Eliciting claims development patterns and costs hidden in backlogs

Filip Lindskog, Stockholm University

Based on joint work with Mario Wüthrich forthcoming in European Actuarial Journal

Stockholm, September 17, 2025

Background facts

- Insurance claims are reported with random delays
- Each reported claim is labeled by occurrence period (with time unit e.g. a year, a quarter or a month) and development period (elapsed time units between start of occurrence period and reporting time)
- Based on what is known now the insurer wants to predict e.g. total ultimate number of claims (or cost) for each occurrence period
- Under idealized statistical assumptions this prediction problem is well studied (various approaches to claims reserving)

The standard setup

	0	1	2	3
0	$R_{0,0}$	$R_{0,1}$	$R_{0,2}$	$R_{0,3}$
1	$R_{1,0}$	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$
2	$R_{2,0}$	$R_{2,1}$	$R_{2,2}$	?
3	$R_{3,0}$	$R_{3,1}$	?	?
4	$R_{4,0}$	?	?	?

- Let $R_{i,j}$ be the number of claims labeled by occurrence period i and reported in development period j
- Predict future $R_{i,j}$s based on observed $\{R_{i,j} : i + j \leq 4, j \geq 0\}$
- Assuming independence between rows and e.g. assumptions like $R_{i,j+1}/R_{i,j} \approx R_{i',j+1}/R_{i',j}$ (stable development patterns) this prediction problem is well understood

A statistical challange

- A reported (R) claim is processed (P) without delay if there is sufficient processing capacity (C), otherwise it contributes to a backlog (B) of claims waiting to be processed
- Data available as basis for prediction are typically number of claims $P_{i,j}$ processed by claim handlers
- If there is unlimited capacity, then $P_{i,j} = R_{i,j}$
- The fact that often $P_{i,j} \neq R_{i,j}$ leads to statistical challenges
- Backlogs distort data, making numbers of processed claims $P_{i,j}$ unsuitable as the input to standard prediction methods
- How can we predict future $P_{i,j}$ if we do not have access to undistorted variables $R_{i,j}$?

Limited claims processing capacity - facts

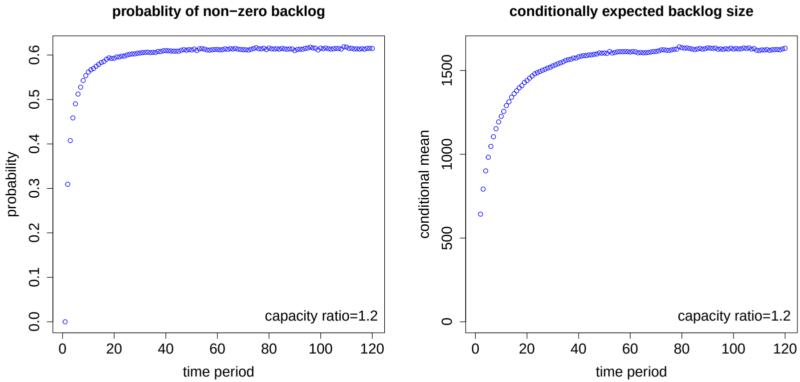
- $P_t = P_{t,0} + P_{t-1,1} + \dots$ and similarly for B_t and R_t
- $P_t = \min(B_t + R_t, C_t)$ (total number of processed claims)
- $B_{t+1} = \max\left(B_t + R_t - C_t, 0\right)$ (total number of forwarded backlog claims, a Lindley recursion)
- $B_{t+1} = B_t + R_t - P_t$
- Claims are labeled by occurrence and development period: $R_{i,j}$, $P_{i,j}$, $B_{i,j}$ satisfy $B_{i,j+1} = B_{i,j} + R_{i,j} - P_{i,j}$ and $B_{i,0} = 0$
- Claims for different occurrence periods share processing capacity according to a claims processing protocol (e.g. first come first served)

