinement

A R Effendie
Actuarial Science Study Program, Department of Mathematics, Universitas Gadjah Mada, Yogyakarta
E-mail: adhityaronnie@ugm.ac.id

A Rahim
Department of Mathematics, Faculty of Mathematics and Natural Sciences University of Hasanuddin, Makassar
E-mail: amran@science.unhas.ac.id

Abstract. We want to extend an individual claim reserving method proposed by [1] and also [2] by using segmented calculation. This method is an individual method of claims reserve estimation which involves detailed condition on claim characteristics in the calculation process. Data is divided into several segments according to combination of background variables. We then apply RDC method to find estimated IBNR and RBNS reserves for each segment. Bayesian Additive Regression Tree (BART) is used to refine the estimated reserves. This is because the estimation become unstable due to a lot of combination factors for each segment.

1. Introduction
Starting point of this paper is the Reserve by Detailed Conditioning (RDC) method, as introduced by [1]. RDC - in its original specification - is a deterministic reserving method, designed for individual claims in discrete time. A remarkable and innovative aspect of the method is its ability to condition on claim characteristics, which are used for identification or clustering of similar claims. Conditional on a specific set of claim characteristics, a best estimate for the reserve attached to an open claim is obtained from the observed, historical development of claims from the same cluster, hence with similar characteristics [3].
2. The Claim Process

In general, the loss reserve is the total outstanding payments of all incurred claims, whether reported or not. In other words, it is an aggregation of the outstanding payments for every single claim. The claim process reflects the dynamics of the development of a single claim and is discussed in [4].

Following [1] we determine reserves by conditioning on claims characteristics. These characteristics summarize information registered during the development of a claim. They allow for the identification of similar claims. In this work we consider the claim length, the last observed cumulative payment and the reporting delay as claim characteristics.

\[ L(k) = F(k) - W(k) + 1. \]
Conditional claim length probability:

\[ P(L = \lambda | L > t) = \left[ \prod_{k=t+1}^{\lambda-1} P(L > k | L \geq k) \right] P(L = \lambda | L \geq \lambda) \] (2)

for \( 0 \leq t \leq n - 1 \) and \( t + 1 \leq \lambda \leq n \).

2.2. Mean payment

Define the sum of amounts paid up to and including period \( t \) from reporting as

\[ H(t) = \sum_{h=1}^{t} Y(h + W - 1), t = 0, 1, \ldots, n \] (3)

where \( h \) is counted from reporting with the reporting period \( W \) having \( h = 1 \). We want to predict the expected remaining payment sum from the known sum. Consider this expression:

\[ E[H(L) - H(t)|L > t, H(t), W] \] (4)

For \( t = n - i - W + 2 \) an estimate of this expression gives the RBNS (Reported But Not Settled) reserve of a reported open claim. For \( t = 0 \) we obtain the IBNR (Incurred But Not Reported) reserve per claim.

2.3. Rosenlund’s Estimator of claim reserve

Define the underlying reserve for a claim as

\[ R(q, w, t) = E[H(L) - H(t)|L > t, Q_t = q, W \wedge w_0 = w] \] (5)

\[ \hat{R}(q, w, t) = \sum_{\lambda=t+1}^{n} \sum_{h=t+1}^{\lambda} \hat{p}_\lambda(q, w, t) \hat{\mu}_{\lambda h}(q, w, t) \] (6)

where

\[ p_\lambda(q, w, t) = P(L = \lambda | L > t, Q_t = q, W \wedge w_0 = w) \] (7)

is the probability of claim length and

\[ \mu_{\lambda h}(q, w, t) = E[Y(h + W - 1)|L = \lambda, Q_t = q, W \wedge w_0 = w] \] (8)

is the expected of claim payment for \( 0 \leq t \leq n - 1, t + 1 \leq \lambda \leq n \) and \( t + 1 \leq h \leq \lambda n \).

3. Data

The dataset that we used in this research contains information on 58,573 individual claims of BPJS Kesehatan (Indonesian Social Health Insurance) during 2015. 51,978 claims are closed and remaining 6,595 claims are open. All claims come from Special Capital Region of Jakarta, reported from 4 major public hospitals with different specialties (Heart center, Cancer center, Children and Mother hospital and general hospital) There are four background variables available from the data:

- **Prov (f1)**: Provider: 1. Cipto Mangunkusumo general hospital, 2. Dharmais National Cancer Center, 3. Harapan Kita National Heart Center and 4. Harapan Kita Children and Mother hospital.
Figure 1. Figure caption for first of two sided figures.

