Predictive modeling in P&C Insurance

Himchan Jeong, University of Connecticut

1 Introduction

1.1 What is Actuarial Science?

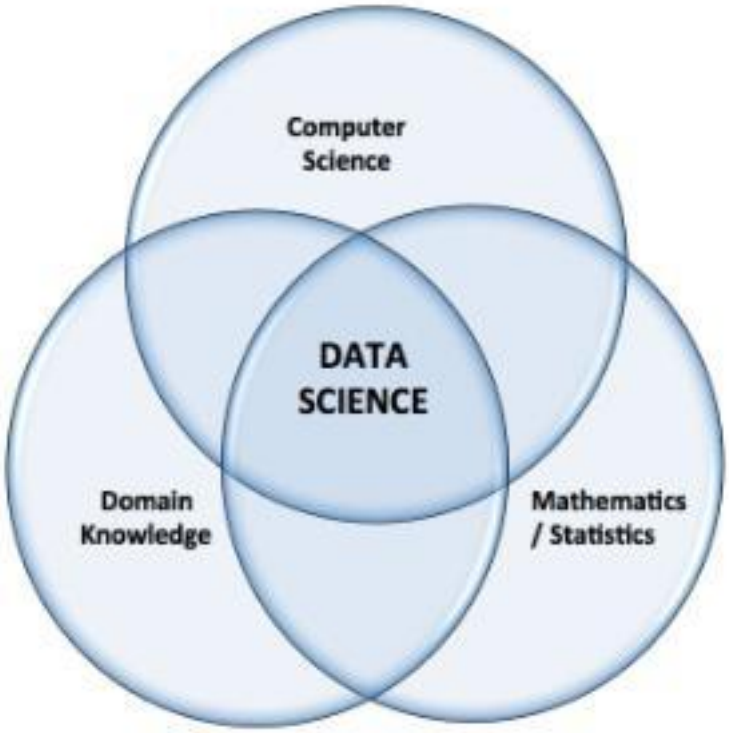

Figure 1: Components of Data Science

"Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in insurance, finance and other industries and professions." (Wikipedia) In short, We need to PRICE given risk for the transaction. Actuary is one of the professions with 'data-driven decision making', for more than 200 years. So actuary can be classified as a type of data scientist whose expertise is in insurance and related industires. Thus, actuaries needs well-developed predictive model both with high predictability and interpretability.

There are a lot of reasons why the interpretability is important in Actuarial Science.

- Tradition
- Internal/External Communication
- Regulation
- Robustness

I want to introduce current practice done by property and casualty (P&C) insurance company, as well as suggest the more sophisticated predictive model which can outperform the benchmarks.

1.2 Common Data Structure

For ratemaking in P&C, we have to predict the cost of claims $S = \sum_{k=1}^{n} C_k$. Policyholder i is followed over time $t = 1, \ldots, T_i$ years. Unit of analysis it – an insured driver i over time t (year) For each it, we could have several claims, $k = 0, 1, \ldots, n_{it}$ Thus, we have available information on: number of claims n_{it}, amount of claim c_{itk}, exposure e_{it} and covariates (explanatory variables) x_{it}, which often include age, gender, vehicle type, building type, building location, driving history and so forth

2 Model Specification

2.1 Current Approches for Claim Modeling

There are two major models which are well-known and widely used in P&C insurance company. First one is two-parts model for frequency and severity, and the other is Tweedie model.

In two-parts model, total claim is represented as following;

$$\text{Total Cost of Claims} \;=\; \text{Frequency} \;\times \text{Average Severity}$$

Therefore, the joint density of the number of claims and the average claim size can be decomposed as

$$\begin{aligned} f(N, \overline{C}|\mathbf{x}) &= f(N|\mathbf{x}) \times f(\overline{C}|N, \mathbf{x}) \\ \text{joint} &= \text{frequency} \times \text{conditional severity}. \end{aligned}$$

In general, it is assumed $N \sim \text{Pois}(e^{X\alpha})$, and $C_i \sim \text{Gamma}(\frac{1}{\phi}, e^{X\beta}\phi)$.

In tweedie Model, instead of dividing the total cost into two parts, we directly entertain the distribution of compound loss S where

$$\begin{aligned} &S = \sum_{k=1}^{N} C_k, \quad N \sim \text{Pois}(e^{X\alpha}) \\ &C_k \sim \text{Gamma}(\frac{1}{\phi}, e^{X\beta}\phi), \quad C_k \perp N \;\; \forall k \end{aligned}$$

in order that it has point mass probability on $\{S = 0\}$ and has the following property.

$$\mathbb{E}[S] = \mu, \quad \mathbb{V}ar(S) = \Phi\mu^p, \quad p \in (1, 2)$$

However, there are some pitfalls in the current practice aforementioned.

