Data Analytics for Non-Life Insurance assignment

Group 7

Eitas Rimkus (SNR: 2070805) Ids de Vries (SNR: 2074298)
Nikodem Baehr (SNR: 2076515)

2023-10-08

Contents

Introduction

Oxford English Dictionary defines insurance as 'a contract by which the one party (usually a company or corporation) undertakes, in consideration of a payment (called a premium) proportioned to the nature of the risk contemplated, to secure the other against pecuniary loss, by payment of a sum of money in the event of destruction of or damage to property (as by disaster at sea, fire, or other accident)[...]'. Car insurance protects the owner from loss, whether it's monetary or health damage, in a case of an unfortunate event involving the car. Since cars have become much more affordable around the World War II, more and more people started buying them. For many of us, the cars have become one of the necessities, essentials in our lives. The car industry has been booming as the global automotive manufacturing market was worth about 2.85 trillion USD in 2021 and 81.6 million units were sold in 2022 (Carlier, 2023). Given the increasing amount of cars, the amount of accidents also increases, making the car insurance market a very interesting and busy one. Driving a car can be dangerous and the damages to both the car and the passengers substantial. A split second may decide if the accident occurs or not. With people driving recklessly on the roads, being distracted or the conditions being harsh in certain situations, people want to insure themselves in a case of an accident. Given there is a certain degree of an event happening, the car insurance market has developed in order to make profits. In order to be profitable, the companies within the market have collected vast amounts of data given the accidents, their reasons, frequencies, damages severity etc. Once the data is collected a company can come up with a certain premium bearing in mind competitiveness and proportional to the probability of an event happening as well as being able to cover all their liabilities. In Europe in 2020, the motor insurance market brought in €149 billion in revenue from premiums whilst the companies lost a total of €97 billion in the claims paid, resulting in profit of around €52 billion (Insurance Europe, 2022). The car insurance market takes up around a third (36%) of the whole insurance market in Europe.

In this paper, we will compare the accuracy of estimates of the claim amount for non-zero claims in a collective model using 2 different methods: compound Poisson with normal approximation and compound Poisson with Monte Carlo simulation. We will discuss the differences between the two models, their assumptions and usage. We will also compare the results and interpret them.

Research Question

In this paper, we will answer the following two research questions:

1) To what extent do the results of normal approximation and Monte Carlo simulation differ for calculating the Capital at Risk values?

2) How can we apply normal approximation for calculating initial capital value as well as the optimal alpha for proportional reinsurance?

Data description and analysis

The analyzed data set contains 7 columns: Calendar Year, volume of the policy in years, indicator for nonzero claims, size of the claim, fuel type, gender, and annual number of miles driven. It represents a cross-section of previous claim amounts between year 2000 and 2022.

Claim Size

In total, there are 570 000 observations. Only 4.1% of the claims are strictly positive. Considering the distribution of nonzero claims (Figure 1), it can be seen that its distribution is slightly skewed to the right.

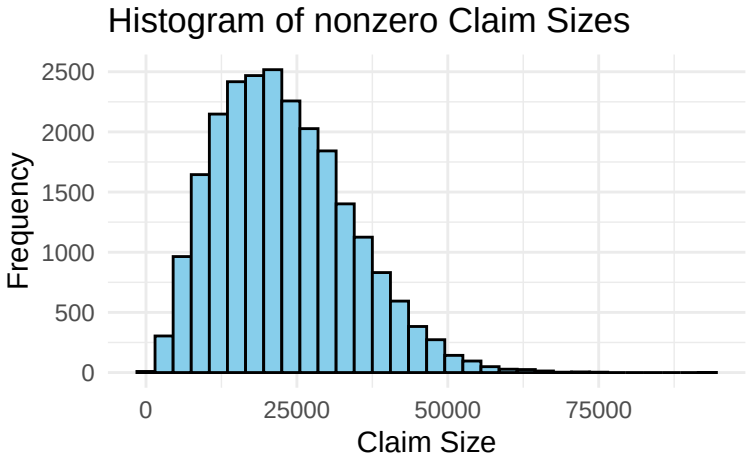

Figure 1: Histogram of non-zero claim sizes between 2000 and 2022

Moreover, the total claim size amount differs by year (Figure 2). However, there is no trend of an increase or decrease throughout the years. Hence, it can be assumed that the total claim size amount per year is rather random.

