Comprehensive insurance method General Project

by Kevin Santoso

29949637

ETC3420 - Applied insurance methods S2 2020

Dan Zhu

Monash University

Clayton Victoria

11-09-2020

Abstract

We are living in an environment where the needs to manage various type of risks with a significant economic impact has been a pivot thing in the society. Humanity has grown accustomed to the need of feeling secured , the need of protection has become more pronounced more than ever , leading to the growing demand of financial security against future possible losses. All the things above has led to the raise and the upgrowing development of insurance industry related to the instinctual need of being protected related in their assets against loss from particular events. The entire process of insurance consists in offering an equitable method of transferring the risk of a contingent or uncertain loss in exchange for premium.

Non-life insurance business has recently got some popularity and increased interest on it because of its ability to manage large number of situations (regarding the increased number of vehicle and the increased number of accidents) with a wide variety of risk. The most basic need and goal of insurance industry is to calculate the appropriate and applicable price to correspond to an insured in order to cover a certain risk both to the insurer and the insured. A well-known method to calculate the premium is to multiply the conditional expectation of the claim frequency with the expected cost of claims. Therefore, modelling frequency of claims, also known in theory as count data, represents an essential step of non-life insurance pricing.

1. Introduction

1.1 Background Research

Comprehensive insurance is a type of policies that helps to cover the cost to replace or repair the vehicle insured due to damages or accident other than collision. Comprehensive insurance , or sometimes called “other than collision” coverage , covers a wide range of damages from fire , vandalism , or falling object (like tree or hail). For people intended to lease of need financing for their vehicle usually required to have comprehensive insurance. If you own your vehicle outright, it's an optional coverage on your car insurance policy.

This type of comprehensive car insurance is typically more popular in America , as in Australia comprehensive insurance sometimes still includes the damages from collision Comprehensive insurance is bought only in addition to the compulsory liability insurance (liability insurance) in the same insurance company. The insurance policy is always one. Comprehensive insurance already includes high coverage limits, coverage in both directions (liability and car) - liability, coverage of a car if it just stands parked and something happens to it - comprehensive, coverage of the liability of an uninsured person who either drove your car or flew into a car - uninsured motorist liability, payment for a rented car for the period of repair after an accident or roadside assistance along with paying for a tow truck.
In GEICO and All State comprehensive insurance has a "voluntary-compulsory" package of insurance products, which is most often drawn up taking into account the wishes not of customers, but of bankers seeking to protect themselves from various risks.

The issue of auto insurance in the United States is complicated by the fact that the country does not have a federal system that regulates this industry. Each state is independently engaged in the formation of requirements for car insurance.

However, in this regard, all states can be divided into 2 categories:

- Regardless of who is to blame, the damage is covered by your insurance (no fault). This type of insurance is common in 12 states - Florida, New York, Michigan and others.

- The responsibility rests with the at fault. This is how they work in 38 states, including California and Washington.

1.2 Project Aim

In this report , we are aiming to investigate the relationship of previous claim data , model the appropriate distribution to capture the claim data and finally discuss the appropriate price of the premium to minimize the probability of ruin and satisfy the need of both insurer and the insured. Data sets used will be extracted from last year's claim experiences data given from 5655 sample policies and 8 variable given.

1.3 Report Structure

The report comprised of 2 main sections ,

- The data analysis and modelling
- The discussion of the model and pricing implications

The report begins with a description of the data used for the modelling and estimation of pricing. The model selection and estimation , also the simulation of the model chosen using Monte Carlo. The result then used to estimate and choose a pricing option for the portfolio. Finally, limitations of the model selection and pricing option are discussed, and recommendations made for future improvement of portfolio.

2. Data and Methodology

In this assignment , we are working with the claim from comprehensive insurance data . The data will be concerned about the number of claims from each policy holder and the severity of claims for each policy holder , for simplicity in the number of claims , we are assuming the maximum number of claim for each policy holder is 1 , that means that each policy holder can only have 2 choices (claim or not claim) and the severity of each claim is the Total amount of claims for each policy holder.

2.1. Sample data

The sample contains 5655 observations from claims2020.xls that use 6 exogenous variable for every policy as well as the number of claim and the total claims for each policyholder. The Number of claims will be the response variable , while the total claims will be assumed independent of the other explanatory variables , and will be modeled as a new distribution. Thus other than the two explained variable , the other variables are considered risk factor that has been accounted priorly by the insurer.

Table 1 summarizes the information available about each policyholder.

Variables	Description	Values
Name	Name of the Policyholder	
Gender	Gender of the Policyholder	Male (M) or Female (F)
Age	Age of the Policyholder	16 to 90 years old
MStat	Marriage Status	Married (M) or Single (S)
PostCode	Postcode of states	
Population	Number of Population in the states	From 150 to 104,000

Among the 6 Variables given Name could not be use as a predictor , as no matter what , name would not impact the number of claims. While the Postcode and population is related to each other , as the people lived in the same postcode will have the same number of population , so we will only use one of the 2 variables , and that will be the number of population.