Numerical illustration

- $\mathbb{E}[R_{i,0}] = 500$, $\mathbb{E}[R_{i,1}] = 300$, $\mathbb{E}[R_{i,2}] = 150$, $\mathbb{E}[R_{i,3}] = 50$, and $R_{i,j} \equiv 0$ for $j \geq 4$.
- The $R_{i,j}$ are simulated as independent $\text{NBin}(\alpha_j, \beta)$,

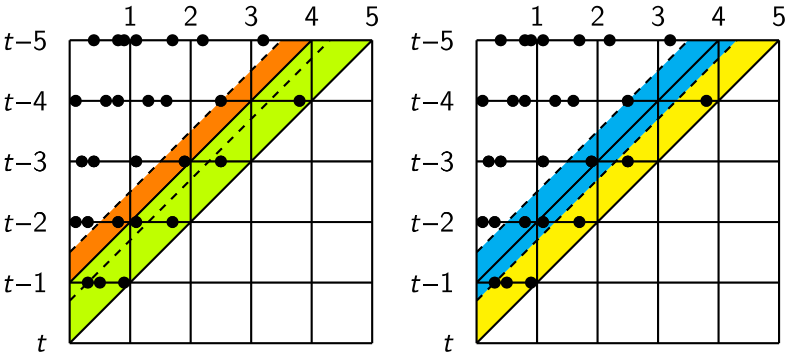
$$\frac{\sqrt{\text{Var}(R_{i,0} + \cdots + R_{i,3})}}{\mathbb{E}[R_{i,0} + \cdots + R_{i,3}]} = 0.7$$

- The capacities $C_t = 1200$ are all equal to 120% of the expected number of newly reported claims in each time period.
- Each reported claim is assigned a time stamp, uniformly distributed in the unit time interval to which it belongs.
- Claims processing according to first come first served.



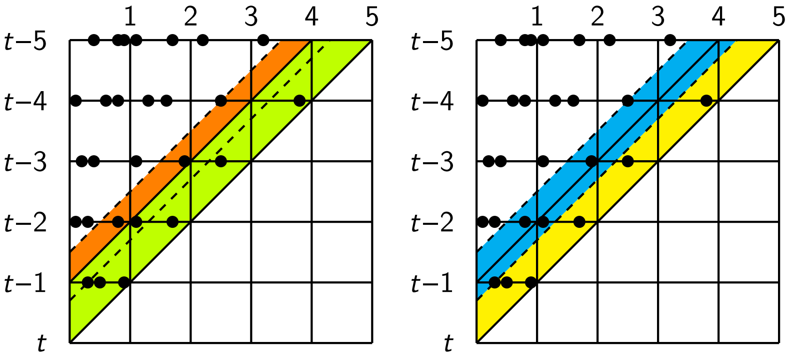
Estimates of $\mathbb{P}(B_t > 0 \mid B_1 = 0)$ and $\mathbb{E}[B_t \mid B_t > 0, B_1 = 0]$

Illustration: first come first served



$B_t = 2$: points in the orange region, $R_t = 7$: points in the lime region, $P_t = \min(B_t + R_t, C_t) = 4$: points in the blue region, $B_{t+1} = B_t + R_t - P_t = 5$: points in the yellow region.

Illustration: first come first served



Here $R_t = 7$ points in the lime region satisfies
$R_t = R_{t,0} + R_{t-1,1} + R_{t-2,2} + R_{t-3,3} + R_{t-4,4} = 3 + 2 + 1 + 1 + 0$
Here $P_t = 4$ points in the blue region satisfies
$P_t = R_{t,0} + R_{t-1,1} + R_{t-2,2} + R_{t-3,3} + R_{t-4,4} = 1 + 2 + 1 + 0 + 0$

The prediction problem

	0	1	2	3
0	$P_{0,0}$	$P_{0,1}$	$P_{0,2}$	$P_{0,3}$
1	$P_{1,0}$	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$
2	$P_{2,0}$	$P_{2,1}$	$P_{2,2}$	?
3	$P_{3,0}$	$P_{3,1}$	?	?
4	$P_{4,0}$	?	?	?