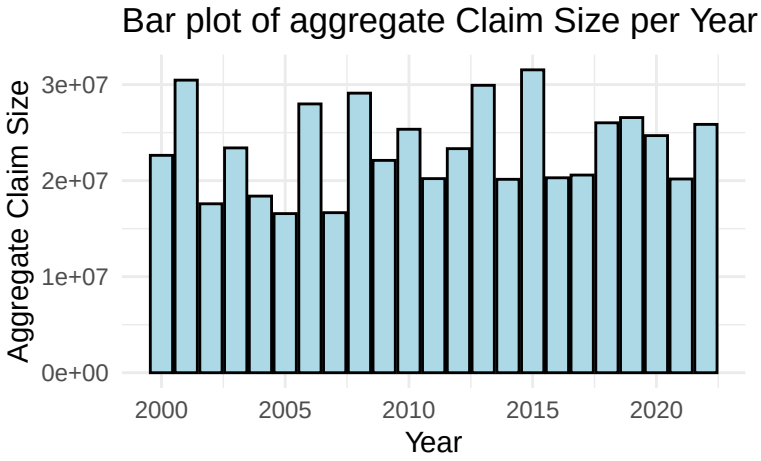

Figure 2: Total claim size per year between 2000 and 2022

The difference between minimum and maximum non-zero claim size is big, especially when the 75% of all nonzero claims is smaller than 29923, which is 3 times as small as the maximum value of the non-zero claim size (Table 1).

Table 1: Description of non-zero claim size distribution

Metric	Value
Minimum value	393
1st Quartile	14516
Median	21630
3rd Quartile	29923
Maximum	93436

Volume

The aggregate volume amount per year, as seen in Figure 3, is nearly constant during 2020-2022. Therefore, the stationarity assumption in the expected claims frequency is not refuted, so $\hat{\lambda}$ is constant.

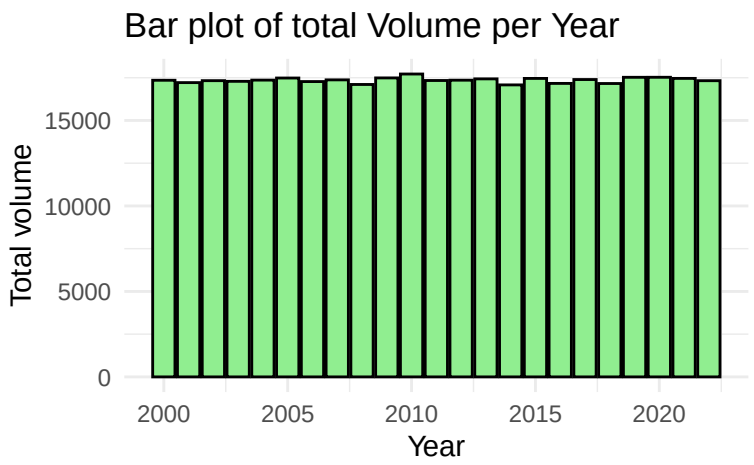

Figure 3: Total amount of volume per year between 2000 and 2022

Gender

The full data set has around 60% of all incidents caused by male (the same proportion applies if we consider only nonzero claims). If we split claim sizes into two categories: large claims (claim sizes greater than 40 000, and small claims (less than or equal to 40 000), one can see that in group of large claims, the aggregate amount of claim sizes, caused by men (Figure 4), is twice as large as for women, whereas for small claims, the difference is not that significant (Figure 4).

Type of Petroleum

In the full data set there are approximately 40% of claims caused by insureds driving diesel cars (same percentage for nonzero claim sizes as well). Figure 5 shows that there is an immense difference in the type of fuel causing small and large claims. Diesel cars are much more prevalent in large claims.

Mileage

There is a slight trend in claim sizes being larger as the mileage gets larger (Figure 6), meaning that there is more risk of car causing an incident if the annual miles driven is larger.