In the data observation , we will be grouping the 2 explanatory variable to make us easier in seeing the relationship between the response variable and the explanatory variable

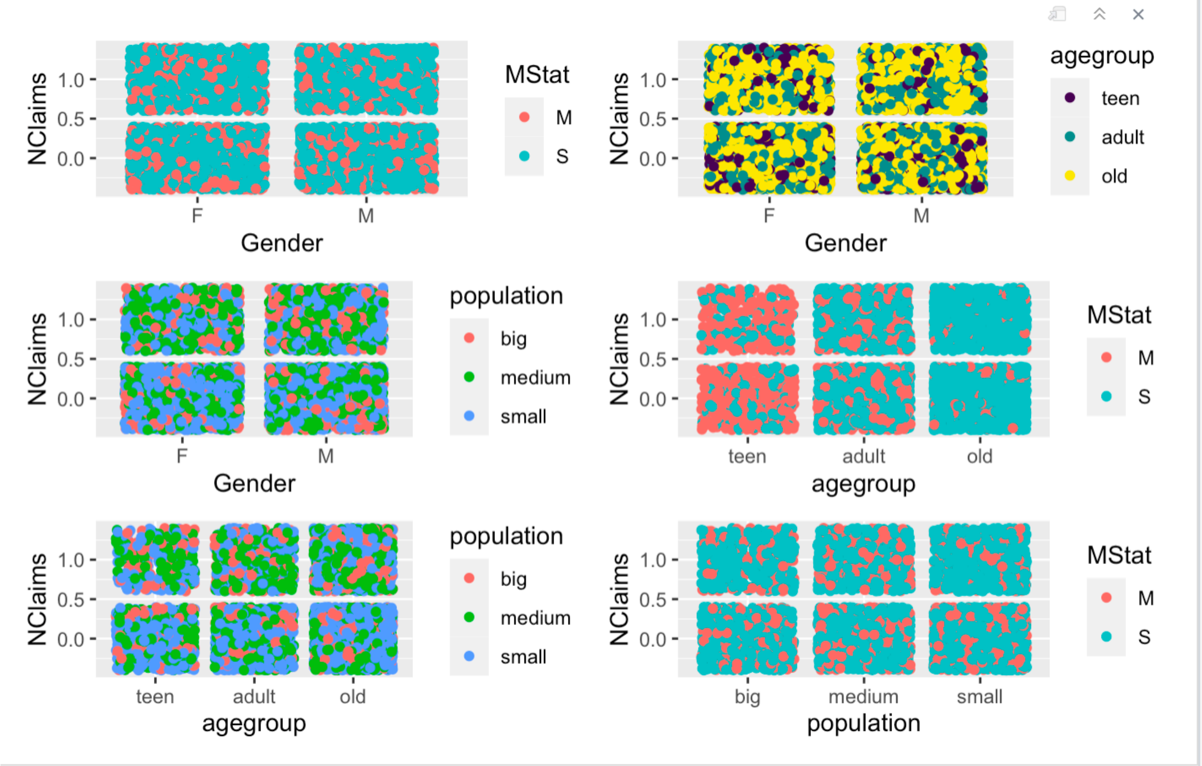
agegroup from Age =
“teen” (16-25) “adult”(26-50) “old”(> 50)

population from Population =
“small”(<35000) “medium”(35000-70000) “big”(>75000)

2.2 Sample regression and quicklook of the data

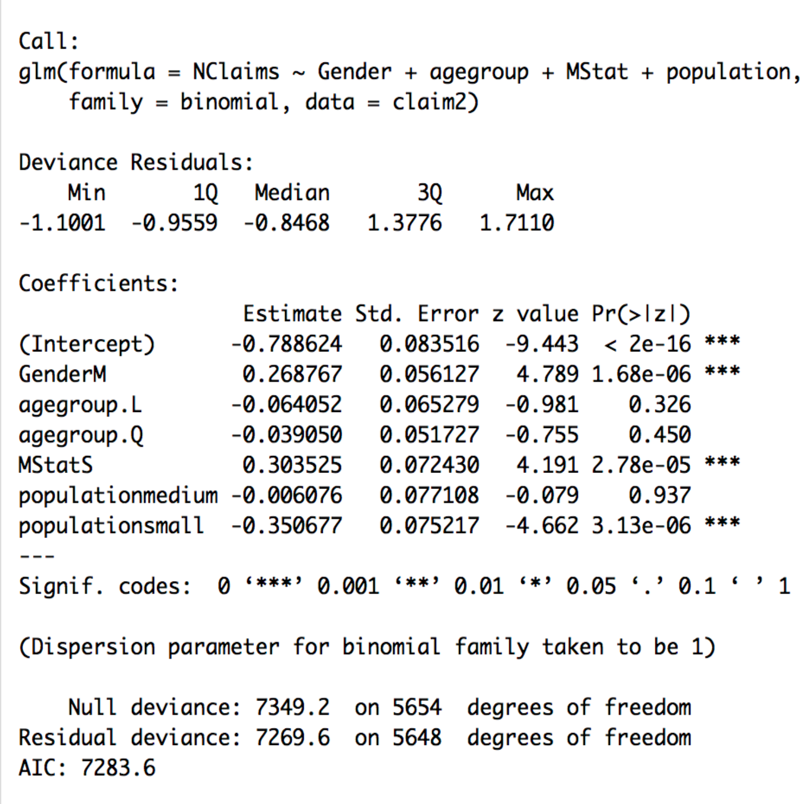
To assess how the explanatory variable impacting the response variable , we are going to plot our sample data against the response variable.

Figure 1 will illustrate the relationship between each explanatory var against NClaims



From the figure m each of the explanatory variables looks like distributed randomly or evenly with no direct impact to the response variables NClaims. To gain further understanding against the data , we are doing a regression analysis against the response variable NClaims.