- Let $P_{i,j}$ be the number of claims labeled by occurrence period i and processed in development period j
- Predict $\{P_{i,j} : i + j > \tau, j \geq 0\}$ based on

$$\{P_{i,j} : i + j \leq \tau, j \geq 0\} \cup \{B_s : s \leq \tau + 1\}$$

i.e. based on numbers of processed claims and total backlog sizes observable at time τ of prediction.

The prediction problem

- If we had access to $\{R_{i,j}, P_{i,j} : i + j \leq \tau, j \geq 0\}$, then by choosing a predictor

$$\sum_{j=\tau-i+1}^{\infty} \widehat{R}_{i,j} \quad \text{together with identity} \quad \sum_{j=0}^{\infty} R_{i,j} = \sum_{j=0}^{\infty} P_{i,j}$$

would naturally give the predictor

$$\sum_{j=\tau-i+1}^{\infty} \widehat{P}_{i,j} = \sum_{j=0}^{\tau-i} R_{i,j} + \sum_{j=\tau-i+1}^{\infty} \widehat{R}_{i,j} - \sum_{j=0}^{\tau-i} P_{i,j}$$

- However, $\{R_{i,j} : i + j \leq \tau, j \geq 0\}$ is not observable.
- Find a way to predict/estimate $\{R_{i,j} : i + j \leq \tau, j \geq 0\}$ from $\{P_{i,j} : i + j \leq \tau, j \geq 0\} \cup \{B_s : s \leq \tau + 1\}$

	processing $P_{i,j}$					backlog $B_{i,j}$				
	0	1	2	3	4	0	1	2	3	4
1	362	946	371	36	0	0	1252	341	36	0
2	0	362	481	2	1	0	68	481	0	1
3	164	462	418	56	0	0	285	364	56	0
4	220	627	41	0	0	0	590	40	0	0
5	153	174	5	222	109	0	174	0	0	109
6	208	175	0	0	0	0	0	0	0	0
7	620	379	254	116	1	0	0	189	116	1
8	599	397	135	259	175	0	323	127	434	175
9	440	922	167	136	74	0	921	263	208	74
10	27	773	524	41	265	0	1288	551	44	265
11	0	365	826	28	0	0	1013	854	28	0
12	0	259	97	8	0	0	305	83	8	0
13	0	436	204	0	0	0	193	203	0	0
14	374	934	421	0		0	311	202	0	0
15	54	53	38			0	17	0	0	
16	572	232				0	0	0		
17	254					0	0			

Shared processing capacity - first come first served

- In time period t, for each occurrence period i,
  - reported claims in the backlog are processed first, generating $P^B_{i,t-i}$ processed claims,
  - followed by processing of newly reported claims, generating $P^R_{i,t-i}$ processed claims
- $P_{i,t-i} = P^B_{i,t-i} + P^R_{i,t-i}$
- $P^B_{i,t-i} \mid B_{i,t-i}, B_t, P_t \sim \text{Bin}\left(B_{i,t-i}, \mathbb{1}_{\{B_t \leq P_t\}} + \frac{P_t}{B_t}\mathbb{1}_{\{B_t > P_t\}}\right)$
- $P^R_{i,t-i} \mid R_{i,t-i}, B_t, R_t, P_t \sim \text{Bin}\left(R_{i,t-i}, \frac{P_t - B_t}{R_t}\mathbb{1}_{\{B_t \leq P_t \leq B_t + R_t\}}\right)$
- At time t, the actuary has access to B_t, B_{t+1}, P_t and therefore also R_t via $B_{t+1} = B_t + R_t - P_t$
- At time t, the actuary has access to $P_{i,t-i}$ but not $B_{i,t-i}$ and $R_{i,t-i}$