Figure 4: Claim Size by gender and claim size group

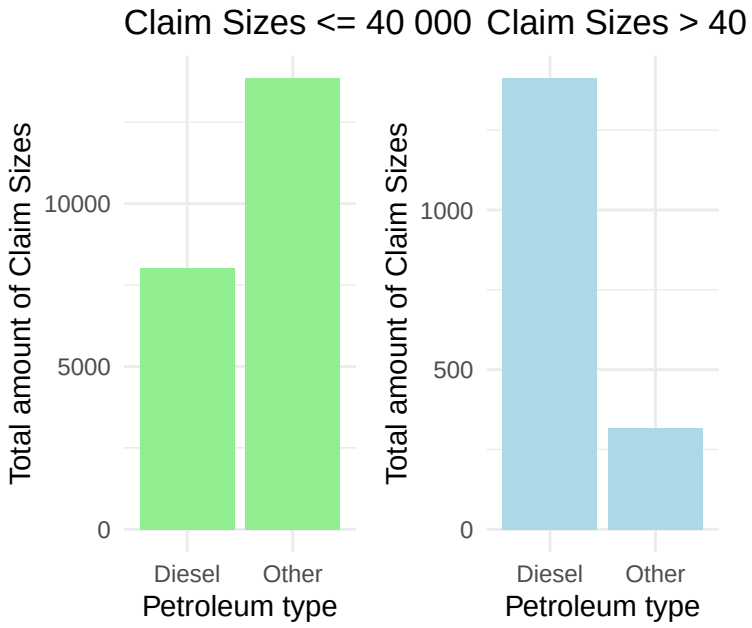

Figure 5: Claim Size by petroleum type and claim size group

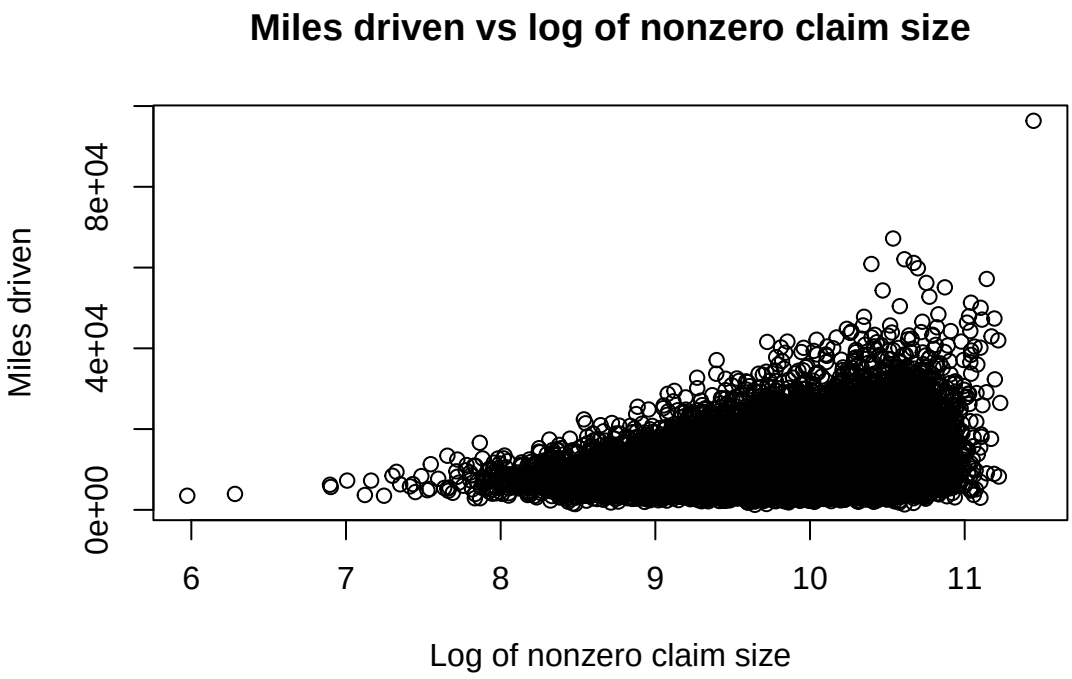

Figure 6: Log nonzero claim size and mileage

Models and Assumptions

To answer the research question, we will need to make use of the collective risk model. Let S be the total claim amount over a one year period, where $S = Y_1 + \cdots + Y_N = \sum_{i=1}^{N} Y_i$ with Y_i the individual claim size and N an integer random variable that counts the number of claims that occur over the one year period. We assume that $Y_1, \ldots, Y_N$ are i.i.d and distributed with function G, with $G(0) = 0$. Furthermore we assume that N and $(Y_1, \ldots, Y_N)$ are independent.