Figure 2 shows the regression analysis of NClaims



```
Call:
glm(formula = NClaims ~ Gender + agegroup + MStat + population,
    family = binomial, data = claim2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1001  -0.9559  -0.8468   1.3776   1.7110

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.788624   0.083516  -9.443  < 2e-16 ***
GenderM           0.268767   0.056127   4.789 1.68e-06 ***
agegroup.L       -0.064052   0.065279  -0.981    0.326
agegroup.Q       -0.039050   0.051727  -0.755    0.450
MStatS            0.303525   0.072430   4.191 2.78e-05 ***
populationmedium -0.006076   0.077108  -0.079    0.937
populationsmall  -0.350677   0.075217  -4.662 3.13e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7349.2  on 5654  degrees of freedom
Residual deviance: 7269.6  on 5648  degrees of freedom
AIC: 7283.6
```

Looking at the regression , the agegroup may not be a significant predictor to NClaims , while the other 3 (Gender , populations , and MStats) may be a significant predictor.

From the gender , we know that Male has 26% higher chance to claim rather than female
From the Mstats , we know that single or unmarried has 30% higher chance to claim
From the populations , we know that for population less than 35000 , it has 35% less chance to claim than a population more than 35000

The 3 of them might be a significant predictor that affect the distributions of NClaims

3. Model Estimation and Simulation

Within non-life insurance, when actuaries are interested in estimating the frequency of claims, Model Estimation is and simulation is often considered. In the model Estimation we often works by estimating the moment of the data and making a new parameter for the model. In this assignment we are working with Individual Risk Model or Compound Binomial Model.

3.1 Model Assumption

In this model estimation we are working with a few assumption

a. The chance of each policyholder making a claim is independent of the explanatory variable , it is independently distributed with the distribution that we are going to estimate.
b. Each policy holder chance of claim and claim size is independent with each other , that means each policyholder has no power to influence other policyholder.
c. Maximum claim for each policy holder is 1 , and the claim size will be the total claim from the policy holder for the specific year
d. The total claim that the insurer must bear will be written as

$$S = \sum_{i=1}^{N} X_i$$

Where S = Total claim the insurer must bear
N = The number of claim made
Xi = The claim amount made by the i-th policy.

All of the assumption made could be justifiable as we assume the claim number is independent on individual trait because in this comprehensive insurance , the insurer does not cover collision , as the accident caused by natural or other caused is random and cannot be predicted by mankind. There is some evident that individual trait may be a good predictor for the accident caused by vandalism , but as the cause is too random and there is no direct relations from age , gender etc to the chance of getting vandalised , we will just assume that the risk is independent of each individual traits.

For the independency of each policyholder claim from each other , we could argue that the claim of a person does not impact another person to make a claim.

For the maximum claim , as we do not know the claim severity for each claim , and could only assume the claim size for a person is always the same , we may get better by assuming there is only maximum of 1 claim for each person and it is the total claim amount to simplify our model & still get a good representation as we want to know the severity claim of a person in a year , not each claim.

3.2 Model for the number of claim made using binomial model

Binomial distribution is one of the most commonly used distribution. Binomial distribution models the probability of obtaining one of the two outcomes (Success/Failure , or in this assignment 0/1) under a given number of parameters. It summarizes the number of trials assuming each trials has the same chance of obtaining one specific outcome.

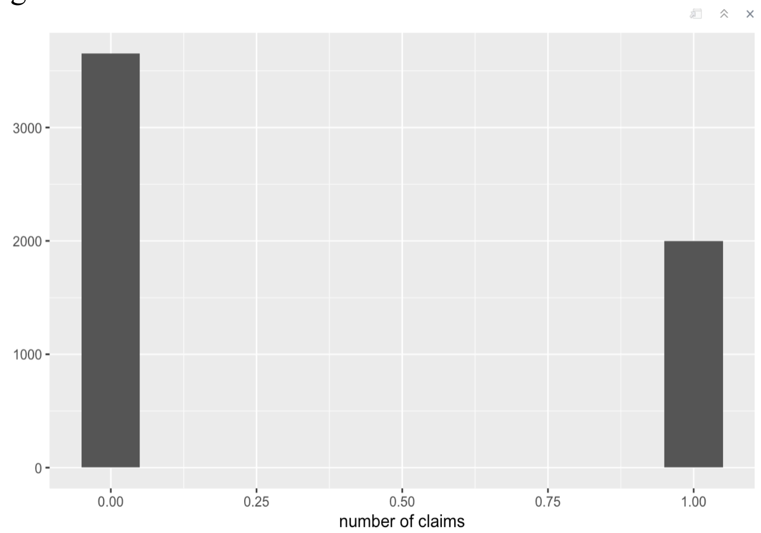
In this assignment , we are interested with the probability of claim for each policyholder and can be modelled after

i ~ Bernoulli(p) $\quad$ i $\begin{cases} 0 \;\; with\, probability\; (1-p) \\ 1 \;\; with\, probability\; p \end{cases}$

Then we summing each policyholder and make a new distribution of

N ~ binomial(n,p) $\quad$ where n = number of samples
p = probability of claim for each policyholder

From the sample we are working with , we are given 5655 observations with 2001 policyholder among them make 1 claim or more.



