# Predictive models for P&C insurance

Himchan Jeong

University of Connecticut

Data Science in Action

6 March, 2018

# What is Actuarial 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.
- Thus, actuaries need well-developed predictive model both with high predictability and interpretability.

# Interpretability Issue on Actuarial Science

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

- Tradition
- Internal/External Communication
- Regulation
- Robustness

# Purpose of the Project

- Introduce current practice done by property and casualty (P&C) insurance company
- Suggest the more sophisticated predictive model which can outperform the benchmarks

# P&C Insurance Claim 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 "$it$", could have several claims, $k = 0, 1, \ldots, N_{it}$
- Have available information on: number of claims $n_{it}$, amount of claim $c_{itk}$, exposure $e_{it}$ and covariates (explanatory variables) $x_{it}$
  - covariates often include age, gender, vehicle type, building type, building location, driving history and so forth

# Current Approches for Claim Modeling

- (1) Two-parts model for frequency and severity
- (2) Tweedie model

# Two-parts Model

- Total claim is represented as following;

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

- 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 $C_i \sim \text{Gamma}(\frac{1}{\phi}, e^{X\beta}\phi)$.

# Tweedie Model

- Instead of dividing the total cost into two parts, Tweedie model directly entertain the distribution of compound loss $S$ where

$$S = \sum_{k=1}^{N} C_k, \quad N \sim \text{Pois}(e^{X\alpha})$$

$$C_k \sim \text{Gamma}(\frac{1}{\phi}, e^{X\beta}\phi), \quad C_k \perp N \ \ \forall k$$

- It has point mass probability on $\{S = 0\}$ and has the following property.

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

# Pitfalls in Current Practices

- (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.

# Premium for Compound Loss under Independence

- 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

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

  In other words, we just multiply the expected values from frequency model and the average severity model.
- In general, $\mathbb{E}[S] \neq \mathbb{E}[N]\,\mathbb{E}[C]$.

# Why is the Dependence Important?

- 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.
- 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.

# Possible Alternatives for the Benchmarks

- 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

# 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 5,677 observation in traning set and 1,098 observation in test set.
- It is a longitudinal data with more or less 1,234 policyholder, followed for 5 years.
- Although the dataset includes information one multi-line insurance, here I only used building and contents (BC) claim information.

# Observable Policy Characteristics used as Covariates

| Categorical variables | Description | | | |
|---|---|---|---|---|
| | | | Proportions | |
| TypeCity | Indicator for city entity: | Y=1 | 14 % | |
| TypeCounty | Indicator for county entity: | Y=1 | 5.78 % | |
| TypeMisc | Indicator for miscellaneous entity: | Y=1 | 11.04 % | |
| TypeSchool | Indicator for school entity: | Y=1 | 28.17 % | |
| TypeTown | Indicator for town entity: | Y=1 | 17.28 % | |
| TypeVillage | Indicator for village entity: | Y=1 | 23.73 % | |
| NoClaimCreditBC | No BC claim in prior year: | Y=1 | 32.83 % | |
| Continuous variables | | Minimum | Mean | Maximum |
| CoverageBC | Log coverage amount of BC claim in mm | 0 | 37.05 | 2444.8 |
| lnDeductBC | Log deductible amount for BC claim | 0 | 7.14 | 11.51 |
| FreqBC | number of BC claim in a year | 0 | 0.88 | 231 |
| log(yAvgBC) | (log) avg size of claim in a year | 5.17 | 8.76 | 16.37 |

# Future Works for this Project

- 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.