Let v be the fixed, positive volume and λ be the fixed, positive expected claims frequency. Then $N \sim \mathsf{Poi}(\lambda v)$ with properties:

$$\mathbb{E}[N] = \lambda v = Var(N)$$

So for the collective risk model, the total claim amount S has a Compound Poisson Distribution.

$$\Rightarrow S \sim \mathsf{ComPoi}(\lambda v, G)$$

For the individual claim size distribution G. Then S has the following properties: 1) $E[S] = E[N] * E[Y_1] = \lambda v E[Y_1]$; 2) $Var(S) = \lambda v E[Y_1^2]$.

Methods

Estimating s-values

To estimate the values of s, such that $\mathrm{P}(S \leq s) = c$ for c= {0.9, 0.95, 0.99}, we will be applying the normal approximation to the compound poisson collective model firstly, $S \sim \mathsf{CompPoi}(\lambda v, G)$. From the CLT we know that as $v \to \infty$:

$$\frac{S - \mathbb{E}(S)}{\sqrt{Var(S)}} = \frac{S - \lambda v\, \mathbb{E}(Y_1)}{\sqrt{\lambda v\, \mathbb{E}(Y_1^2)}} \to N(0,1)$$

Here, $\mathbb{E}(Y_1)$ is the expected claim size. With the available data, we can calculate $\mathbb{E}(Y_1)$ and $\mathbb{E}(Y_1^2)$. After this, we can derive an expression for s. The CLT implies

$$\mathrm{P}(S \leq s) = \mathrm{P}(\frac{S - \lambda v\, \mathbb{E}(Y_1)}{\sqrt{\lambda v\, \mathbb{E}(Y_1^2)}} \leq \frac{s - \lambda v\, \mathbb{E}(Y_1)}{\sqrt{\lambda v\, \mathbb{E}(Y_1^2)}}) \approx \Phi(\frac{s - \lambda v\, \mathbb{E}(Y_1)}{\sqrt{\lambda v\, \mathbb{E}(Y_1^2)}})$$

So

$$\begin{aligned} &\mathrm{P}(S \leq s) = c \\ \Rightarrow\ &\Phi(\frac{s - \lambda v\, \mathbb{E}(Y_1)}{\sqrt{\lambda v\, \mathbb{E}(Y_1^2)}}) = c \\ \Rightarrow\ &\frac{s - \lambda v\, \mathbb{E}(Y_1)}{\sqrt{\lambda v\, \mathbb{E}(Y_1^2)}} = \Phi^{-1}(c) \\ \Rightarrow\ &s = \sqrt{\lambda v\, \mathbb{E}(Y_1^2)} * \Phi^{-1}(c) + \lambda v\, \mathbb{E}(Y_1) \end{aligned}$$

for every value of c={0.9, 0.95, 0.99}. After calculating the values of n, $\mathbb{E}(Y_1)$, $\mathbb{E}(Y_1^2)$, $\lambda \approx \hat{\lambda}_{MLE} = \frac{1}{v}\sum_{t=1}^{n} N_t$ and $v = \sum_{i=1}^{n} v_i$ we can substitute that in the equation of s to derive the expressions for $s_{0.9}, s_{0.95}$ and $s_{0.99}$.