From the sample size , we could find the mean and the variance of the sample data

Sample mean $\quad \bar{x} \quad$ = 0.3538462
Sample Variance $\quad s^2 \quad$ = 0.2286795

Using the Sample mean and variance we could use Maximum Likelihood Estimator to find the parameter of the Bernoulli distribution

$$L(p \mid x_i, \dots, x_n) = \prod_{i=i}^{N} p_i^n (i - p_i)^{1-n}$$

Setting $\frac{d}{dp}\, L(p \mid x_i, \dots, x_n) = 0$, we get $\hat{p} = \sum_{i=i}^{N} \frac{X_i}{N}$
So $\hat{p} = \; \bar{x} = 0.3538462$

Thus the distribution of N can be simulated with
$$N \sim Binomial\;(\,n\,, 0.3538462\,)$$

3.3 Model the amount of claim size

The severity of claim is another important part to model. The severity of claims depend harshly with the number of claim from each policy holder , and the number of claim is limited to 0 or 1 , the distribution of claim severity can be written as

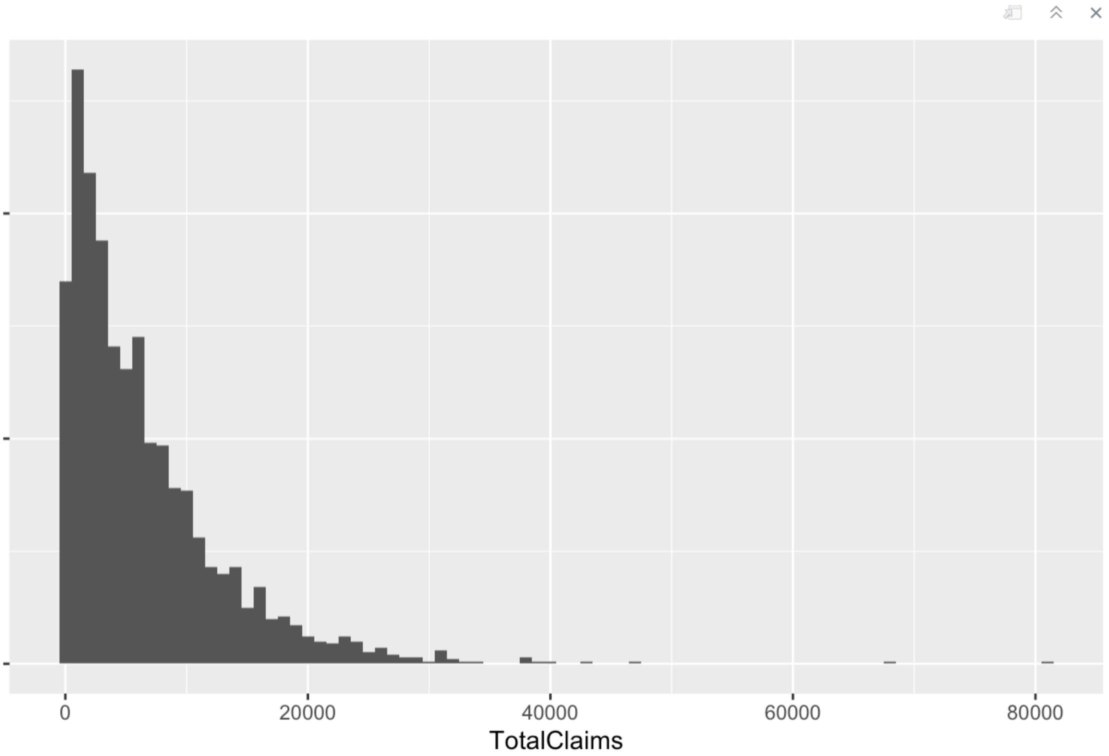
$$f_{Yi}(y) = \begin{cases} (1 - p_i) & if\ y = 0 \\ p_i\big(f_i(y)\big) & if\ y = 1 \end{cases}$$

Or if we include the information from N_i

$$f_{Yi}(y \mid N_i) = \begin{cases} 0 & if\ N_i = 0 \\ f_i(y) & if\ y = 1 \end{cases}$$

Looking at the sample data , from the sample observation we have 2001 policyholder having more than 1 claim , with each of the policyholder claim size is independent from each other.

The figure 3 below shows the histogram from the claim size



From the sample data , we could calculate the sample mean of the sample data

Sample mean $\bar{x}$ = 6524.98

Using the Sample mean we could use Maximum Likelihood Estimator to find the parameter of the gamma distribution

$$L(p \mid x_i, \dots, x_n) = \prod_{i=i}^{N} \frac{1}{\Gamma(k)\vartheta^k}\, {x_i}^{k-1} e^{-\frac{x}{\vartheta} x_i}$$

Setting $\frac{d}{dp}\ L(p \mid x_i, \ldots, x_n) = 0$,

We get $\hat{k} = \frac{1}{2\ln\left(\frac{\bar{x}}{G.M}\right)}$ where G.M is the geometric mean of the sample

$\hat{k} = \frac{1}{2\ln\left(\frac{6524.98}{3757.89}\right)} = 0.906$

And $\hat{\vartheta} = \frac{\bar{x}}{k} = \frac{6524.98}{0.906} = 7200.71$

So the distribution of claim size can be modelled as

$$f_{Yi}(y \mid N_i) = \begin{cases} 0 & if\ N_i = 0 \\ gamma(\ 0.906\ , 7200.71) & if\ y = 1 \end{cases}$$

Where

- Y_i is constructed via I the indicator random variable
- N_i I a strictly positive random variable Xi