Figure 2. Figure caption for second of two sided figures.

- **LofB (f2):** Line of Business: 1. PBI (Funded by Government) 2. PPU (Employee obligation) 3. PBPU (Other sources)
- **MedBen (f3):** Type of Medical Benefit: 1. Procedure, 2. Non-procedure, 3. Maternity
- **Resint (f4):** Resource intensity: 1. Low, 2. Medium, 3. High, 4. Outpatient

Claim currency was converted to Canadian dollar with currency rate on September 1, 2018. The mean, median and standard deviation of severity was $2,245.9, $722.3 and $7,585.99, respectively.

Figure 3. Figure caption for first of two sided figures.

Figure 4. Figure caption for second of two sided figures.

3.1. **Aggregate Incremental Run-Off Triangle**

We will show the following aggregate incremental Run-off triangle of the data:
Here we give full (cumulative) triangle, computed by classical (Chain Ladder) method:

The outstanding reserve is CAD 27,425,948

4. RDC Method

Expand the triangle into individual run-off triangle:

| Row | ID | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|-----|----|---|---|---|---|---|---|---|---|---|---|----|----|----|
| 1   | 6451| 1  | 12| NA| O  | 1  | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 2   | 6452| 1  | 12| NA| O  | 1  | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 3   | 6453| 1  | 12| NA| O  | 1  | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 4   | 6454| 1  | 12| NA| O  | 1  | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 5   | 6455| 1  | 12| NA| O  | 1  | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| 6   | 6456| 1  | 12| NA| O  | 1  | 3 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |

The outstanding reserve is CAD 27,425,948

4.1. Claim characteristics

Basic statistics of payment delay (developments):

Some options: \( w_0 = 3.7 \) and \( q_0 = 10, 15, 37 \) (Sturgess)

4.2. RDC results

\[
\begin{array}{|c|c|c|c|c|}
\hline
w_0 & q_0 & IBNR & RBNS & Total \\
\hline
3 & 10 & 18,553,258 & 6,749,386 & 25,302,644 \\
3 & 15 & 18,553,258 & 6,986,395 & 25,539,653 \\
3 & 37 & 18,553,258 & 6,814,946 & 25,368,204 \\
7 & 10 & 17,913,593 & 6,988,734 & 24,903,327 \\
7 & 15 & 17,913,593 & 7,255,622 & 25,169,215 \\
7 & 37 & 17,913,593 & 7,074,267 & 24,987,860 \\
12 & 10 & 17,478,848 & 6,895,006 & 24,373,854 \\
12 & 15 & 17,478,848 & 7,150,855 & 24,629,703 \\
12 & 37 & 17,478,848 & 6,965,673 & 24,444,521 \\
\hline
\end{array}
\]
4.3. Comments on RDC results

In general, RDC method has lower claim reserve estimation compare to Chain Ladder method. This is agree with some opinion that Chain Ladder method is over estimate. At the same maximum claim reported period, \( w_0 \), IBNR result is not change. IBNR result going down as \( w_0 \) increase. Within the same \( w_0 \), RBNS initially increase but at some quantiles it reaches its asymptotic value. As \( w_0 \) increases, RBNS result is also increase. As this model doesn’t count the effects of several background variables (rating factors) and the estimate is much lower than standard method (Chain Ladder), we may think that the result of standard RDC method is under estimate. Need a new method that count the effect of rating factors and adjust the estimate value from its base factor.

5. RDC Segmented Calculation

The basic idea of RDC segmented calculation method is, we calculate RDC claim estimation for every segment (i.e. every combination of background variable). In this case we will have response variable for every combination of background variable (excluding the zero combinations). We can calculate claim estimation directly from this result, but will give ”raw estimate”. We need a ”smoothing” method, to smooth the result from segmented calculation.

5.1. Bayesian Additive Regression Tree (BART) overview

BART is a Bayesian approach to nonparametric function estimation using regression trees. Regression trees rely on recursive binary partitioning of predictor space into a set of hyper-rectangles in order to approximate some unknown function \( f \) [5]. Predictor space has dimension of the number of variables. Tree-based regression models have an ability to flexibly interactions and nonlinearities. Models composed of sums of regression trees have an even greater ability than single trees to capture interactions and non-linearities as well as additive in \( f \). We choose package bartMachine from R library

5.2. RDC-BART Segmented Calculation

We divide the method into the following stages:

- **RDC segmentation** Group data into segments. Each segment represents unique combination of background variables. In this case we have \( 4 \times 3 \times 3 \times 4 = 144 \) combination of background variables. Then choose appropriate \( w_0 \) and \( q_0 \) and apply RDC method to get IBNR and RBNS estimate for each segment.

