Market-based insurance ratemaking

Pierre-O. Goffard
(Joint work with G. Peters and P. Piette)

Université de Strasbourg
goffard@unistra.fr

Insurance Data Science

DIALog
Digital insurance and long term risk
Chaire d'Excellence

[redacted]

Vet Bill.
Oil on canvas.

[redacted]

Risk model

The medical expenses of a pet, over a year, are given by

$$X = \sum_{i=1}^{N} U_i$$

where

- N is the number of visit to the vet
- U_i are the associated expenses

Pure premium

The insurance company offer a coverage of the risk

$$g(X) = \min[\max(r \cdot X - d, 0), l],$$

where

- r is the coverage rate
- d is the deductible
- l is the limit

The pure premium is given by

$$p = \mathbb{E}[g(X)].$$

Using historical data $\Rightarrow \widehat{p}$

Commercial premium

Expectation principle

The final quote is given by

$$\widetilde{p} = f(p) \geq p$$

For instance,

$$\widetilde{p} = (1+\eta)p, \; \eta > 0$$

Market data

Let $\mathscr{D} = \{\widetilde{p}_1, \ldots, \widetilde{p}_n\}$ be a collection of insurance quotes with

$$\widetilde{p}_i = f_i\left\{\mathbb{E}\left[g_i(X)\right]\right\},\ i = 1, \ldots, n,$$

where

- The loading functions f_i are unknown
- The insurance coverages g_i are known
- The risk X is parametrized by an unknown parameter θ.

Optimization problem

Find $\theta \in \Theta \subset \mathbb{R}^d$ and $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ to minimize

$$d\left[\widetilde{p}_{1:n}, f\left(p^{\theta}_{1:n}\right)\right],$$

where

- $p_i^{\theta} = \mathbb{E}_{\theta}\left[g_i(X)\right],$ for $i = 1, \ldots, n$, associated to X parametrized by θ
- f is applied elementwise on $p^{\theta}_{1:n}$
- $d(\cdot,\cdot)$ distance function over the observation space

Subject to

$$\widetilde{p}_i \geq p_i^{\theta}, \text{ and } f\left(p_i^{\theta}\right) \geq p_i^{\theta}, \text{for } i = 1, \ldots, n.$$

ABC algorithm

1. Sample a parameter value
$$\theta \sim \pi(\theta)$$
2. Compute
$$p_i^\theta = \mathbb{E}_\theta\left[g_i(X)\right] \text{ for } i = 1, \ldots, n$$
3. Fit an isotonic regression model
$$\widetilde{p}_i \sim f(p_i^\theta) \text{ for } i = 1, \ldots, n$$
4. If
$$d\left[\widetilde{p}_{1:n}, f\left(p_{1:n}^\theta\right)\right] < \epsilon$$
then we store θ and f.

Repeat the procedure to have many parameter values and build
$$\pi_\epsilon(\theta|\widetilde{p}_{1:n}),$$
the proxy of the posterior distribution.

Isotonic regression

How to learn the link between pure and commercial premium?

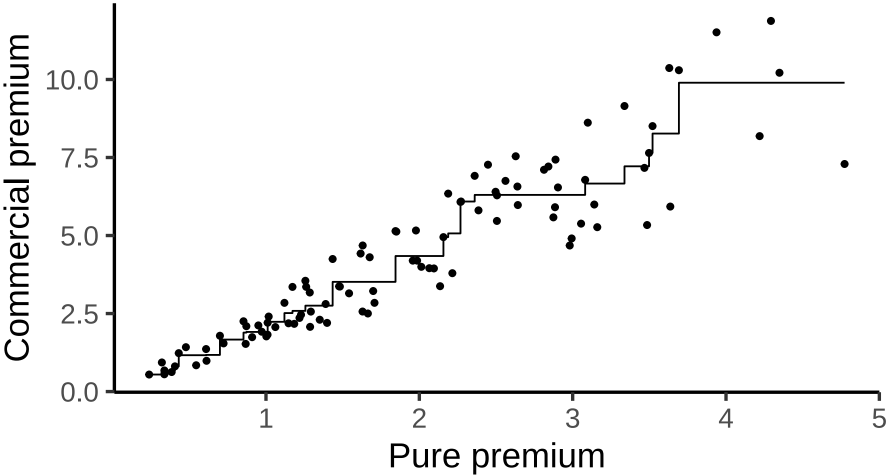
- The data is made of pairs $(p_i, \widetilde{p}_i)_{i=1,\dots,n}$
- Assume that $p_1 < \ldots < p_n$
- Find $\widetilde{p}_1^*, \ldots, \widetilde{p}_n^*$ that minimize

$$\sum_{i=1}^{n} (\widetilde{p}_i^* - \widetilde{p}_i)^2, \text{ subject to } \widetilde{p}_i^* \leq \widetilde{p}_j^* \text{ whenever } p_i \leq p_j.$$