And we create a new variable X_i such that

- $X_i = (Y_i | N_i = 1.\ )$
- $X_i \sim f_i(x) =$ as the amount of the claim made by the i-th policy such that
  - E[X_i] = μ_i $= k\vartheta$ = 6,523.843
  - V[X_i] = σ^2 $= k\vartheta^2$ = 46,976,303.4 , or $\sigma = 6853.93$

3.4 Combined Model and simulation

3.4.1 Combined Model

Our Total claim model is written as

$$S = \sum_{i=1}^{N} Y_i$$

Considering the distribution of number of claims (N) and the claim size (Y_i)

$$N \sim Binomial\ (\ n\ , 0.3538462\ )$$

$$f_{Yi}(y \mid N_i) = \begin{cases} 0 & if\ N_i = 0 \\ gamma(\ 0.906\ , 7200.71) & if\ y = 1 \end{cases}$$

We are going to combine the number of claims and the claim size to make a new separate model as a compound binomial model as such that

Y_i is a Compound Binomial random variable with parameters (1, 0.3538462 and $f_{Yi}(y\ )$)

Which we can derive the expectation and variance

$$\begin{aligned} \mathrm{E}[Y_i] &= \mathrm{E}[\mathrm{E}(Y_i|N_i)] = \mathrm{E}(Y_i\ |N_i{=}\ 1)]\Pr(N_i = 1) + \mathrm{E}(Y_i\ |N_i = 0)]\Pr(N_i = 0) = p_i\mu_i \\ &= (0.3538462)(6{,}523.843) = 2308.437055 \end{aligned}$$

$$\begin{aligned} \mathrm{V}(Y_i) &= \mathrm{E}[Y_i^2] - E[Y_i]^2 = \mathrm{E}[\mathrm{E}[Y_i^2|N_i)])] - p_i^2\mu_i^2 = p_i\ [((\sigma_i^2 + \mu_i^2)) - p_i\mu_i^2] \\ &= p_i\ [\sigma_i^2 + (1 - p_i)\ \mu_i^2] = 26353385.73 \text{ or } \sigma_i = 5133.55 \end{aligned}$$

3.4.2 Model Simulation with Monte Carlo simulation

Monte Carlo simulation is one of the commonly used method to understand the impact of risk and uncertainty in models forecasting. Monte Carlo simulator provide visualization of the potential outcome to give better understanding on the risk of a decision.

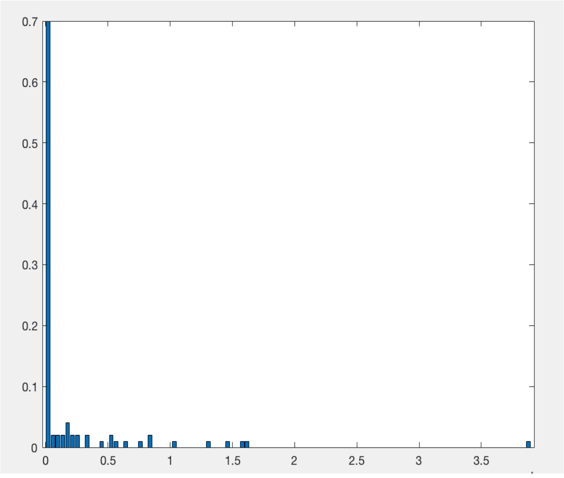
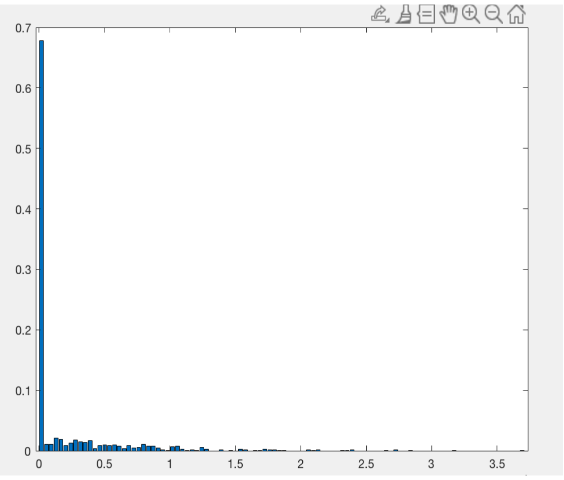
Monte Carlo does the risk analysis modelling by building models of possible results by substituting a range of values from the probability distribution for any factor that has inherent uncertainty. Monte Carlo then keep repeating the process over and over , each time using a different set of random values from the probability functions. Depending upon the number of uncertainties and the ranges specified for them, Monte Carlo simulation will simulate the process thousand or tens of thousands according on what we specified before it is complete. Monte Carlo simulation then will visualize the distribution of possible outcome value.