Secondly, we want to estimate the values of s in $\mathrm{P}(S \leq s) = c$ by means of a Monte Carlo simulation in order to check the validity of our estimations made in the first part. Let J denote the number of scenarios in the Monte Carlo simulation, and let $j = 1, \ldots, J$. We will start with using $J = 10000$ observations, but this number can easily be adapted. Assumed is that we draw the number of claims for each scenario j out of a Poisson distribution, so $N \sim \mathsf{Poi}(\lambda v)$. Furthermore we assume that the random variable Y_i for every observation i is Gamma distributed, $Y_i \sim \mathsf{Gamma}(\gamma, 1/c)$ where γ and c are estimated by

$$\hat{\gamma} = \frac{\mathbb{E}(Y_{dataset}^2)}{Var(Y_{dataset})}, \hat{c} = \frac{\mathbb{E}(Y_{dataset})}{Var(Y_{dataset})}$$

After drawing from the Poisson distribution, for each observation we will draw from the Gamma distribution to determine a value of $S = \sum_{i=1}^{n} Y_i$ and use this to estimate the $s_{0.9}, s_{0.95}$ and $s_{0.99}$ quantiles.

Proportional Reinsurance

To determine the amount of reinsurance, we first need to know what our initial capital u is, given that we have safety load $\theta = 0.05$ and we aim for a ruin probability of 1%.

Assume that without reinsurance the ruin probability is $P[S > u + \pi] = 0.01$. We determine the u using the normal approximation, i.e.

$$P[S > u + \pi] = P\left[\frac{S - E[S]}{\sqrt{Var(S)}} > \frac{u + \pi - E[S]}{\sqrt{Var(S)}}\right] \approx 1 - \Phi\left(\frac{u + \pi - E[S]}{\sqrt{Var(S)}}\right)$$

Let the premium $\pi = (1+\theta)E[S]$. Then, since the ruin probability is equal to 0.01, we have

$$1 - \Phi\left(\frac{u+\theta E[S]}{\sqrt{Var(S)}}\right) = 0.01 \iff \frac{u+\theta E[S]}{\sqrt{Var(S)}} = \Phi^{-1}(0.99)$$

After reworking the equation for u, we obtain

$$u = \Phi^{-1}(0.99)\sqrt{Var(S)} - \theta E[S]$$

Since $S \sim \mathsf{CompPoi}(\lambda v, G)$, the expectation of S is $E[S] = \lambda v E[Y_1]$ and the variance is $Var(S) = \lambda v E[Y_1^2]$, where $Y_i \sim \mathsf{G}$ is the random variable of an individual claim size, we can simplify the calculation of u to

$$u = \Phi^{-1}(0.99)\sqrt{\lambda v E[Y_1^2]} - \theta\lambda v E[Y_1]$$

Now for the reinsurance, consider that we want to reduce the default probability to 0.5% by a proportional reinsurance contract of all policies. Let $\alpha \in [0,1]$ be the proportionality factor, such that $S^c = \alpha S$ and $S^r = (1-\alpha)S$. We may assume that $\xi = 0.07$. For the premiums we have $\pi^r = (1+\xi)(1-\alpha)E[S]$ and $\pi^c = \pi - \pi^r = [1+\theta-(1+\xi)(1-\alpha)]E[S]$. We are interested in finding the optimal α.

Given the initial capital u, we need to maximize the expected profit of the cedent under the constraint that $P[S^c > u + \pi^c] \leq \epsilon \ \forall \epsilon > 0$, i.e.

$$\max E[B^c] = \pi^c - E[S^c] = [\theta - \xi + \xi\alpha]E[S]$$

$$P[U^c < 0] = P[S^c > u + \pi^c] \leq \epsilon$$

Since $\xi > 0$, the expected profit function $E[B^c]$ is increasing in α, hence the maximization problem becomes to find maximal α such that

$$P[S^c > u + \pi^c] \leq \epsilon \iff P[S^c > u + \pi^c] \leq 0.005$$

Using the CLT, and the fact that the profit function of cedent is $E[B^c] = \pi^c - E[S^c]$, we obtain

$$P\left[\frac{S^c - E[S^c]}{\sqrt{Var(S^c)}} > \frac{u + \pi^c - E[S^c]}{\sqrt{Var(S^c)}}\right] \leq 0.005$$

Define $\alpha_0^c = \frac{u+E[B^c]}{\sqrt{Var(S^c)}}$ to get

$$P[U^c < 0] = 1 - \Phi^{-1}(\alpha_0^c) \leq 0.005 \iff \Phi(\alpha_0^c) \geq 0.995 \iff \alpha_0^c \geq \Phi^{-1}(0.995)$$