- The link is estimated by

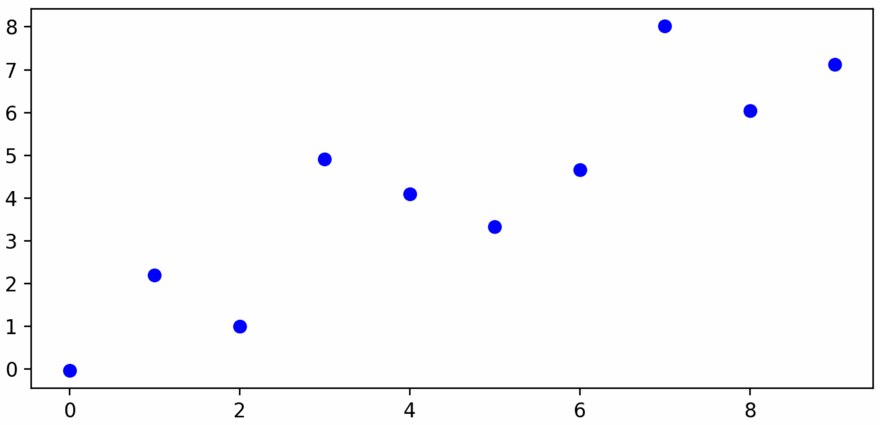
$$f(p) = \sum_{i=1}^{n} \widetilde{p}_i^* \mathbb{I}_{[\widetilde{p}_i^*, \widetilde{p}_{i+1}^*]}(p)$$

Isotonic regression $\Rightarrow \widetilde{p}_1^*, \ldots, \widetilde{p}_n^* \Rightarrow f$



Commercial premium
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Pure premium

PAVA Algorithm



Non identifiability

Ill posed inverse problem

⚠ Many parameters can be equally good

We address this issue by

- adding a regularization term so that

$$\text{LR}_i = \frac{p_i}{\widetilde{p}_i} \in [\text{LR}_{\text{low}}, \text{LR}_{\text{high}}] \text{ for } i = 1, \dots, n$$

- Using a particle based optimization algorithm

Data

1,080 pet insurance quotes from 5 insurance companies

Variable	Description	Example
specie	Specie of the pet	dog
breed	Breed of the pet	australian sheperd
gender	Gender of the pet	female
insurance_carrier	identification number of the insurance company	1
age	Age of the pet (in years)	4 years
r	Value of the coverage rate	0.6
l	Value of the limit of the insurance coverage	1100
d	Value of the deductible of the insurance coverage	0
x	Yearly commercial premium	234.33

Table: List of the variables of our dataset

Several risk classes / One model

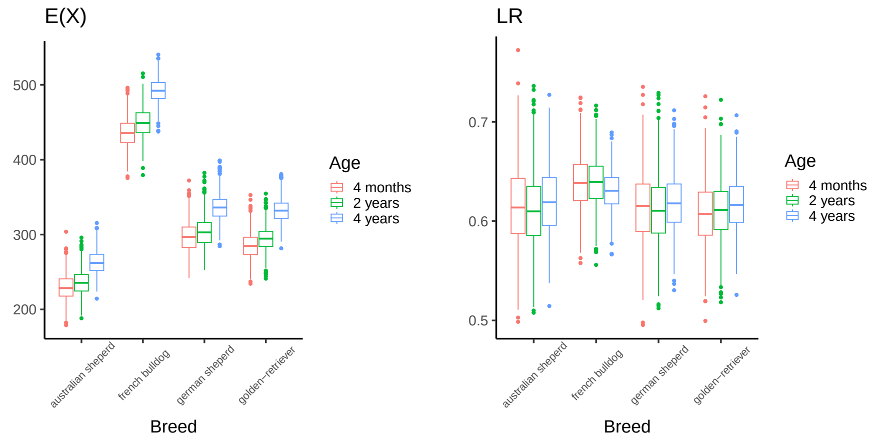
Risk class #	specie	breed	gender	age
1	dog	australian sheperd	female	4 months
2	dog	australian sheperd	female	2 years
3	dog	australian sheperd	female	4 years
4	dog	french bulldog	female	4 months
5	dog	french bulldog	female	2 years
6	dog	french bulldog	female	4 years
7	dog	german sheperd	female	4 months
8	dog	german sheperd	female	2 years
9	dog	german sheperd	female	4 years
10	dog	golden-retriever	female	4 months
11	dog	golden-retriever	female	2 years
12	dog	golden-retriever	female	4 years

Table: The 12 risk classes under study

We fit the $\text{Pois}(\lambda) - \text{LogNorm}(\mu, \sigma = 1)$ with prior settings

$$\lambda \sim \text{Unif}([0,10]), \text{ and } \mu \sim \text{Unif}([-10,10]).$$

Risk classes comparison



Limitations

- Model misspecification
- Premium principle misspecification
- Heterogeneous data
- Computing time
- Identifiability issue

Conclusion

- `IsoPriceR` package
- Theoretical study
- Model selection
- Other applications than pet insurance
- Market data + historical data

  $\hookrightarrow$ credibility framework

The preprint is avalable at

Pierre-Olivier Goffard, Pierrick Piette, and Gareth W. Peters.
Market-based insurance ratemaking: Application to pet insurance.
ASTIN Bulletin, pages 1–24, April 2025.