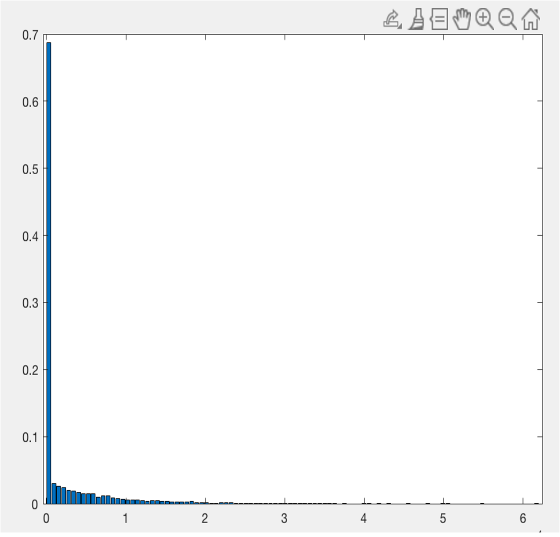
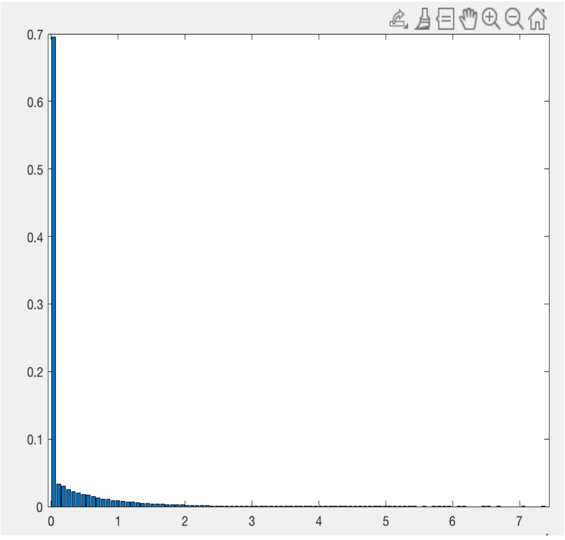
By including probability distributions , variables can have different probabilities of different outcomes occurring. Probability distributions are a much more realistic way of describing uncertainty in variables of a risk analysis.

In this assignment we are going to simulate the Monte Carlo simulation of our distribution and compare it with the sample data

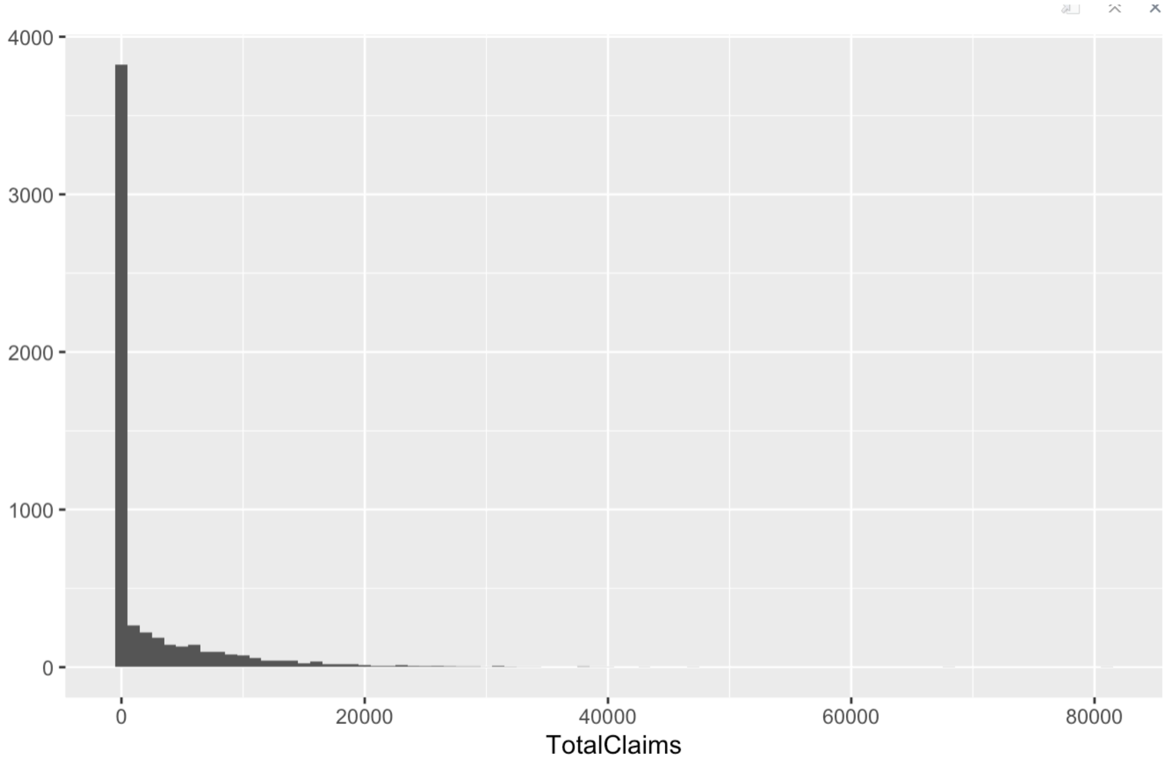
$$Y_i \sim \text{CBin}(1,\ \ 0.3538462 \text{ and } f_{Yi}(y\ )\ ) \text{, where}$$
$$f_{Yi}(y \mid N_i) = \begin{cases} 0 & if\ N_i = 0 \\ gamma(\,0.906\,, 7200.71) & if\ y = 1 \end{cases}$$