Now, we simplify and rework the equation to solve for optimal α

$$\alpha_0^c = \frac{u + (\theta - \xi + \xi\alpha)E[S]}{\alpha\sqrt{Var(S)}} = \frac{u + (\theta-\xi)E[S]}{\alpha\sqrt{Var(S)}} + \frac{\xi E[S]}{\sqrt{Var(S)}} \geq \Phi^{-1}(0.995)$$

Assume $\Phi^{-1}(0.995)\sqrt{Var(S)} - \xi E[S] < 0$, then we obtain the following expression for α:

$$\alpha \leq \frac{u + (\theta-\xi)E[S]}{\Phi^{-1}(0.995)\sqrt{Var(S)} - \xi E[S]}$$

Results

Parameter estimation

We want to know n, the number of policies. Which, in our case, is $n = 1154$ in the year 2022. Furthermore we want to know the total volume $v = \sum_{i=1}^{n} v_i = \sum_{i=1}^{1154} v_i = 17327$.

Now, we want to estimate λ. To estimate λ we will use Maximum Likelihood Estimation (MLE). We know $\hat{\lambda}_{MLE} = \frac{1}{v} \sum_{t=1}^{n_{2000-2022}} N_t$ which implies that $\hat{\lambda}_{MLE} = 0.05905$, where $n_{2000-2022}$ is the number of policies in the years from 2000 up until 2022.

Also, since we have $E[Y] = 22406$, and $Var(Y) = 116870345$, we obtain that $E[S] = 22923820$ and $Var(S) = 633103186536$.

Then, we know that for the Monte Carlo simulation we have hat $N \sim \mathsf{Poi}(\lambda \mathsf{v})$, where the values of λ and v are as stated above. The Gamma distribution parameters c and γ have to be computed in order to approximate the distribution of the individual claim sizes. After computing, we get that $\hat{c} = 0.000192$ and $\hat{\gamma} = 4.295700$. In Figure 7, it is seen that the fitted Gamma distribution fits almost perfectly the empirical distribution of the individual claim sizes, therefore we assume that it is reasonable to use it for the further calculations.