- (1) Dependence between the frequency and the severity
- (2) Longitudinal property of data structure.
  - For example, if we observed a policyholder i for T_i years, then we have following observation $N_{i1}, N_{i2}, \ldots, N_{iT_i}$, which may not be identically and independently distributed.

For the first problem, if we assume that N and $C_1, C_2, \ldots, C_n$ are independent, then we can calculate the premium for compound loss as

$$\begin{aligned} \mathbb{E}[S] &= \mathbb{E}\left[\sum_{k=1}^{N} C_k\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{k=1}^{N} C_k|N\right]\right] \\ &= \mathbb{E}[\mathbb{E}[C_1 + \cdots + C_N|N]] = \mathbb{E}[N\mathbb{E}[C_1|N]] \\ &= \mathbb{E}[N\mathbb{E}[C]] = \mathbb{E}[N]\,\mathbb{E}[C] \end{aligned}$$

In other words, we can just multiply the expected values from frequency model and the average severity model to get the estimate for compound loss. However, in general N and C_k are correlated so that $\mathbb{E}[S] \neq \mathbb{E}[N]\mathbb{E}[C]$. If we have positive correlation between N and C, then

$$\mathbb{E}[S] > \mathbb{E}[N]\mathbb{E}[C]$$

so the company suffers from the higher loss relative to earned premium.

On the other hand, if we have negative correlation between N and C, then

$$\mathbb{E}[S] < \mathbb{E}[N]\mathbb{E}[C]$$

so the company confronts the loss of market share due to higher premium.

2.2 Possible Alternatives

There are some possible alternatives which can be used for dealing with the pitfalls.

- For dependence between the frequency and severity
  - Set $\mathbb{E}\left[\overline{C}|N\right] = e^{X\beta+N\theta}$
  - Copula for N and $\overline{C}$
- For longitudinal property
  - Random effects model
  - Copula for multiple claim observation
- Non-traditional approaches
  - Neural networks
  - Regression for each group classified by decision tree

3 Analysis

3.1 Data Description

Here I use a public dataset on insurance claim, provided by Wisconsin Propery Fund. (https://sites.google.com/a/wisc.edu/jed-frees/) It consists of 5,656 observation in traning set and 1,098 observation in test set. It is a longitudinal data with more or less 1,232 policyholder, followed for 5 years. Although the dataset includes information one multi-line insurance, here I only used building and contents (BC) claim information.

Here we can see very high overdispersion. (Variance is much larger than mean.) Therefore, use of Negative Binomial distribution is recommended rather than Poisson distribution. Moreover, there are some 'outliers' which looks too big. (231 claim per year) So they might be wrong records on the claim file.

Usually, for the positive severity, it is traditional to use either log-normal distribution or gamma distribution with log-link.