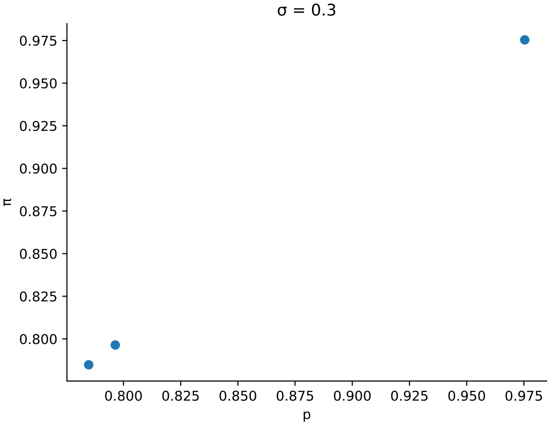
Model identifiability?

Let $X \sim \text{LogNorm}(\sigma = 0.3)$ and consider 3 insurance coverages

$$g_r(x) = 0.75 \cdot X, g_d(x) = \max(X - 0.25, 0), \text{ and } g_l(x) = \min(X, 1.5)$$

We have

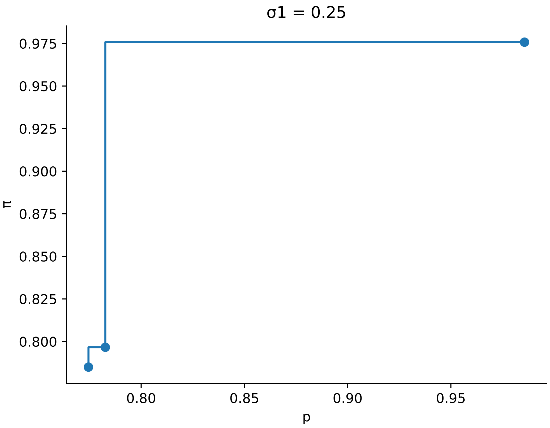
$$p_r = 0.78 < p_d = 0.79 < p_l = 0.97$$



$\sigma_1 = 0.25$

Let $X \sim \text{LogNorm}(\sigma_1 = 0.25)$, we have

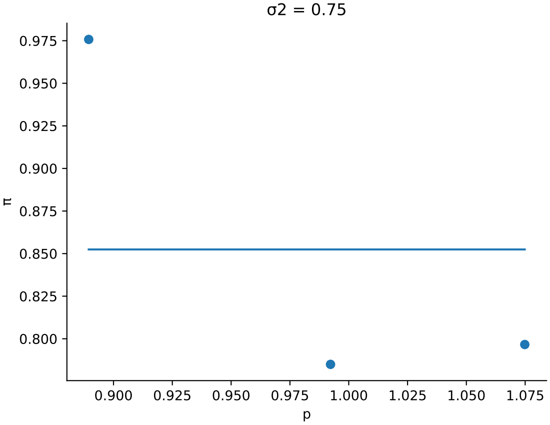
$$p_r = 0.77 < p_d = 0.78 < p_l = 0.98$$



$\sigma_2 = 0.75$

Let $X \sim \text{LogNorm}(\sigma_2 = 0.75)$, we have

$$p_d = 1.07 > p_r = 0.99 > p_l = 0.89$$



Population Monte Carlo Algorithm

Algorithm ABC-PMC

set $\epsilon_0 = \infty$ and $\pi_{\epsilon_0}(\theta|\mathscr{D}) = \pi(\theta)$
for $g = 1$ to G **do**
  for $j = 1$ to J **do**
    repeat
      generate $\theta^* \sim \pi_{\epsilon_{g-1}}(\theta|\mathscr{D})$
      compute $p_i^{\theta^*} = \mathbb{E}_{\theta^*}[g_i(X)]$, for $i = 1, \ldots, n$
      fit the isotonic regression model $\widetilde{p}_i = f(p_i^{\theta^*}) + e_i$, for $i = 1, \ldots, n$
      compute $d\left[\widetilde{p}_{1:n}, f\left(p_{1:n}^{\theta^*}\right)\right]$.
    until $d\left[\widetilde{p}_{1:n}, f\left(p_{1:n}^{\theta^*}\right)\right] < \epsilon_g$
    set $\theta_j^g = \theta^*$ and $d_j^g = d^*$
  end for
  find $\epsilon_g \le \epsilon_{g-1}$ so that $\widehat{\text{ESS}} = \left[\sum_{j=1}^J \left(w_j^g\right)^2\right]^{-1} \approx J/2$, where $w_j^g \propto \frac{\pi(\theta_j^g)}{\pi_{\epsilon_{g-1}}(\theta_j^g|\mathscr{D})} \mathbb{I}_{d_j < \epsilon_g}$
  compute $\pi_{\epsilon_g}(\theta|\mathscr{D}) = \sum_{j=1}^J w_j^g K_H(\theta - \theta_j^g)$
end for