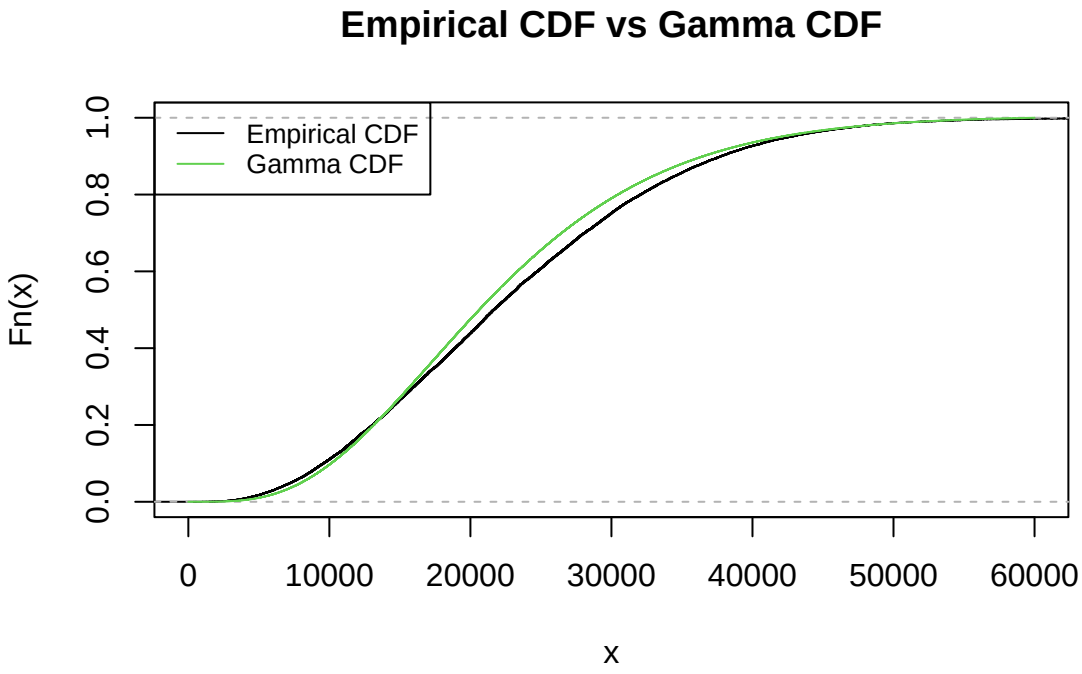

Figure 7: Empirical CDF vs approximated Gamma distribution

Estimating s-values by two methods

After estimating the values and calculating everything as discussed in the section Methods, we obtain the final results for the Capital at Risk values. As seen in Table 2, the differences between the Monte Carlo simulation and normal approximation for the capital at risk values are very small; all of them are identical except the value of $s_{0.99}$, which slightly differs. The differences are very small because the sample size is quite

large; the results of the normal approximation and simulation are likely to be very similar when a sample size is large. However, that would not necessarily be the case if our sample size was significantly smaller.

Table 2: Results of s using different methods of estimation

Method	s at 0.90	s at 0.95	s at 0.99
Normal approximation	23 943 522	24 232 593	24 774 843
Monte Carlo simulation	23 943 522	24 232 593	24 775 256

Even though the results are almost identical, we consider the Monte Carlo simulation a more convincing method as it does not assume a certain distribution for the variable - meaning it is (more) robust and (more) flexible.

Proportional Reinsurance

We obtain the initial capital value to be $u \approx 704832$. Hence, this is the initial capital value that the insurer must have in order to avoid the default at 1% probability.

For the given $\theta = 0.05$, $\xi = 0.07$, and the reduced default probability being 0.005 as well as the initial capital value, we obtain that the optimal alpha $\alpha = 0.554$. This means that the cedent can take around 55% of the expected total claim size to receive the maximum profit subject to the ruin probability of 0.5%. It can be also seen from Figure 8. Since the ruin probability function is increasing, all alpha values that are below the intersection between the line of target value and the ruin probability function are feasible. For the optimal value of α we need to look at the intersection between these two values, which results in the maximum profit.