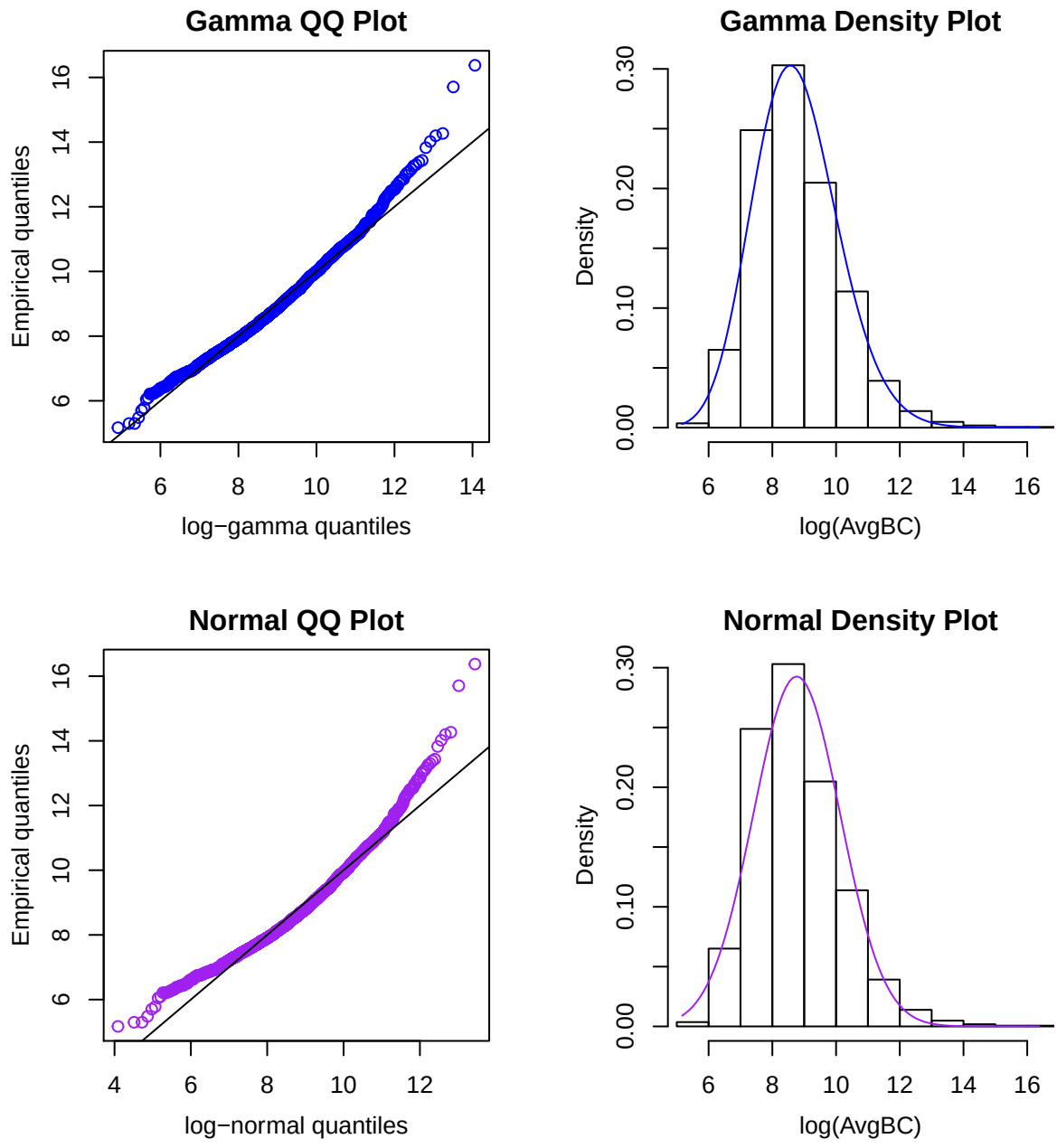

Figure 2: Plots of fitting normal and gamma to average severity

Table 1: Observable policy characteristics used as covariates

Categorical variables	Description		Proportions	
TypeCity	Indicator for city entity:	Y=1	13.97 %	
TypeCounty	Indicator for county entity:	Y=1	5.66 %	
TypeMisc	Indicator for miscellaneous entity:	Y=1	11.09 %	
TypeSchool	Indicator for school entity:	Y=1	28.13 %	
TypeTown	Indicator for town entity:	Y=1	17.34 %	
TypeVillage	Indicator for village entity:	Y=1	23.82 %	
NoClaimCreditBC	No BC claim in prior year:	Y=1	32.94 %	
Continuous variables		Minimum	Mean	Maximum
CoverageBC	Log coverage amount of BC claim in mm	0	34.1	1420.63
lnDeductBC	Log deductible amount for BC claim	0	7.13	11.51

Table 2: Summary statistics for claim frequency

		Minimum	Mean	Variance	Maximum
FreqBC	number of BC claim in a year	0	0.59	2.12	17

Table 3: Goodness-of-fit test for the frequency component

Count	Observed	Poisson	Negative Binomial
0	3993	3126.4	4026.7
1	997	1853.4	855.3
2	333	549.4	365.3
3	136	108.6	182.1
4	76	16.1	97.3
5	31	1.9	54.1
6	19	0.2	30.9
7	19	0	17.9
8	16	0	10.5
9	5	0	6.3
>9	31	0	5.8
χ^2		14392.3	11377.3

Table 4: Summary statistics for claim severity

		Minimum	Mean	Variance	Maximum
log(yAvgBC)	(log) avg size of claim in a year	5.17	8.77	1.86	16.37

3.2 Future Works

- Deal with 'outliers' on the observations for claim frequency.
- Provide methodologies for modelling the claim and compare their performance with those of the benchmark models.
- If possible, suggest a model with higher predictability and interpretability which can be used in P&C insurance company.