Notes : All X are in 10000

n = 100	n = 1000
	

n = 10000	n = 100000
	

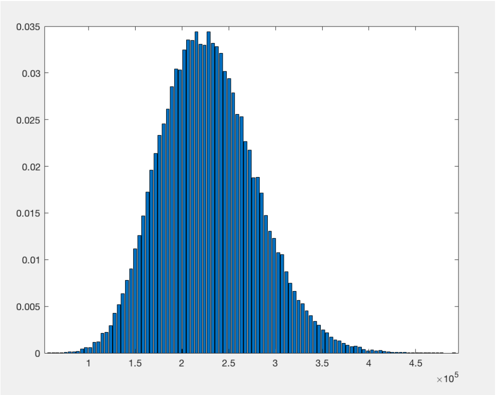
Sample data distribution



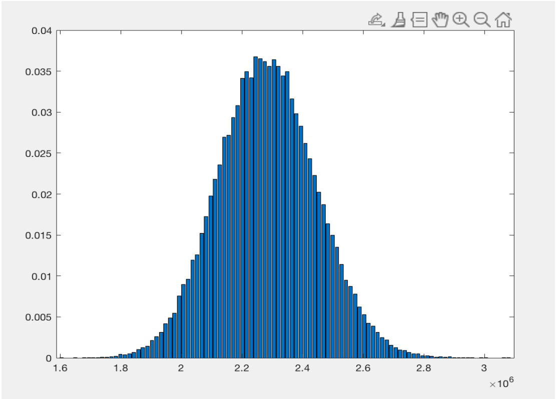
From the simulation and comparing it with the sample data , we could see that while the Monte Carlo simulation has longer tail than the sample data, when n is large enough (> 10000) the Monte Carlo simulation has a similar distribution to the sample data. It means that our model estimation of Y_i ~ CBin(1, 0.3538462 and $f_{Yi}(y\ )$) is a good enough model for forecasting the claim number and claim size if the number of data is high.

Now we are looking at the distribution of S
Since S the sum of all individual claim in the portfolio and the number of policies will be sufficiently large , we could theorize that as n grow larger , the distribution of S will follow normal distribution , so doing a Monte Carlo simulation with 100 , 1000 , 10000 and 5655 like the sample number policies give result as below

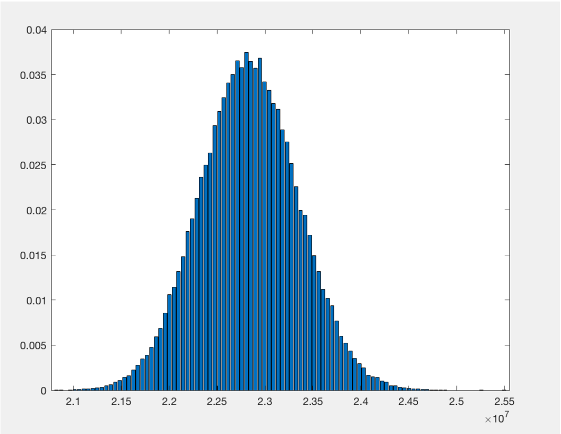
100 policies



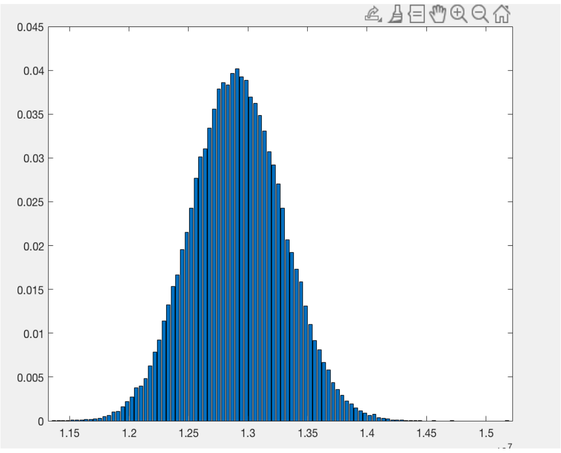
1000 policies



10000 policies



5655 policies



From the result of the Monte Carlo simulation , we could see that the distribution looks approximately normal , now we are going to estimate the mean and variance of S

E(S) = E(N). E(Y)
= 13054211.55

Var(S) = Var(N).E(Y^2)
= 340245.5956

3.5 Pricing

In this pricing section we are trying to use 3 simple pricing option which is the mean , mean plus standard deviation , and mean plus 2 standard deviations. All of data used from the modelling of S and will be named price A,B,C

We are going to try 3 pricing method in this assignment which is

Prem	A. $= E(S)/N$	$= 2308$
	B. $= (E(S)+s.d(S))/N$	$= 2369$
	C. $=(E(S)+2s.d(S))/N$	$= 2429$

For in a , there is 50% chance of the total claim greater than the total premium received , 32% in b and 5% in c

3 of them are good pricing option example as in A the expected profit is 0 , it is in line with insurance principle of equity , also represent the minimum amount of premium the insurer could give. In B & C the insurer wants to get some profit and minimize the probability of loss, and 1 standard deviation or 2 standard deviation away is a good choice as it represent the threshold of the variations from the estimated model.

4. Discussion of the method and implication in pricing and risk management

From the Data modelling , we have found that the model Y_i ~ CBin(1, 0.3538462 and $f_{Yi}(y\,)$) , and S (13054211.55 , 340245.5956) could pretty much capture the distribution of the sample data. In this part of the assignment , we are going to discuss the strength and the weakness of our model , and discussing the appropriate value of the premium including the implication of the pricing and the risk management.

4.1 Discussion of the model

We have estimate the model and chose the most appropriate model from our assumption in this assignment. There is several implication as well as advantages in choosing this model.

For the implication , using this model we have to assume a few things , including the independency of each policyholder & there are no risk that affect the probability of claim , which could be not the case for the real world application. If we got more data & more variable , there might be some factor that could directly impact the possibility of claim making our assumption not true and change the distribution. Using this model we also could not take account of some immediate changes in the global condition , like the pandemic or demonstration that may improve the rate of vandalism.