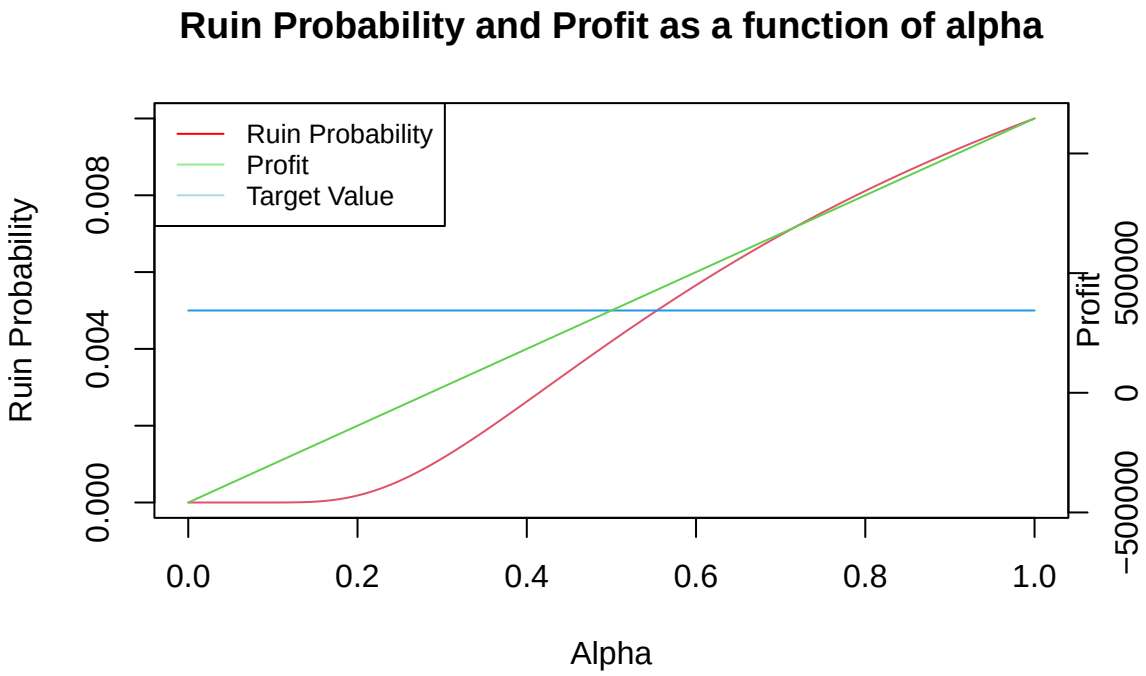

Figure 8: The plot of ruin probability and profit as a function of alpha

Possible Improvements

In order to improve our assignment, we could have added different risk classes to the data. In this way, we differentiate the premiums per (homogeneous) class. To divide the different risk classes, we would probably would have distinguished classes according to how many claims one person would have made in previous years. Furthermore, we saw in the Data Description and Analysis that there is a (slight) positive trend between the mileage and the claim size. This means that when someone is on the road a lot, the risks of having an accident increases. We could also take this into account by adding extra classes based on how many miles someone has driven in the past. To illustrate how the different risk classes could look like, we made an example table (Table 3):

Table 3: Possible risk classification

	0 claims previous years	1 claim previous years	2 or more claims previous years
Driven miles previous year <= 20 000	Class(1,1)	Class(1,2)	Class(1,3)
Driven miles previous year > 20 000	Class(2,1)	Class(2,2)	Class(2,3)

In addition, splitting claims by their size as well as other characteristics, e.g. miles driven, would allow to determine separate claim size distributions for different risk classes. This would result in a more accurate calculations of Capital at Risk, etc.

Also, risk classes would be helpful for tariffication. One can calculate the different premiums by means of the Bailey&Simon method, Bailey&Jung method, or by GLM estimation. This would result in a more personal tariffication and could therefore lead to a higher profit for the insurance company.

Because climate change also is an (big) issue nowadays, we could even charge everyone that drives diesel a higher premium than others, or even lower the premium of the ones who drive electric cars. This probably would not really increase the profit, but it stimulates people to drive environmentally friendly. However, charging males a higher premium than females, or the other way around, would not be a great idea we think. This will deliver a lot of problems and eventually resulting in people running away.

Conclusion

In conclusion, we applied the collective risk model to the car insurance problem, in which the individual risks are aggregated together and the aggregate claim size distribution follows compound Poisson distribution. We used this model to evaluate s values in two different methods, namely normal approximation and the Monte Carlo simulation.

It turned out that the results of both approximation methods are very similar because there were plenty of data. The more data we have the better approximations we can make. In other words, as volume $v \to \infty$, the claim size distribution approaches the standard normal distribution, whereas for the Monte Carlo simulation the large number of simulations allowed us to approximate the aprroximately real distribution of the aggregate claim size amount.

For the calculation of initial capital u, the normal approximation method was used as well. The initial capital value was needed in order for the insurer to ensure that the risk of the insurance company going default did not exceed 1%. When one wanted to reduce its ruin probability to 0.5%, the insurer could apply proportional reinsurance. This resulted in the optimal alpha value $\alpha = 0.554$.

Instead of charging everyone the same premium (Egalitarian method), we could split people to homogeneous groups, i.e. risk classes, for example, by distinguishing how many miles someone has driven or amount of claims someone has made. This would allow the insurer to charge a more appropriate premium, and thus potentially increase profit by getting more customers.

Bibliography

- Insurance Europe (2022), Motor Insurance https://www.insuranceeurope.eu/priorities/20/motor-insurance
- Oxford Dictionary, Insurance, Definition 4. Commerce https://www.oed.com/dictionary/insurance_n?tab=meaning_and_use&tl=true#401861
- Mathilde Carlier (2023, August 28), Revenue - automobile manufacturing industry worldwide 2019-2022, Statista https://www.statista.com/statistics/574151/global-automotive-industry-revenue/#:~:text=The%20global%20automotive%20manufacturing%20market,trillion%20U.S.%20dollars%20in%202022
- Mathilde Carlier (2023, May 24), Motor vehicle sales worldwide by type 2016-2022, Statista https://www.statista.com/statistics/1097326/worldwide-motor-vehicle-sales-by-type/