Shared processing capacity - first come first served

With σ-algebras

$$\mathcal{F}_t = \sigma(R_s, B_s, P_s : s \leq t), \quad \mathcal{B}_t = \sigma(B_{i,j} : i + j \leq t, j \geq 0),$$
$$\mathcal{R}_t = \sigma(R_{i,j} : i + j \leq t, j \geq 0), \quad \mathcal{P}_t = \sigma(P_{i,j} : i + j \leq t, j \geq 0),$$

we therefore get

$$\begin{aligned}&\mathbb{E}\big[P_{i,t-i} \mid \mathcal{F}_t \vee \mathcal{B}_t \vee \mathcal{R}_t\big] \\ &\quad = B_{i,t-i}\left(\mathbb{1}_{\{B_t \leq P_t\}} + \frac{P_t}{B_t}\mathbb{1}_{\{B_t > P_t\}}\right) + R_{i,t-i}\left(\frac{P_t - B_t}{R_t}\mathbb{1}_{\{B_t \leq P_t\}}\right)\end{aligned}$$

where $B_{i,t-i} = \mathbb{1}_{\{t-i>0\}} \sum_{j=0}^{t-i-1}(R_{i,j} - P_{i,j})$. Note that

$$\mathbb{E}\big[P_{i,t-i} \mid \mathcal{F}_t \vee \mathcal{B}_t \vee \mathcal{R}_t\big] = \mathbb{E}\big[P_{i,t-i} \mid \mathcal{F}_t \vee \mathcal{P}_{t-1} \vee \mathcal{R}_t\big]$$

is linear in $R_{i,j}$s

Estimating unknown reportings

- Write $\ell_{i,t-i}(R_{i,0},\dots,R_{i,t-i}) := \mathbb{E}\big[P_{i,t-i} \mid \mathcal{F}_t \vee \mathcal{P}_{t-1} \vee \mathcal{R}_t\big]$
- Equations

$$P_{i,t-i} = \ell_{i,t-i}(R_{i,0},\dots,R_{i,t-i}) + \varepsilon_{i,t-i}, \quad t \le \tau, i \le t$$

- Solving

$$\min \sum_{t\le\tau}\sum_{i\le t}\Big(P_{i,t-i} - \ell_{i,t-i}(r_{i,0},\dots,r_{i,t-i})\Big)^2$$

$$\text{subject to } r_{i,j} \ge 0, \quad \sum_i r_{i,t-i} = R_t$$

  yields estimates $\widehat{r}_{i,j}$ of unseen $R_{i,j}$
- Refinement by considering conditional variance is possible

	true reporting $R_{i,j}$					estimated reporting $\widehat{R}_{i,j}$				
	0	1	2	3	4	0	1	2	3	4
1	1614	35	66	0	0	1606	51	58	0	0
2	68	775	0	3	0	70	772	3	0	1
3	449	541	110	0	0	448	597	0	55	0
4	810	77	1	0	0	794	95	0	0	0
5	327	0	5	331	0	322	5	6	332	0
6	208	175	0	0	0	208	175	0	0	0
7	620	568	181	1	0	620	575	174	1	0
8	922	201	442	0	0	920	205	434	120	0
9	1361	264	112	2	0	1354	285	111	35	0
10	1315	36	17	262	0	1306	146	0	145	85
11	1013	206	0	0	0	592	634	0	0	0
12	305	37	22	0	0	276	71	18	0	0
13	193	446	1	0	0	189	451	0	0	0
14	685	825	219	0		683	826	219	0	
15	71	36	38			71	36	38		
16	572	232				572	232			
17	254					254				

Summary

- Numbers of processed claims are identified as output from a queueing system with limited shared service capacity
- Statistical challenges are due to labeled input and output (labeling by occurrence and development period)
- An actuary's prediction problem corresponds to predicting the output without observing the input or service capacity
- We propose a prediction method based on estimating unseen input that performs well
- The setup allows for claims cost optimization by treating the capacity as a control variable