The model also assume that the maximum amount of claim for each policyholder to be 1 , making it sometimes has a few disadvantage compared to the model that can take account the number of claim for each policyholder. As for the more number of claim there may be more cost implicated to the insurer instead of grouping it into 1.

On the other side , this model also has several advantages . Aside than this model is easier to model compared to the model without assumption , this model could also represent more diverse type of population / people , since we leave out the specific personal information for each policyholder & assume them independent. From the Monte Carlo and comparing it to sample , we could also see that although it could not capture all the data , this model could at least capture the shape of the sample data.

For the modelling of S this model have some advantages that is , it is easy to apply , it also require a little information to do as it only need the expected value and variance of S . Another thing is this approximation is reasonable for large n, since S is the sum of a number of random variables. The larger the number of variables summed, the closer is the distribution of their sum to a Normal distribution by the Central Limit Theorem.

The Model chosen may not be the most appropriate model for the data , but the model provide some advantages that make this model usable , such as this model is easy to use and interpret. Assuming the distribution of S is not normal but more to truncated gamma , may make the distribution skewed and could capture the claim distribution which is positively skewed , but using this assumption make this model harder to interpret and not look as nice as the normally distributed one. This model is simple but the simpleness is its own advantages as some people said , The simpler the better.

4.3 Discussion of the pricing and implications of the pricing

In part 3 we have found 3 method of pricing (A,B and C) , which all of them implied the same price for each policyholder , neglecting their individual character , like ages gender etc it could be justified as we could argue that the accident happens randomly as collision is not included in our coverage, but this may led to some future implication if for some reason the individual traits have impact in the chance of claim . Because if the individual traits matter , we could be overcharging some policyholder , which may led to they don't consider to buy our product & undercharge some policyholder which has higher chance of accident .Because of this , the undercharge category would think our product is cheaper which led to most of our policyholder belong to this category, resulting in we get more claim than expected and may led to ruin.

The price a have 50% chance of the total claim of the portfolio greater than the total premium received , making it an undesirable price for the insurer part, as we have 50% chance that we lost more than we get paid making it looks like a gamble , and expected profit of 0.

The price of b has 32% of chance the total claim greater , it is more understandable for the insurer part but , the price also increase by $61 from A. The expected profit of this price is $34955.

The price C has 5% chance the total claim greater than the premium received , making it the most desirable for the insurer part and got the most expected profit of , but it also has some drawback which is more expensive than the other 2 counterpart , which is by $121 from A and $60 from B making the expected profit of $684255.

Another issues is that if the distribution of S is proven to be not normally distributed as we assumed. If it is skewed we may have overcharge / undercharge our policyholder that also may led to ruins.

For all the implications above , we could alleviate the risk by sharing our risk with reinsurer , we assume that we insure 30% of our portfolio to reinsurer, in a we would get A negative expected profit if we get reinsurer . In B we could take up to 7.5% retention level before we got loss , and 15% in C. Another option of reinsurer is we take a stop-loss policy from reinsurer to control the severity of claim.

Highlighting all of the information above we could see that price B may be the best option as the insurer company still have positive expected profit and we also could spare some of the profit for taking reinsurer policy. To improve our policies , we also could take up some research on the policyholder political & geographic condition so we could diminish the probability of claim and decrease our ruin probability.

5. Conclusion

A good and accurate insurance policy pricing allows insurance companies to minimize expected loss and make enough provision for contingencies. The first step to get a good pricing in Comprehensive insurance is to model the claim frequency and severity, which represents the vital point in generating a reasonable price for the policies.

In this Report , we have considered the claim frequency and severity as a separate model of binomial to estimate the probability of claim and a gamma distribution model to estimate the claim severity. We also has considered the joint model of compound binomial to get further understanding of the portfolio, all assuming the risk are independent and cannot be explained by individual trait.

After modelling each model , we also could see that our model could capture most of the properties from sample data making it an adequate model to be a base in the pricing. We also has proved that the distribution of S would follow normal distribution as the number of policyholder grows by using Monte-Carlo simulation, and estimating the expected value and the variance of S.

In the pricing section , we has estimate some pricing strategy with the combination of mean and the standard deviation of the distribution. We also have made some discussion in the future implication that may have happened due to change in the distribution. The result of the pricing also suggest that taking some reinsurer may be an interesting strategy to diminish future loss and share the risk. Based on the pricing strategy and distribution , all of this aspect aim to obtain reasonable premium corresponds to the distribution of risk and thereby respecting the principle of equity in insurance.

Our empirical study hopes to be useful to policymaker by allowing a better understanding on the insure risk and accurate assessment of the insurance company on the distribution of risk leading to solvency , profitability and healthiness of insurance industry.

Bibliography

Esfandabadi, Z.S., Ranjbari, M. and Scagnelli, S.D., 2020. Prioritizing Risk-level Factors in Comprehensive Automobile Insurance Management: A Hybrid Multi-criteria Decision-making Model. *Global Business Review*, p.0972150920932287.

David, M., & Jemna, D. (2015). Modeling the frequency of auto insurance claims by means of poisson and negative binomial models

Kim, J., Kim, S., Park, E., Jeong, S. and Lee, E., 2017. Policy issues and new direction for comprehensive nursing service in the national health insurance. *Journal of Korean Academy of Nursing Administration*, *23*(3), pp.312-322.

Liu, F., 2016. Individual Risk Preferences and Better Car Replacement. *Journal of Finance and Economics*, *4*(4), pp.1-10.