- **bartMachine** setup: clean data that was obtained from previous stage from NAs, build response vector (we take log of the reserve as continuous response) and predictor matrix (set of four categorical variables), setting Java heap (up to 5GB of RAM) and setting number of core used (we use 4 cores).

5.3. RDC-BART Segmented Calculation

BART model building: Setting hyperparameters (in our case) : \( m = 50, \alpha = 0.95, \beta = 2, k = 2, q = 0.9, \nu = 3 \). Set probabilities of the GROW/PRUNE/CHANGE steps to 28% / 28% / 44%. Set the number of burn-in Gibbs samples to 250 and number of post-burn-in samples to 1,000. Set the covariates to be equally important \textit{a priori}.

5.4. RDC-BART Segmented Calculation

R output

```r
# for IBNR
> bart_machine1
bartMachine v1.2.3 for regression
t training_data n = 709 and p = 26
```
5.5. **RDC-BART Segmented Calculation Steps**

- Find the best model
- Predict the reserve (IBNR and RBNS) based on the winning model, (R2)
- Choose a rating factor as a base factor. Here we choose Prov as it is most directly connected to the expected loss rather than other available rating factors.
- Calculate the reserve with data from segmented base factor only (R1). Here we use standard RDC method
- Calculate final reserve estimation as \( R_3 = R_2 \times \frac{\sum_{u=1}^{n} R_1(u)}{\sum_{u=1}^{n} R_2(u)} \) with \( n \) is the number of total period (12 in this case).

5.6. **Result**

| Month1 | Month2 | Month3 | Month4 | Month5 | Month6 | Month7 | Month8 | Month9 | Month10 | Month11 | Month12 | Total |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|---------|---------|---------|-------|
| CAD    |        |        |        |        |        |        |        |        |         |         |         |       |
| 3367946 | 3749727 | 2103851 | 2040325 | 1950846 | 1567745 | 5270556 | 3 | 1181286 | 6715422 | 5761650 | 1147126 | 1863662 | 19039721.9 |
| RBNS   |        |        |        |        |        |        |        |        |         |         |         |       |
| 877371,5 | 1447691 | 120357 | 138748 | 320957 | 52988,3 | 295371,2 | 421208,1 | 25099,4 | 40474,0 | 23598,86 | 331794,6 | 6778976,75 |
| Total   | 4243436,8 | 5197130 | 4222208 | 4416879 | 2232783 | 208843 | 1223330 | 2005293 | 92243,7 | 101095,7 | 56060,22 | 51817,7 | 2487408,7 |

So this method gives total reserve CAD 25,879,699 with \( w_0 = 3 \) and \( q_0 = 10 \)

6. **Summary**

RDC-BART Segmented Method gives estimate slightly greater than standard RDC but still lower than Chain Ladder method. RDC-BART Segmented Method depends on several factors, including \( q_0 \) and \( w_0 \), hyper-parameters, number of trees, BART winning model. RDC-BART Segmented Method also depends on base factor we choose. Some improvements could be made including adding inference, RMSE, MSEP, change the base to categorical level and better algorithm to shorten execution time.

7. **References**

[1] Rosenlund 2012, *Bootstrapping individual claim histories* (ASTIN Bulletin) vol 42 pp 291-324
[2] Godecharle and Antonio 2014, *Reserving by conditioning on markers of individual claims: a case-study using historical simulation* (KU Leuven Faculty of Economics and Business)
[3] Effendie, A.R., Pebriawan, R. 2017, *Estimation of IBNR and RBNS reserve by detailed conditioning method* (Far East Journal of Mathematical Sciences) vol 101(12), pp. 2785-2801
[4] Wüthrich and Merz 2008, *Stochastic claims reserving methods in insurance* (John Wiley & Sons)
[5] Chipman, H.A., George, E.I., McCulloch, R.E. 2010, *BART: Bayesian Additive Regression Trees* (The Annals of Applied Statistics) Vol 4 No. 1, pp 266-298